

ESTADÍSTICA APLICADA A PSICOLOGÍA Y EDUCACIÓN

Teoría y ejercicios con aplicaciones en Excel



Agustín Dousdebés Boada

ESTADÍSTICA APLICADA A PSICOLOGÍA Y EDUCACIÓN

TEORÍA Y EJERCICIOS CON APLICACIONES EN EXCEL

AGUSTÍN DOUSDEBÉS BOADA

ESTADÍSTICA APLICADA A PSICOLOGÍA Y EDUCACIÓN.

Teoría y ejercicios con aplicaciones en Excel

de Agustín Dousdebés Boada

Primera edición: PUCE, 2021

© 2021 Agustín Dousdebés Boada

Centro de Publicaciones PUCE

www.edipuce.edu.ec

Quito, Av. 12 de Octubre y Robles

Apartado n.º 17-01-2184

Telf.: (5932) 2991 700

e-mail: publicaciones@puce.edu.ec

Dr. Fernando Ponce, S. J.

Rector

Dr. Fernando Barredo, S. J.

Vicerrector

Dra. Graciela Monesterolo Lencioni

Directora General Académica

Mtr. Paulina Barahona

Decana de la Facultad de Psicología

Mtr. Santiago Vizcaíno Armijos

Director del Centro de Publicaciones

 Centro de
Publicaciones
PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

Comité Ejecutivo de Publicaciones:

Andrea Muñoz Barriga

César Eduardo Carrión

Santiago Vizcaíno Armijos

Diseño de portada:

Carlos Andrade Dousdebés

Diagramación:

Rafael Castro

Corrección:

Centro de Publicaciones

ISBN: 978-9978-77-530-1

Derecho de Autor:

Impresión:

Tiraje: 300 ejemplares

Quito, abril 2021

Impreso en Ecuador. Prohibida la reproducción de este libro, por cualquier medio,
sin la previa autorización por escrito de los propietarios del Copyright.

ESTADÍSTICA APLICADA A PSICOLOGÍA Y EDUCACIÓN

TEORÍA Y EJERCICIOS CON APLICACIONES EN EXCEL

INTRODUCCIÓN

En la antigüedad no tenían estadísticas
por lo que tuvieron que recurrir a la mentira.
Stephen Leacock

CÓMO SE DESARROLLA EL LIBRO

La intención de este libro es presentar un formato más amigable en el estudio de la Estadística si se compara con otros textos; además, he tratado de utilizar un lenguaje formal pero con la intención de ser lo más sencillo de entender y evitando un nivel de desarrollo matemático muy elevado, ya que considero no necesario incursionar en este ámbito para lograr el objetivo de la presente obra.

Cualquier tratado sobre Estadística suele provocar dos tipos de reacción, la primera y creo yo la más común es de alejamiento y hasta de cierta aversión dado que es una materia tradicionalmente considerada difícil y, según muchas personas alejadas de la Matemática, poco útil para sus intereses.

La segunda, aunque menos común, es a la que pertenecemos algunos seguidores de esta ciencia quienes la acogemos con cariño, entusiasmo y necesidad de incursionar en ella de tal manera que hasta nos atrevemos a escribir y proponer, en función de la visión que cada uno tenga, de tal manera que tratamos de dar un aporte para su aplicación en muchos y variados ámbitos del quehacer profesional y estudiantil.

¿Qué propongo con este libro? Pues no otra cosa que el estudio de la Estadística a nivel de grado con aplicación a las ciencias Psicológicas y de Educación; con una visión muy práctica en el estudio de los distintos temas a estudiar a nivel universitario, sin complicarlos con una Matemática avanzada pero tampoco dejando de tener un buen nivel de profundidad en el tratamiento de cada uno de los temas planteados aquí.

En esta publicación propongo el uso de Excel como herramienta básica para la resolución numérica de los distintos ejercicios en cada tema propuesto y además, explico con detalle algunas funciones que no siempre están desarrolladas de manera explícita ni en otros

textos, ni en guías de uso de este software, de tal manera que quienes no conocen esta herramienta puedan con facilidad acercarse con mayor seguridad al desarrollo de los ejercicios planteados; es por ello que antes de iniciar los ejercicios de cada tema, se hace una explicación del proceso a seguir en Excel.

En cuanto a esto, el libro proporciona dos tipos de ejercicios, estos son: ejercicios resueltos y ejercicios propuestos. Al final de cada tema se plantean unos y otros y, además, se desarrollan los ejercicios impares de aquellos que se han propuesto para que sean una guía para el alumno

Debo confesar también que estoy consciente de las posibles críticas que puede tener esta publicación, especialmente por el hecho de que consideré innecesario tratar ni desarrollar específicamente un capítulo referente a Probabilidades y también por no utilizar otra herramienta de mayor complejidad como es el paquete Estadístico SPSS.

En cuanto a lo primero, debo comentar que, si no se van a utilizar temas como Probabilidad, tipos de distribución probabilística como: Bernoulli, Binomial, Poisson, Exponencial, Geométrica u otras, ¿qué sentido tiene dedicar páginas enteras a algo que no tendrá aplicación directa dentro del ámbito y propósito de este libro?

En cuanto al segundo aspecto, considero que el paquete Estadístico SPSS es extraordinariamente bueno, pero para el propósito de esta publicación no hace falta, además si comparamos la accesibilidad al Excel por parte de cualquier estudiante que quiera incurrir en el estudio de la Estadística con la de este software, coincidirán conmigo que el segundo es menos alcanzable por muchas razones y que su uso puede tener cierta complicación.

Estoy consciente que se quedan algunos temas por tratar, la Estadística es muy amplia y abarca gran cantidad de aspectos que otros libros sí los abordan, pero la intención de esta publicación es desarrollar los temas que normalmente se alcanzan a revisar en un programa de grado en las dos carreras a las que está dirigida esta publicación; será motivo de otro emprendimiento el revisar temas que abarquen otros aspectos de aplicación más amplio y específico para otro nivel.

Por último, considero necesario hacer una puntualización respecto a un término que con alguna frecuencia utilizo en distintos temas que se abordan en la presente publicación y es el referente a “cociente intelectual”. En algunos libros este término se utiliza igual que en el presente, pero en otros, al igual que muchos profesionales también, utilizan el término “coeficiente intelectual”.

Hay que recordar que el cociente intelectual (CI) se obtiene en base a pruebas de inteligencia y para ello existen varias pruebas psicológicas, pero en cuanto al proceso numérico, el CI es el resultado de la división entre la edad mental y la cronológica y el nombre que se da al resultado de una división se llama “cociente”; es por ello que a mi parecer el término a usar debería ser “cociente intelectual” y no coeficiente ya que este último se usa para determinar un factor numérico que afecta a una variable.

Según lo explicado por Ramón Pérez Juste:

La Edad Mental (EM) es, también una puntuación individual, y su cociente con la edad cronológica (EC) otra diferente, el Cociente Intelectual (CI): $= EM / EC$. La EM indica que una persona tiene una inteligencia propia de una determinada edad (2012, p. 19).

LOS NÚMEROS, LA ESTADÍSTICA Y LA PSICOLOGÍA

Normalmente ocurre que el estudiante que ha escogido una carrera dentro del área o del ámbito de los estudios sociales, suele argumentar que dada su formación y profesión futura no necesitará de los números, además supone (erróneamente) que no le serán de utilidad en el ámbito profesional. A esto hay que añadir que seguramente su experiencia con los números siempre le ha significado una gran dificultad y por qué no decirlo se ha desarrollado cierta “repulsitis” a cualquier estudio que signifique trabajar con números.

Si bien es cierto que esto último no tiene remedio y por tanto al ingresar a la universidad se mantendrá este tácito rechazo a cualquier materia que tenga que ver con cálculos numéricos (obviamente la Estadística está dentro de esta categoría), esto no significa que los esfuerzos que debemos realizar los profesores de estas materias para tratar de convencer a los estudiantes de la importancia y real aplicación de esta ciencia, se vean ya derrotados ante ese generalizado rechazo histórico y temor a cualquier rama de la Matemática.

Es obvio pensar que los números y la Estadística van de la mano, esta última no tendría vida sin los primeros; pero lo que normalmente los estudiantes – y muchos profesionales de las ramas sociales – no alcanzan a visualizar objetivamente, es que la Estadística se refleja positivamente en el quehacer de los estudios de estas ciencias y entre ellas encontramos también a la Psicología y a las Ciencias de la Educación.

Pero no se trata que los estudiantes y profesionales de estas ramas u otras en el ámbito social desarrollen destrezas en cálculos numéricos ni manejo de fórmulas, unas más complicadas que otras; no se trata tampoco de que aprendan (creo yo algo tardíamente) a manejar ciertos procesos numéricos en los que se basan muchos de los cálculos Estadísticos.

La relación de los números, la Estadística y la Psicología está en la capacidad de interpretación que desarrolle el estudiante o el profesional en base a la problemática concreta de estudio dentro del ámbito de las ciencias sociales.

De lo dicho en los dos últimos párrafos se desprende que lo importante no es manejar las fórmulas si no que en base a los resultados obtenidos, el estudiante o profesional pueda ser capaz de interpretar, concluir y recomendar objetivamente sobre temas referentes a cualquier problemática estudiada.

Para esto ventajosamente existen paquetes estadísticos o sin ir más lejos las herramientas que ofrece el Excel, que “facilitan la vida” a cualquier persona que no maneje bien los procesos numéricos ni los – a veces – engorrosos cálculos numéricos. Por tanto, en realidad

el tema numérico estaría “resuelto” con estas herramientas y la relación Estadística – Ciencias Sociales se basará en el desarrollo de la destreza de interpretación objetiva y la relación con las teorías a manejar; de esto debe encargarse el profesor en su práctica diaria en clase con los estudiantes para que, entre otras cosas, éstos últimos vean la aplicación y estrecha relación que en verdad existe entre la Estadística y la Psicología en cualquiera de sus tres especialidades: Clínica, Organizacional o Educativa.

Propongo un ejemplo inicial para aclarar estos puntos:

Supongamos que se requiere determinar el tiempo de reacción de un grupo de personas ante un estímulo externo determinado; ¿qué debemos hacer?, pues nada más que someterles a dicho estímulo y medir el tiempo de reacción. Ya hemos obtenido un conjunto de números (tiempo medido en segundos), ¿y ahora qué?, será necesario entonces establecer “reglas” (las llamaremos parámetros más adelante) para que tenga algún sentido el estudio.

Estas reglas pueden ser por ejemplo que si se demoran más (o menos) de cierto tiempo deben hacerse otros estudios o si al compararlos con otro conglomerado de personas el grupo de estudio reacciona, en general, en un tiempo aceptable.

Aquí es donde entra en escena la Estadística, esta ciencia propone entre otras cosas, establecer relaciones numéricas de datos obtenidos en un estudio; pero para los fines de este libro no debemos quedarnos en la obtención simple de resultados (primer punto de la perspectiva propuesta anteriormente), si no que debemos pasar al segundo punto, es decir, aprender a interpretar con lógica qué nos dicen esos resultados y con ello establecer cuatro aspectos que tienen niveles distintos de complejidad y hay que aprender a manejarlas: observar, comentar, concluir y recomendar. No es suficiente por ejemplo decir que el promedio de tiempo de reacción es “x”, con eso no se ha dado ningún valor agregado al problema; el investigador debe ser capaz de ver y explicar a detalle lo que en verdad se encuentra en los resultados y esta capacidad es lo que hay que desarrollar en quienes incursionan en esta noble ciencia.

Es el momento ya de que las ciencias sociales reclamen su espacio en estas primeras líneas, luego de haber obtenido los resultados numéricos y relacionarlos en base a ciertos criterios, el Psicólogo o educador tendrá una base para poder opinar sobre las condiciones en las que el conjunto de personas en cuestión se encuentra y también sobre la situación particular de los individuos en comparación con el grupo mismo, otros grupos o la teoría base.

Pero la intervención (observar, comentar, concluir, recomendar) del profesional (específicamente en Psicología), depende de dos factores para que el análisis tenga mayor y mejor aceptación:

1. La visión teórica que maneje y
2. La experiencia que posea

Considero que la mayor dificultad realmente radica en la capacidad de interpretar los datos y poder decidir objetivamente en base a los resultados obtenidos, reconozco también que no es fácil ya que requiere entrenamiento y saber ver más allá de lo evidente.

QUÉ SÍ Y QUÉ NO CONTIENE ESTE LIBRO

Una pizca de probabilidad tiene tanto valor
como una libra de quizá.
James Thurber

La verdad al cien por cien existe tan poco
como el alcohol al cien por cien.
Sigmund Freud

En estos más de veinte y tantos años de experiencia como profesor de esta materia en la facultad de Psicología de la Pontificia Universidad Católica del Ecuador (PUCE), he ido cambiando y sobre todo puliendo ciertos temas que conforme iba conociendo más la práctica de los Psicólogos y también de los docentes especialmente a nivel medio, notaba que la Estadística debe centrarse en temas que al final del día realmente sean de utilidad para su práctica profesional; consciente también de que en verdad la Estadística es una herramienta que sirve fundamentalmente para la investigación, he visto que en el currículo de las facultades de Psicología no solo de la PUCE sino de otras, la Estadística debe centrarse en estudiar capítulos que en verdad el estudiante vea lo práctico y útil aunque los libros tradicionales dediquen páginas enteras a estudios complementarios como por ejemplo Probabilidad.

Debemos tomar en cuenta que la Estadística al nivel de grado debe considerarse un instrumento ya sea para el análisis que he descrito en párrafos anteriores o para su aplicación en temas de Psicometría, Evaluación, Análisis Salarial o Desarrollo por ejemplo; y también en el ámbito educativo, para interpretar y decidir sobre aspectos del desarrollo del aprendizaje o para el análisis de situaciones concretas como el rendimiento académico, problemáticas conductuales y mejoramiento de procesos de formación entre otros temas.

Con especial dedicación quiero referirme brevemente al tema de Probabilidad para los Psicólogos y Educadores. He conversado y consultado con muchos profesionales de estas ciencias y cada vez me convenzo más de lo delicado que es este tema en su práctica profesional. Ejemplos como el siguiente referente a probabilidad condicional no reflejan un sentido

práctico de estos estudios: “si a la consulta de cierto profesional de la Psicología llegan consecutivamente 5 pacientes con síntomas de depresión, ¿cuál es la probabilidad de que el siguiente también manifieste esta patología?

Pregunto dos cosas:

1. ¿Qué importancia tiene calcular esto?
2. ¿En verdad esto es calculable?

Respecto al primer cuestionamiento: salvo que el Psicólogo se dedique a investigar las razones de tanta recurrencia de pacientes depresivos, el valor académico de este estudio es, insisto, a mi juicio, nulo; sin embargo, considero que esta herramienta sí sirve para profundizar en temas referentes a investigación en general y también para la Psicología en particular, pero no como parte de los estudios dentro de la carrera.

Y en cuanto a la segunda pregunta, la respuesta en verdad es NO, no se puede calcular por razones de desconocimiento real del universo de pacientes.

Otro asunto respecto a este mismo tema radica en la diferencia que hay entre probabilidad y posibilidad ya que existe en realidad un abuso del lenguaje; sobre esto expongo a continuación un ejemplo muy simple relativo al maravilloso mundo del fútbol: si se realiza un partido en el estadio Camp Nou entre el Barcelona de España (Barça) y el Chacarita Junior (Ecuador), ¿cuál es la probabilidad de que el equipo ecuatoriano gane?, la respuesta inmediata que me han dado todas las personas a quienes he preguntado (conozcan o no de fútbol) es: “obviamente el Barcelona”.

La respuesta no es la correcta por lo siguiente: en términos de probabilidades se debe definir lo que se conoce como Espacio Muestral, esto no es otra cosa que el universo de posibles resultados del partido a jugarse, por tanto el espacio muestral en este caso es el siguiente: gana, pierde o empata, por lo tanto numéricamente el universo es 3 y aplicando la relación matemática de probabilidad que no es otra cosa que la relación numérica dada entre el evento a estudiar dividido para el universo; tendríamos que la probabilidad de que gane el Chacarita Junior es la siguiente:

$$p(x) = \frac{1}{3}$$

¡Que es la misma probabilidad que tiene el Barça!

Es decir, tiene un 33.33% de probabilidad de ganar, pero contrastando esto con la **posibilidad** “intuitiva” de que en realidad gane, el criterio (supongo unánime) será que no lo hará.

En términos de situaciones prácticas en pedagogía un ejemplo puede ser: “si un estudiante terminó la prueba en 30 minutos y lo hizo bien, ¿cuál es la probabilidad de que la siguiente vez ocurra lo mismo? Mi pregunta es: ¿y esto para qué sirve?

Por tanto en este libro he evitado tratar algunos conceptos y temas que en la práctica del Psicólogo y Educador considero no tiene sentido tratarlos (salvo que éste se dedique a la investigación), entre ellos están: cálculo de Probabilidades, estudio de la Varianza y los capítulos referentes a la Estadística no Paramétrica como por ejemplo: pruebas de signo, de Rachas, U de Mann-Whitney, Kruskal-Wallis entre otras; esto último fundamentalmente porque no corresponde normalmente a las mallas curriculares de las distintas facultades de Psicología y Ciencias de la Educación y tampoco a la constante aplicación de estudio.

Pero sí contiene y desarrolla a profundidad los temas clásicos de la Estadística Descriptiva y de la Inferencial proponiendo su aplicación a temas que tienen que ver tanto con la profesión del Psicólogo en cualquiera de sus tres áreas de acción: Psicología Clínica, Psicología Organizacional y Psicología Educativa como en el quehacer del Educador a cualquier nivel: Primario, Secundario o Universitario.

LA OBSERVACIÓN COMO ELEMENTO FUNDAMENTAL DEL ANÁLISIS

Como base para cualquier proceso de investigación, siempre he sugerido que la persona debe tener un buen entrenamiento para desarrollar su capacidad de observación; pero pudiendo esto ser solo una opinión de mi parte, he buscado si algunos personajes dentro del ámbito educativo y otros opinan lo mismo, es por ello que presento a continuación lo que algunos pensadores opinan y recomiendan al respecto.

“La percepción de las cosas se distorsiona cuando no tienes el enfoque ideal de lo que observas. Es importante ver cada ángulo para poder dar una mejor opinión de las cosas, no dejarnos guiar por la primera impresión”, Hishee Salgado Morán (“Frases Célebres aplicadas a la Percepción,” n.d.).

“No podemos crear observadores diciendo “observar”, pero dándoles el poder y los medios para esta observación, y estos medios son adquiridos a través de la educación de los sentidos”, María Montessori (“frases sobres observacion - Buscar con Google,” n.d.).

“La observación indica cómo está el paciente; la reflexión indica qué hay que hacer; la destreza práctica indica cómo hay que hacerlo. La formación y la experiencia son necesarias para saber cómo observar y que observar; cómo pensar y qué pensar”, Florence Nightingale (“frases sobres observacion - Buscar con Google,” n.d.).

“Se puede afirmar, sin exagerar, que la observación y la búsqueda de similitudes y diferencias son la base de todo conocimiento humano”, Alfred Nobel (“frases sobres observacion - Buscar con Google,” n.d.)

Las frases de estos pensadores reflejan la gran necesidad de desarrollar capacidades de observación en todos los ámbitos profesionales y en lo referente al Educador y al Psicólogo, considero que es un tema de muchísima importancia, ya que de una buena y objetiva observación se podrá obtener valiosa información que le permitirá tener una visión más amplia de alguna problemática a tratar.

Los procesos estadísticos tienen dos elementos que considero fundamentales para el análisis de variables:

- i. La capacidad de observación
- ii. El análisis numérico de los datos

Respecto al primer punto debo decir que esto no es cuestión de “ver” como ocurren las cosas, realmente va más allá ya que se debe tener un juicio crítico muy agudo y esto permitirá ser muy objetivo; estos atributos la persona los puede ir desarrollando con base a diario entrenamiento y paciencia; el sistema podría llamarse tipo “espía” ya que las personas observadas no deben darse cuenta de que son sujetos de estudio.

Surgiría entonces una pregunta: ¿qué observar?, en términos del objetivo de este libro la respuesta sería: actitudes, conductas que todos tenemos en situaciones de trabajo, estudio, familia, amigos, en fin, en todos los ámbitos que cada persona se desenvuelve.

Por ejemplo, el desarrollar esta capacidad de observación permitirá al Psicólogo Educativo conocer más y mejor a los estudiantes especialmente en las relaciones que mantiene con sus compañeros, pero lo debe hacer tanto en los momentos de esparcimiento como en los estadios formales, esto ayudará a darse una imagen más real y obviamente complementaria a las entrevistas y datos adicionales que pueda recabar.

En cuanto al Psicólogo Clínico esto le será un poco más complejo ya que difícilmente tendrá acceso a situaciones específicas de sus pacientes, pero la observación puede darse en cada consulta y dentro de ella el profesional debe desarrollar la capacidad de observación del lenguaje corporal.

El caso del Psicólogo Organizacional tiene mucha similitud con el Educativo ya que él se encuentra en el mismo medio de trabajo de los colaboradores y, por tanto, tiene tiempo para su observación. Una recomendación es utilizar el sistema de “administración por recorrido”, que en resumen no es otra cosa que visitar las distintas dependencias de la organización en un sistema de visita informal y aleatoria sin necesidad de tener un tema específico a tratar.

Para el educador la capacidad de observar es básica y sumamente necesaria a cualquier nivel (primario, secundario o superior), considero que los dos primeros tienen un peso mucho más fuerte ya que permitirá obtener valiosa información que permitirá un acercamiento a la realidad de ciertas problemáticas tanto individuales como grupales.

¿PARA QUÉ SIRVE LA ESTADÍSTICA EN CIENCIAS SOCIALES Y PARTICULARMENTE EN PSICOLOGÍA O EDUCACIÓN?

Una de las quejas de los estudiantes de grado y posgrado (y aunque no se crea también de profesionales) es que la Estadística no les ayuda en nada para su profesión; a los estudiantes les entiendo porque todavía no conocen el alcance real y la necesidad de análisis numérico en la Psicología, pero ¿a los profesionales? Y qué decir de los educadores, lastimosamente el poco apego que muchos tienen a los números les venda los ojos y esto no les permite utilizar los datos para un mejor quehacer educativo.

Como he dicho anteriormente es fundamental que la visión que deba darse al estudio de la Estadística en cada rama de las Ciencias Sociales, dependerá del profesional que la aborde y aplique y sobre todo anime a los estudiantes a ver que sí tiene influencia real en su formación profesional.

Pero se debe tener en cuenta la diferencia que realmente hay según el nivel de estudio, ya que es importante delimitar la temática y, sobre todo, la aplicación dependiendo del nivel (grado o posgrado).

Por ejemplo, a nivel de posgrado sí será necesario profundizar un poco más en temas como el análisis de varianza, tipos de distribución como la *chi* cuadrada, distribución F, pruebas “t” y otras e inclusive si es necesario el estudio de series temporales; temas que a nivel de grado no son necesarios por la poca aplicación que pueda observar el estudiante ya que sus conocimientos dentro de la profesión no son lo suficientemente amplios aún y en algunas ramas sociales no son muy aplicables; pero debo también insistir en que estos últimos temas expuestos sí tienen aplicación práctica en temas de investigación.

Por tanto, la Estadística tendrá sentido (servirá) en la medida del conocimiento y aplicación a la ciencia de estudio; por ejemplo en Psicología es obvio que la aplicación es inmediata en temas de Psicometría, Psicología Social, Evaluación entre otras; en ciencias como la Sociología tendrá aplicación por ejemplo para el levantamiento de información y estudio de poblaciones marginales, la Historia se beneficia de la Estadística para relacionar hechos históricos e inclusive para entender (¿traducir?) ciertos escritos en lenguas antiguas, aquí por ejemplo sí se utilizan temas como los de probabilidad y frecuencia; en cuanto a temas de Ciencias de la Educación en el análisis de los resultados académicos la Estadística es una herramienta fundamental y que da muchas luces si se hacen estudios rigurosos sin necesidad de tener conocimientos muy profundos en esta ciencia.

¿EN QUÉ MOMENTO DEBE INICIARSE EL ESTUDIO DE LA ESTADÍSTICA EN PSICOLOGÍA?

Esta es LA pregunta. Dado que en los primeros niveles el estudiante aún no tiene conocimientos referentes a su carrera, no es conveniente iniciar el estudio de Estadística ya que no podrá apreciar la aplicación que tiene esta materia en la Psicología o Pedagogía; si se coloca a mitad de la malla curricular paralelamente coincidirá con materias que ya necesitarían de ciertos conocimientos estadísticos para su desarrollo, tal vez la mejor opción sería al final de la carrera ya que tiene las siguientes ventajas:

1. Tendrá los conocimientos suficientes para entender la aplicación a su especialidad.
2. La Estadística le ayudará en el proceso de investigación de su trabajo final de grado.
3. Podrá relacionar la aplicación práctica y profesional de esta herramienta con bases teóricas referentes a su profesión inmediata.

Pero esta pregunta y los razonamientos argüidos aquí intentando responderla, deberá ser parte de un profundo análisis en base a las necesidades reales que se establezcan en la malla curricular de las distintas carreras y que deberá ser argumentada y consensuada con los profesores involucrados en estos procesos.

En lo que sí estoy en total desacuerdo, es que en los programas de estudio a nivel medio (colegios) introduzcan la Estadística como parte de un “relleno” con poca o ninguna aplicación para este nivel; ya que esta materia se transforma en aplicar fórmulas, encontrar resultados (respuestas aplicando fórmulas) y nada más; es decir, el objetivo es resolver ejercicios sin ningún tipo de análisis.

HERRAMIENTAS A UTILIZAR PARA EL ANÁLISIS DE DATOS

Hay muchas herramientas informáticas que ayudan a realizar los cálculos que se necesitan y también para graficar resultados, entre ellas tenemos dos que se han vuelto clásicas para estos fines: Excel y el SPSS (paquete estadístico que es el más utilizado para ciencias sociales).

Pero (como dijera unas líneas atrás) voy a poner a consideración las ventajas que tiene el Excel frente al SPSS especialmente tomando en cuenta el perfil que normalmente tienen los estudiantes de carreras sociales, entre las más destacadas tenemos:

1. Todo estudiante tiene acceso al office y por ende al Excel, no así al SPSS.
2. El manejo de Excel en general es mucho más sencillo y en los establecimientos de educación media ya reciben nociones sobre el manejo de esta herramienta.
3. En cuanto a gráficos, el SPSS es muy limitado no así el Excel.
4. Para utilizar el SPSS hay que tener un mejor conocimiento de los procesos Matemáticos de la Estadística.

5. El SPSS expone resultados de pruebas estadísticas que no siempre son estudiadas en las carreras de Psicología y Educación (u otras de las ciencias sociales) y por tanto pueden confundir.
6. Aunque mecánicamente el SPSS pueda con el tiempo ser fácilmente manejable, su estructura física y su presentación en pantalla no es muy amigable para quien no está familiarizado con él.
7. Todos los temas que se manejan en los programas académicos referentes a Estadística, se pueden tratar muy bien con Excel y, como he venido diciendo, la profundidad y aplicación a las ciencias sociales dependerá del profesor.

CAPÍTULO 1:

ANTECEDENTES

LA ESTADÍSTICA COMO CIENCIA

“No te dejes engañar, el sentido común es demasiado común para ser realmente sentido, en el fondo no es más que un capítulo de la Estadística, y el más vulgarizado de todos”.

José Saramago

“No te conviertas en un mero registrador de hechos, intenta penetrar en el misterio de su origen”

Ivan Pavlov

BREVE HISTORIA DE LA ESTADÍSTICA

Al revisar algunos libros y direcciones electrónicas sobre la historia de esta Ciencia, he encontrado las más variadas y diversas “versiones” de las que uno puede imaginarse sobre los pasos que se han dado para llegar a la actualidad. Cada autor que incursiona en el interesante devenir de la Estadística aporta datos sugestivos, por tanto he realizado mi propia recopilación de hechos considerados para mí como lo más destacable en cuanto al desarrollo histórico de esta ciencia.

La palabra Estadística procede del vocablo “Estado”, ya que una de las preocupaciones de los gobiernos en general radica en conocer datos sobre condiciones materiales, nacimientos, muertes, producción agrícola, ingresos, impuestos, etc.

Como cualquier otra ciencia, la Estadística ha tenido que recorrer un largo camino a través de procesos de desarrollo y evolución; a partir del simple conteo de animales, riquezas, habitantes de la tribu entre otros.

Se conoce a través de crónicas escritas por algunos historiadores y por datos arqueológicos que en tiempos remotos y por el interés de algunos emperadores era de interés conocer

por ejemplo datos sobre las riquezas del imperio, el número de soldados y otros recursos que podían servir tanto para su interés personal como para la defensa de sus territorios (Spiegel, 1970).

En la Biblia también se puede encontrar un curioso dato sobre esto ya que al parecer Moisés realizó un censo después de salir de Egipto, observamos en uno de los libros del Pentateuco, bajo el nombre de Números, el censo que realizó Moisés después de la salida de Egipto. Textualmente dice: “1:1 Habló Jehová a Moisés en el desierto de Sinaí, en el tabernáculo de reunión, en el día primero del mes segundo, en el segundo año de su salida de la tierra de Egipto, diciendo:

1:2 Tomad el censo de toda la congregación de los hijos de Israel por sus familias, por las casas de sus padres, con la cuenta de los nombres, todos los varones por sus cabezas.

1:3 De veinte años arriba, todos los que pueden salir a la guerra en Israel, los contaréis tú y Aarón por sus ejércitos (Biblia Reina Valera 1960 (RVR 1960), n.d., p. Núm. 1.1-3).

También se puede encontrar datos históricos muy interesantes tanto en Grecia como en Roma referentes al desarrollo y aplicación de procesos estadísticos, en estas dos importantes culturas se realizaron observaciones estadísticas en lo que refiere a distribución de terreno, servicio militar, entre otros.

Roma también puso su “grano de arena” ya que debió utilizar esta herramienta dada su compleja organización política, jurídica y administrativa. Una evidencia de esto es que cada 5 años realizaba un censo para conocer no solo datos sobre sus habitantes en cuanto a nacidos y fallecimientos y también en lo referente a bienes de cada familia.

Pero fue bajo Antoninos que la declaración de nacimientos adquirió una verdadera institución legal que era necesaria hacerla ante el “prefecto del Erario” en el templo de Saturno y no después de 30 días de nacimiento. Con la caída del Imperio Romano las estadísticas se pierden en Europa, floreciendo más bajo la civilización árabe (“Estadística para todos,” n.d.).

A todo esto se debe sumar el aporte de estudiosos y políticos en Francia, Italia con Carlo Magno, en Inglaterra Guillermo el Conquistador mandó a realizar una especie de catastro, que constituye un documento estadístico administrativo, en Alemania los aportes de Vito Seckendorff y sobre todo de Hermann Conring son muy importantes y especialmente a este último se le atribuye ser el fundador de la Estadística, Conring tuvo un discípulo (Godofredo Achenwall) que siguió muy de cerca sus pasos y fue quien consolidó definitivamente los postulados de esta nueva ciencia y también de haberle dado el nombre de “Estadística” (Noronha, n.d.).

Aunque todos estos datos históricos indican la evolución de esta ciencia en las distintas culturas y épocas, en realidad no había un solo camino, ni específico interés ni acuerdos

sobre el verdadero alcance que la Estadística estaría por llegar a ser muchos años más tarde; de todas maneras cada cultura dio su aporte según el avance de las complejas transacciones comerciales y del crecimiento de los estados, que cada vez veían la necesidad de encontrar una herramienta que permita cuantificar el avance en cada proceso.

Y es por ello que:

La Estadística pasó así a ser la descripción cuantitativa de las cosas notables de un estado. Von Scholer separó la teoría de la estadística de la aplicación práctica de la misma. Todos ellos formaron parte de la tendencia de la Estadística Universitaria Alemana, conocida como la Estadística Descriptiva (“Estadística para todos,” n.d.).

DEFINICIONES DE ESTADÍSTICA

Conceptualmente hay muchas definiciones, entre ellas se pueden citar las siguientes:

“Estadística es la ciencia que se ocupa de la ordenación y análisis de datos procedentes de muestras, y de la realización de inferencias acerca de las poblaciones de las que éstas proceden.” (Botella, León, & San Martín, 1997, p. 20).

Acoto sobre esta definición dos cosas:

1. El tema de ordenar los datos es bueno, pero no muy importante debido a la utilización de herramientas informáticas, y
2. Una palabra clave sí es “análisis” ya que a mi juicio es lo más importante del quehacer Estadístico, he sostenido en varios ámbitos que la Estadística en sí no es difícil como muchos lo aseguran, las herramientas tecnológicas con las que hoy se cuenta eliminan el muchas veces tedioso proceso que fue causa de dolores de cabeza en décadas pasadas. Lo que considero resulta complejo para muchos es el qué hacer y decir con los resultados obtenido, esa es una capacidad a desarrollar en el investigador que utiliza la Estadística como herramienta, reconozco que no es fácil y que requiere de una capacidad de juicio, en base a los números, que lastimosamente no es muy sencilla de adquirir.

La Estadística entendida en el ámbito de las ciencias de la salud y, por ende, en la Psicología, es un instrumento que nos permite organizar, ordenar, resumir e interpretar información derivada del estudio de fenómenos propios de la Psicología y tomar decisiones a partir de evidencias y criterios empíricos derivados de una serie de datos medidos (Guardia, Freixa, Pero, & Turbany, 2008, p. 2).

Respecto a la definición anterior debo rescatar lo siguiente: concuerdo que la Estadística es un instrumento y que permite como elemento fundamental interpretar la información que nos ayude a tomar decisiones y la palabra clave tal vez sea “interpretar”.

La teoría y el método de analizar datos cuantitativos obtenidos de muestras de observaciones, para estudiar y comparar fuentes de variancia de fenómenos, ayudar a tomar decisiones sobre aceptar o rechazar relaciones hipotéticas entre los fenómenos y ayudar a hacer inferencias fidedignas de observaciones empíricas (p.192) (Kerlinger (1985), citado en Pérez Juste, 2012, p. 2).

De la definición anterior hay algunas puntualizaciones clave a resaltar como son: “aceptar o rechazar relaciones hipotéticas” y “hacer inferencias fidedignas de observaciones empíricas”, esto refleja uno de los objetivos básicos de esta ciencia.

He citado algunas definiciones y las he colocado en orden cronológico para que el lector compare y según su criterio vea si considera ha existido o no una evolución en los conceptos de esta ciencia.

Se podría decir entonces que, sin que se tome esto como una definición específica, la Estadística es la ciencia que expresa en números los resultados de la observación objetiva del comportamiento de alguna variable a analizar, y con ellos trata de interpretar situaciones reales de dicho evento y como consecuencia propone conclusiones y sobre todo recomendaciones sobre el hecho investigado.

DIVISIÓN DE LA ESTADÍSTICA

La Estadística se divide en dos grandes estudios

1. Estadística Descriptiva: Su principal aporte es explicar lo que ocurre en un determinado grupo de estudio, puede valerse de tablas y gráficos que le ayuden a comprender mejor.
2. Estadística Inferencial: Su propósito es establecer inferencias o predicciones en base a los datos obtenidos de una población; permite ya relacionar variables y ayuda a tomar decisiones más específicas.

Para el estudio de cualquiera de ellas hay que aclarar que el proceso se basa en tomar cierta cantidad de datos que proporcionen información veraz y suficiente para analizar y darse una idea lo más cercana posible a la realidad específica de un grupo de estudio.

CONCEPTOS INICIALES

“El muestreo no es la simple sustitución de una cobertura parcial para una cobertura total. Muestreo es la ciencia y el arte de controlar y medir la fiabilidad de la información estadística útil a través de la teoría de la probabilidad”.

Deming (1950)

Población y muestra

Pero ¿cuál es esta realidad?, técnicamente la llamamos **Población** (universo) que será en definitiva la razón de nuestro estudio y para su análisis generalmente debemos tomar un cierto número de valores, que le llamaremos **Muestra** de tal manera que esta última cumpla con una condición fundamental: **debe garantizar representatividad**, por tanto tiene que reunir las **mismas características** de la población.

Debemos tomar en cuenta lo siguiente: todo subconjunto de la población es una muestra, pero no toda muestra es representativa, por tanto, para garantizar que el estudio tenga validez, se debe cumplir con la condición antes establecida; esto se trata de ejemplificar en la Figura 1.

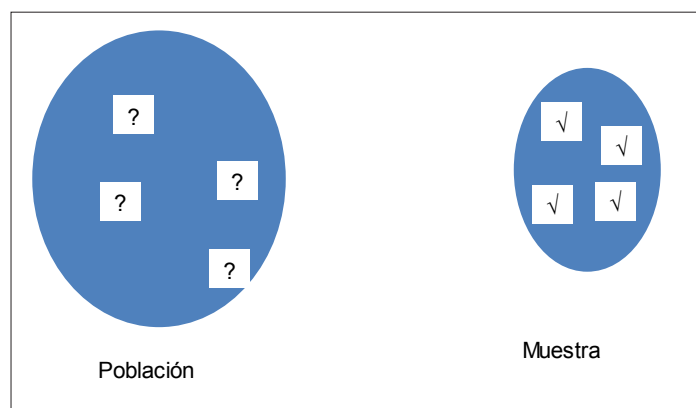


FIGURA 1: LA MUESTRA COMO UN SUBCONJUNTO DE LA POBLACIÓN

Para efectos de proceso, la idea es la siguiente:

El investigador se hace preguntas sobre las características y comportamiento que puede tener una determinada población (objeto de estudio), para ello obtiene una muestra **representativa** y en ella realiza los estudios, obtiene resultados y hace conclusiones que permitan establecer, con determinada seguridad, que lo encontrado en la muestra ocurre también en la población.

Población

Es importante entonces establecer la población a la cual se va a dirigir un determinado estudio y para ello se debe delimitar con la mayor precisión posible cuál sería este grupo objetivo, es decir a qué conglomerado de personas va dirigida específicamente la investigación a realizar.

Para ello pongo los siguientes ejemplos comparativos sobre la identificación de la población objetivo:

1. Estudio del síndrome de Down.
2. Estudio del síndrome de Down en niños.
3. Estudio del síndrome de Down en niños de 4 a 8 años.
4. Estudio del síndrome de Down en niños de 4 a 8 años en establecimientos educativos.
5. Estudio del síndrome de Down en niños de 4 a 8 años en establecimientos educativos particulares.
6. Estudio del síndrome de Down en niños de 4 a 8 años en establecimientos educativos particulares de la zona norte.
7. Estudio del síndrome de Down en niños de 4 a 8 años en establecimientos educativos particulares de la zona norte de la ciudad de Quito.

Fíjese que en cada redacción se ha ido puntualizando la población objetivo, por ejemplo en el punto uno no se establece nada de hacia qué población va dirigido el estudio y en los sucesivos puntos se va aclarando cada vez mejor y determinando con mayor precisión la población en la que se va a realizar la investigación y a pesar de ello en algún momento del proceso el investigador debería establecer geográficamente cuál es esa “zona norte” (numeral 7).

Ahora bien, si ya se ha superado de la mejor manera este primer paso, ¿cuál es entonces la forma de obtener esos datos de la población que nos permita realizar un análisis fiable de ella? Para esto debemos referirnos a un segundo concepto muy importante en el ámbito de la Estadística, esto es la muestra y sus distintos métodos de encontrarla.

Muestra

Como se indicara en párrafos anteriores, la muestra es un subconjunto representativo de la población, esto significa que este grupo de datos deberá asegurar tener las mismas características del objetivo de estudio; por ejemplo si la población está conformada por hombres y mujeres, niños, adolescentes, personas mayores, etc., la muestra deberá reunir también a personas que representen a cada grupo, además, tomando en cuenta el mismo ejemplo, se deberá establecer la proporción de dichos elementos, es decir si hay un número muy grande de niños, esto deberá reflejarse proporcionalmente en la muestra.

Esta propiedad (la representatividad) es el requisito más importante que debe cumplir la muestra en referencia al universo de estudio.

Muestreo

¿Por qué tomar muestras? La principal razón es que normalmente los estudios e investigaciones se realizan en poblaciones muy grandes y por tanto es prácticamente imposible obtener datos de todos sus elementos.

El muestreo consiste en recopilar información confiable de una población utilizando técnicas que garanticen esta condición.

Para ello deben cumplirse requisitos como el número adecuado de datos, la idoneidad en la recopilación de los datos, garantizar que, para obtener los datos, no se haya realizado manipulación alguna, entre otros.

El establecer el número adecuado de datos depende de varios aspectos, entre ellos tenemos:

- i. Cuánto margen de error se permite.
- ii. Qué nivel de confianza se desea para el análisis (este tema se profundizará más adelante cuando se aborden temas referentes a la Estadística Inferencial).

iii. Si se considera que la población es finita o infinita.

Hay varias fórmulas a utilizar según el caso; las más utilizadas son las siguientes:

- Para poblaciones infinitas (más de 100.000 elementos):

$$n = \frac{z^2 * p * q}{e^2}$$

- Para poblaciones finitas (menos de 100.000 elementos, por tanto, es la fórmula más utilizada):

$$n = \frac{z^2 * p * q * N}{e^2(N - 1) + z^2 * p * q}$$

Explicación:

n = Número de elementos de la muestra a obtener

N = Número de elementos de la población

p y q = representan la probabilidad de que se presente un evento o también se dice que es la proporción en la que ocurre un evento y se ha establecido de antemano.

Para los valores de “ p ” y “ q ” se suele determinar igualdad, es decir: $p = q = 0.5$ esto es porque muy pocas veces se conoce la verdadera proporción en que ocurre un evento.

Z = Valor (en tablas) correspondiente al nivel de confianza elegido. Suelen hacerse los análisis al 95% de confiabilidad. (El tema correspondiente al valor “ z ” se tratará en otro capítulo más adelante)

e = Margen de error permitido (imprecisión)

NOTA: los valores de “ e ” varían en función del criterio de quien realice la investigación, pero lo que más se aplica es un valor no mayor al 5% ($e = 0.05$); y aunque puede ser más, no

conviene ya que el error será mayor entre lo establecido en la muestra y lo que ocurra en la población.

Para la obtención de la muestra en sí, existen varias técnicas que para el objetivo de este libro vamos a revisar tres de ellas:

Muestreo Aleatorio simple

Consiste en lo siguiente:

1. La población debe estar numerada, si no lo está habría que hacerlo, en caso de que no sea posible, se recomienda utilizar otro tipo de técnica.
2. Utilizando una calculadora o con ayuda del Excel se procede a la obtención de los valores

Si es con la calculadora, debe buscarse la función “Ran#”

Si es con el Excel debe identificarse la función “aleatorio” o “aleatorio.entre” (se sugiere acceder a la siguiente dirección donde se explica con detalle la mejor manera de obtener la muestra) <https://www.youtube.com/watch?v=Movj5ujvSWM>

Principal ventaja de este sistema de muestreo: no es posible manipular la obtención de los datos.

Desventajas:

- i. la población no siempre estará numerada o se podrá numerar con facilidad.
- ii. Cuando se hace el proceso por cualquiera de los medios electrónicos indicados, se debe tener cuidado en que no se repitan elementos (esto para el caso de personas).

Muestreo Sistemático

Consiste en lo siguiente:

1. Hay que establecer un “sistema” para obtener la muestra, dicho sistema puede realizarse de varias maneras, entre otras:
 - a1. Tiempo: por ejemplo, se puede escoger a un elemento de la población cada “x” minutos.
 - b. Espacio: por ejemplo, puede obtener la muestra cada 100 metros, esto se puede aplicar si se hace una investigación de campo cuya población está muy distanciada.
 - c. Numérico: para esto la población debe estar numerada (o se la puede numerar) y se obtiene la muestra por ejemplo escogiendo a los múltiplos de “x” cantidad.
 - d. Posición: se puede escoger a personas sentadas a tres puestos unas de otras.

Ventaja: es rápida y fácil de aplicar debido a la variedad de sistemas a usarse.

Desventajas:

- i. A veces debe numerarse y por tanto tendría la misma dificultad que el método anterior
- ii. No siempre se puede cumplir con el sistema escogido, por ejemplo, si se establece escoger a una persona cada 2 minutos, tal vez nadie se presente en este tiempo.

iii. Puede manipularse el escogimiento de los datos y por tanto no se garantiza su imparcialidad.

Respecto a este método debo indicar que, si bien es cierto lo dicho en el literal iii, esto algunas veces puede ser utilizado con prudencia para lograr ciertos objetivos. Suelo recomendar para temas pedagógicos y de terapia grupal casos como el siguiente:

Suponga que está trabajando con un grupo de personas y ha decidido sentarse en forma circular (no es fundamental) y que durante el proceso nota que una o dos o más personas no colaboran ni participan activamente aportando al objetivo del trabajo.

La sugerencia para “obligar” a su participación sería que enumere al grupo y observe qué números le tocaron a cada participante que no haya aportado, observe la Figura 2.

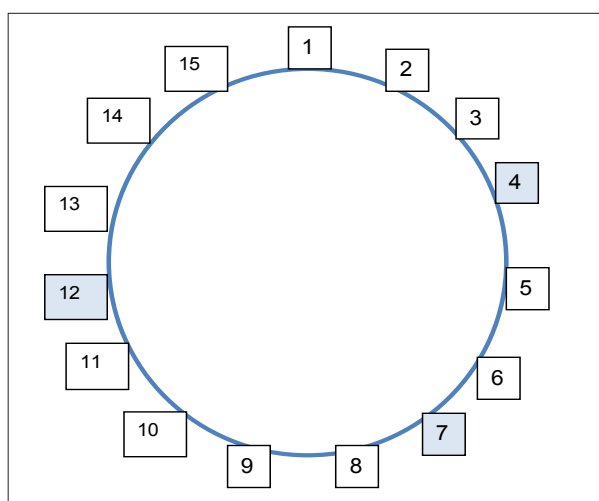


FIGURA 2: REPRESENTACIÓN DE UN SISTEMA DE MUESTREO SISTEMÁTICO

Supongamos que los números 4, 7 y 12 corresponden a esas personas, puede decir “los múltiplos de 4 (sistemático numérico) nos van a dar su opinión”, con esto al menos ya ha logrado que de los tres, dos de ellos participen y de una forma simulada dado que fue “al azar”; para el otro participante en otro momento puede pedir que participen los que correspondan a números primos.

Muestreo Estratificado

Este sistema se debe utilizar cuando la población está dividida en estratos (grupos) plenamente identificados, es decir ningún elemento puede pertenecer a dos grupos al mismo tiempo, por ejemplo, por segmentos de edad (valores enteros), ingresos económicos, cociente intelectual, resultados de evaluación entre otros.

Para ello se debe establecer la proporción de cada estrato respecto a la población y luego de conocer el tamaño de la muestra, escoger esa misma proporción de cada grupo, por ejemplo:

La variable a estudiar es la edad y la población tiene 6 grupos bien definidos, si el tamaño de la muestra a escoger es de 140 personas (aplicando la fórmula para poblaciones finitas), entonces estará compuesta según se muestra en la Tabla 1:

EDAD	FRECUENCIA	PROPORCIÓN	MUESTRA
10 – 15	12	5,48%	8
16 – 21	25	11,42%	16
22 – 27	28	12,79%	18
28 – 33	39	17,81%	25
34 – 39	50	22,83%	32
40 – 45	65	29,68%	42
219			141
Población			Muestra

TABLA 1: PROPORCIONALIDAD DE CADA INTERVALO Y NÚMERO DE PERSONAS A ESCOGER EN CADA GRUPO

Los valores en la columna de “MUESTRA” se obtienen multiplicando 140 (tamaño de la muestra) por cada uno de los valores de la columna “PROPORCIÓN” utilizando en Excel la fórmula: “+REDONDEAR” con cero decimales; aunque según la fórmula el “tamaño” de la muestra debía ser 140, por efectos de redondeo el total resulta ser de 141 para este caso.

Ventaja: es un sistema rápido de obtención y muy justo en cuanto a la representatividad.

Desventaja: con este sistema se conoce sólo el número de elementos de cada grupo, hay que utilizar una técnica adicional para determinar a quiénes se escogerá dentro de ellos.

Problemas del muestreo

¿Le han aplicado a usted alguna vez una encuesta? ¿Ha respondido con total veracidad?, si la respuesta a la segunda pregunta es “no”, entonces comprenderá que no será la única persona que haya respondido así (imagínese si son muchos).

¿Qué significa esto y qué consecuencias tiene? para usted de hecho no tendrá ninguna implicación, pero para la persona a la que usted respondió y para el objetivo de la investigación que se esté realizando sí tendrá consecuencias dado que los resultados no reflejarán auténticamente la realidad de la problemática a estudiar y por tanto no serán muy fiables sin interesar la técnica utilizada.

Casos como este se presentan según el objeto de estudio, es decir si en el tema sobre el cual se investiga hay preguntas que el entrevistado puede considerar comprometedoras y por tanto prefiere dar respuestas que no le perjudiquen o que sean vagas.

Preguntas como la edad, el ingreso familiar, la afiliación política y otras, muchas veces no son veraces; en otros casos no siempre se conoce en realidad sobre lo indagado y se dan datos que tampoco son ciertos por ejemplo el peso corporal, el gasto semanal en alimentación, el número de horas que le atendió el terapeuta; todo esto hace que la calidad de los datos

recabados en un muestreo no dependan del sistema utilizado y por tanto haya cierto escepticismo en los resultados que arroje la investigación, es por ello que en los procesos estadísticos debe determinarse un cierto porcentaje admisible de error.

Entre otras razones lo expuesto anteriormente es un ejemplo de las dificultades que se presentan al obtener datos en una muestra, por tanto, en un proceso de investigación debe tenerse muy en cuenta lo siguiente:

- i. Determinar de manera muy específica la población a estudiar.
- ii. Establecer qué tipo de muestreo es el más conveniente o si es necesario aplicar varios procesos para obtener la muestra adecuada.
- iii. Decidir sobre el nivel de confianza y el error permitido para encontrar el tamaño de la muestra.
- iv. Respecto al instrumento a utilizar si es una encuesta procure hacer la menor cantidad de preguntas posible y no introduzca preguntas que no van a dar un valor agregado a la investigación. Por ejemplo, si para su objetivo no es necesario conocer el sexo o la edad, no pregunte, esto solo aumentará el volumen y no aportará a los resultados.

ESCALAS DE MEDIDA

El tratar este tema permitirá establecer las grandes diferencias que existen en la asignación de números a distintos tipos de variables y por tanto, la forma de “manejar” esos números e interpretarlos deberá mantener la particularidad en cada una de ellas; por ello deberemos tener cuidado en el manejo en cuanto a niveles o escalas de medida ya que no todas permiten – por ejemplo – realizar operaciones aritméticas que representen un proceso lógico.

Así tenemos casos en los cuales se podrán hacer operaciones básicas cuyos resultados sí serían lógicos como es el caso de establecer el promedio de edad o de estatura de un grupo de personas; pero si se asignan números (codificación de la variable) al sexo y se dice que los hombres estarán representados por el 1 y las mujeres por el 2 sería absurdo realizar una operación como el promedio o cualquier otra; así también tenemos variables intangibles que para poder establecer una medida deberán ser previamente definidas con claridad, por ejemplo tenemos la inteligencia, la autoestima u otras como el rendimiento académico que aunque es una variable mucho más “común” (por su uso) se deberá establecer bajo qué parámetros se realiza tal medición.

Medir determinados objetos de los ámbitos en que trabajamos – Educación, Economía, Medicina, Psicología, Sociología...- implica **definir** el objeto a medir, encontrar manifestaciones de tal objeto o reactivos adecuados y decidir la **regla de medida**, la regla que nos permitirá atribuir un valor a cada manifestación o reactivo (unidad de medida) (Pérez Juste, 2012, p. 15).

TIPOS DE VARIABLES Y CLASIFICACIÓN

El estudio de la Estadística requiere que se establezca previamente qué es lo que se desea analizar, es decir hay que determinar el objetivo de estudio, y a ese elemento lo llamamos **variable**, existen a su vez varios tipos, a saber:

Tipos de variable

Las variables se clasifican en Cualitativas y Cuantitativas

Variables cualitativas. Son aquellas cuyos valores no se pueden asociar a un número ya que solo expresan la cualidad.

Tienen dos escalas de medida:

Nominales: si los datos no pueden ser ordenados universalmente por alguna condición, por ejemplo: nacionalidad, sexo, grupo sanguíneo, barrio donde habita, raza, verdadero – falso. En este caso no es posible realizar operaciones. Esto no significa que no se puedan contar, es decir sí es posible decir “hay 20 personas de sexo femenino y 15 de masculino”.

Ordinales: si los datos sí pueden ser ordenados universalmente por alguna condición, por ejemplo: intensidad de dolor, nivel de trauma, nivel de felicidad, nivel de satisfacción de un tratamiento. Sus valores suelen ser codificados. Por ejemplo cuando vamos al médico con alguna dolencia, suelen preguntarnos “en una escala del 1 al 10, cuánto te duele”

Variables Cuantitativas (numéricas): Son aquellas cuyos valores sí son numéricos y pueden hacerse operaciones con ellos.

Se dividen en dos tipos:

Discretas: siempre y cuando sus valores sean exclusivamente números enteros, por ejemplo: número de pruebas aplicadas, número de hijos, número de pacientes atendidos, número de aciertos o errores. En ninguno de estos casos caben números decimales, por ejemplo no hay 15,5 pacientes atendidos.

Continuas: Si los valores pueden expresarse también con decimales o fracciones, por ejemplo: ingresos, edad, altura, peso, dosis de un medicamento, precio. En el caso de la edad, esta variable suele expresarse en valores enteros, es decir cuando nos preguntan sobre los años vividos no decimos tengo 25 años, tres meses y 18 días o tengo 25,34 años,

El resumen de lo indicado anteriormente se muestra en la Tabla 2:

TIPOS DE VARIABLE	ESCALA DE MEDIDA	CARACTERÍSTICA	EJEMPLO
Cualitativas	Nominales	No se puede establecer un orden específico aceptado universalmente	Color, gusto, género
	Ordinales	Sí se pueden ordenar	Estatura, ingreso económico, rendimiento
Cuantitativas	Discretas	La variable debe expresarse en valores enteros	Número de alumnos, número de camas en el hospital, cantidad de niños de la comunidad
	Continuas	La variable puede expresarse en valores decimales o fracción	Resultados de un test, notas de alguna materia, costo de una carrera

TABLA 2: TIPOS DE VARIABLE Y ESCALAS DE MEDIDA

En el uso de las distintas escalas también deberá tomarse en cuenta si la variación entre las medidas dentro de la variable debe interpretarse de la misma forma aún si dicha variación es numéricamente igual.

En las ciencias del comportamiento, muchas de las escalas utilizadas son tratadas con frecuencia como si fueran de intervalos sin establecer con claridad que la escala en realidad no posee intervalos iguales entre unidades adyacentes. Las mediciones de coeficiente intelectual (CI), variables emocionales como la ansiedad y la depresión, las variables de personalidad (p.e., autosuficiencia, introversión, extroversión y dominio), variables de excelencia al final del curso o de logro, variables de actitud, etc., corresponden a esta categoría. Con todas estas variables, resulta claro que las escalas no son de proporción. Por ejemplo, con el CI, si un individuo obtiene cero en la Escala Weschler de Inteligencia para adultos (mejor conocida como WAIS, por sus siglas en inglés), no concluiríamos que tiene cero en inteligencia. Es presumible que descubriéramos que dicho individuo pudo responder a algunas preguntas que quizá indicaran un CI mayor que cero. Por consiguiente, la WAIS no tiene un punto cero absoluto y las proporciones no son adecuadas. Entonces, no es correcto indicar que una persona con CI de 140 es el doble de inteligente que una persona con CI de 70 (Pagano, 2011, p. 33).

De los siguientes enunciados, clasifique como continuo o discreto:

- i. Cantidad de mujeres en la clase
- ii. Número de veces que el ratón en una caja de Skinner presiona la palanca
- iii. Edad de los participantes en un experimento
- iv. Cantidad de palabras recordadas
- v. Peso de los alimentos a ingerir
- vi. Porcentaje de estudiantes en clase mayores a 20 años

De los siguientes enunciados, clasifique como nominal u ordinal:

- i. Cantidad de bicicletas utilizadas por los alumnos
- ii. Tipo de bicicletas utilizadas por los alumnos
- iii. Dominio de la materia de Estadística entre los estudiantes en categorías de deficiente, regular y bueno
- iv. Ansiedad de hablar en público en una escala entre 1 y 100

CAPÍTULO 2:

ESTADÍSTICA DESCRIPTIVA

El uso que debe darse a la Estadística Descriptiva dentro de la Psicología y Educación es de suma importancia debido a su aplicación directa en medición de variables del comportamiento humano, La Figura 3 hace referencia a la implicación de los distintos temas descriptivos con algunas áreas dentro de la Psicología

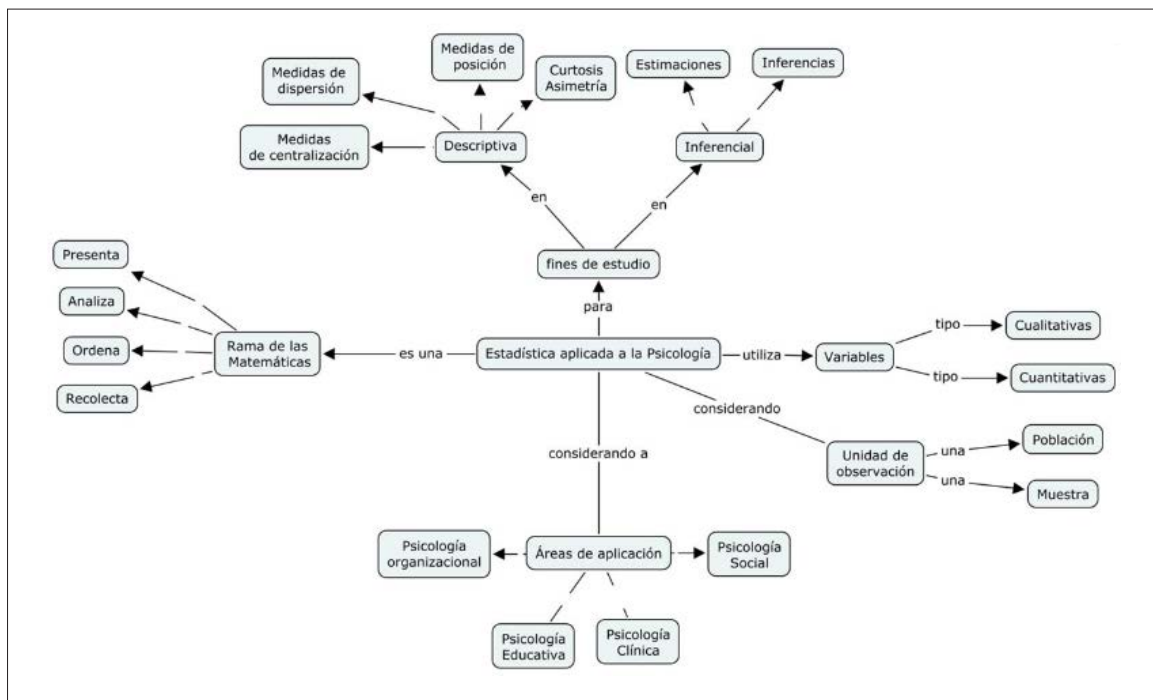


FIGURA 3: RELACIÓN DE LA ESTADÍSTICA CON LA PSICOLOGÍA (Rodríguez Monterrosa, n.d.)

Lo que se presenta en este cuadro es un buen resumen de lo que se ha desarrollado en las páginas anteriores, es clara la importancia de la observación, de la identificación del tipo de variable a estudiar, de la división general de la Estadística en Descriptiva e Inferencial, de los pasos lógicos que la Matemática exige: recolectar, ordenar, analizar y presentar los resultados tanto de manera analítica como gráfica y por último la aplicación a las distintas áreas dentro de la Psicología; aunque este cuadro-resumen no hace referencia a la aplicación en ciencias sociales en general ni en educación en particular, lo que se ha venido tratando en este libro bien se puede incluir sin afectar su esencia.

Procederemos entonces al estudio de cada una las dos grandes áreas en las que se divide la Estadística, iniciando con los temas referentes a la Descriptiva.

ESTADÍSTICA DESCRIPTIVA

La Estadística Descriptiva utiliza algunas herramientas de análisis que permiten obtener una visión general de la variable a estudiar, como su nombre lo indica trata de **describir** el comportamiento del objeto de estudio en base a varias medidas que permitirán tomar decisiones sobre lo estudiado.

En la Figura 4 se resume las partes en las que esta rama de la Estadística se divide y que las estaré tratando en las siguientes páginas.

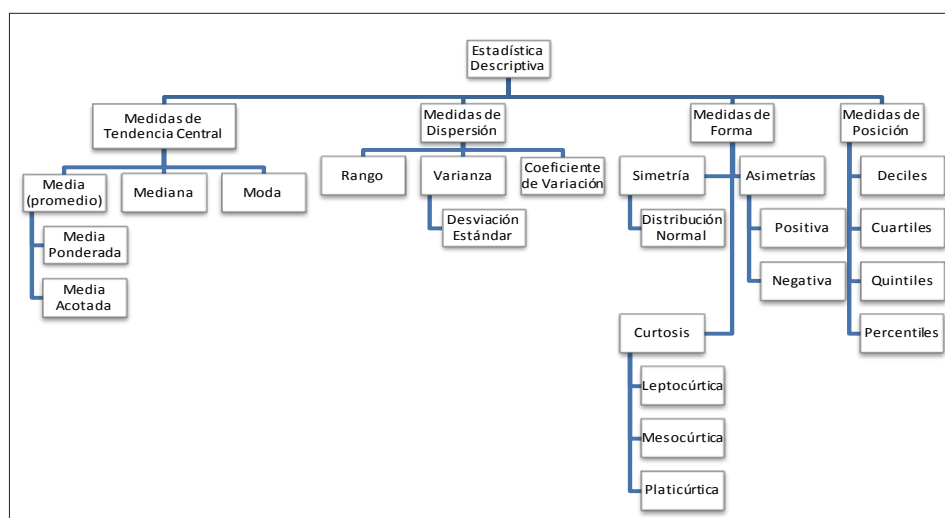


FIGURA 4: DIVISIÓN DE LA ESTADÍSTICA DESCRIPTIVA

MEDIDAS DE TENDENCIA CENTRAL

Estas medidas ayudan a “percibir” y hasta “intuir” alrededor de qué números se aglutinan los valores obtenidos en la muestra estudiada (y supondríamos también en su población), es decir son los resultados que pueden tomarse como representantes del “estado” de la variable y sobre ellos determinar el comportamiento de la esta y describir sus características.

Aunque estos datos nos dan una primera idea sobre el objeto de estudio, representan al mismo tiempo una visión algo “miope” en la evaluación que podemos tener de la muestra en particular y de la población en general, pero aun así estas medidas son muy utilizadas para dar una apreciación sobre el objeto de estudio (variable), aunque a mi juicio erróneamente si el análisis solo se basa en ellas.

Las medidas de Tendencia Central son tres: Media (promedio), Mediana y Moda; como casos especiales de la Media también se estudian otros tipos como son: Media Geométrica (utilizada para fines que salen del alcance de este libro, como son temas referentes a progresiones geométricas y de análisis financiero), Media Acotada y Media Ponderada (sobre estas dos últimas trataré en páginas posteriores).

Media (promedio)

Es la medida más utilizada en los análisis estadísticos y se la obtiene de la sumatoria de todos los datos numéricos de la muestra dividida entre el tamaño de la misma.

Característica principal: se ve afectada por valores extremos, esta afectación puede distorsionar mucho su interpretación.

Por ejemplo, si tenemos los siguientes datos:

1 1 3 5 9 11

La Media de estos valores es: 5

Pero si hacemos solo un cambio en uno de los extremos, suponiendo por ejemplo que hubo un error de digitación y faltó incluir un dato (digamos 100) entonces la muestra pasa a ser la siguiente:

1 1 3 5 9 11 100

Con este cambio, la Media pasa a ser: 18.57; como se puede notar, un valor extremo como el 100 hizo que el promedio se vea aumentado en casi cuatro veces.

Muchas críticas y hasta bromas se producen debido a esta particularidad del promedio, cito la siguiente frase de Nicanor Parra (Poeta, Matemático y Físico chileno): “Hay dos panes. Usted se come dos. Yo ninguno. Consumo promedio: un pan por persona.”

Es por esto que, al utilizar solo el valor de la Media como indicador del comportamiento de alguna variable, debe tenerse mucho cuidado al tomar decisiones en base a ella, se recomienda entonces verificar si los datos se distribuyen más o menos equitativamente alrededor del valor encontrado, sobre esto trataremos más adelante.

Mediana

Esta medida conceptualmente divide en dos partes exactamente iguales a la muestra, por tanto, indica que tomando en cuenta este dato, el 50% de la muestra tendrá valores inferiores y el otro 50% valores superiores al valor encontrado.

En el caso de esta medida, es importante señalar que los datos deben estar ordenados (preferiblemente en forma ascendente); además, para su cálculo existen dos posibilidades según el número de datos, es decir si la muestra está compuesta por un número par o impar de elementos.

Si ocurre que hay un número impar de elementos, la Mediana será simplemente aquel valor que se encuentra exactamente en la mitad (recuerde los datos deben estar ordenados, si es en Excel esto no hará falta) y para el caso contrario (número par de datos) se deberá hacer un promedio entre los dos datos centrales.

A diferencia de la Media, la Mediana se afecta muy pocas veces por valores extremos de la variable, y cuando lo hace, su valor no cambia tanto como la Media, por ello se dice que es la medida de tendencia central más estable.

Si utilizamos el ejemplo propuesto anteriormente para la Media:

1 1 3 5 9 11

En este ejemplo hay seis datos, por lo tanto, hay que obtener el promedio de los valores centrales (el 3 y el 5) por tanto la Mediana será 4

Y si de igual manera aumentamos el valor de 100, los datos se transforman en esta secuencia:

1 1 3 5 9 11 100

Y la Mediana será ahora 5 dado que es el valor central; como se puede notar, el haber aumentado un valor extremo no afectó mayormente el valor inicial, en este caso pasó de un valor inicial de 4 al nuevo valor: 5.

Pero el estudio de la Mediana en realidad no debe hacerse de una manera tan simple como la expuesta, aunque lastimosamente esto es lo que normalmente ocurre. ¿Cuál es entonces la propuesta para analizar los datos con esta medida? Voy a poner un ejemplo para establecer que la Mediana no debe interpretarse solo como un valor que divide a la muestra en dos partes iguales.

La variable a analizar es el rendimiento de la materia de Matemática en segundo curso de básica de un determinado centro educativo a través de las notas publicadas para tres paralelos distintos:

Paralelo 1					Paralelo 2				
1	1	8	1	9	1	7	2	2	3
9	2	9	9	8	4	6	5	8	6
2	1	2	8	5	9	5	7	9	5
6	2	8	1	9	3	8	2	8	8

Paralelo 3				
7	5	7	5	5
7	7	4	7	5
5	5	6	5	7
7	5	7	7	5

El valor de la Mediana para los tres casos es el mismo: 5.5. De acuerdo al concepto revisado se diría entonces que en los tres casos el 50% de las notas de Matemática están por debajo de 5.5 y el otro 50% está por encima de esa nota y nada más.

Visto así los tres paralelos tendrían la misma problemática. Pero pongamos “una lupa” a los datos de cada grupo.

Si ordenamos los datos de menor a mayor se podrá apreciar lo siguiente:

Primer grupo: las notas están “repartidas” desde el 1 hasta el 9, es decir hay notas muy variadas

Segundo grupo: la concentración de notas se distribuye entre dos extremos: [1 y 2] y [8 y 9], con tan solo dos notas intermedias: un 5 y un 6

Tercer grupo: Las notas están repartidas entre el 4 y el 7, es decir no hay notas a ninguno de los dos extremos mínimo y máximo.

Tratando de graficar esto, podríamos ver lo siguiente:

Primer grupo

1 2 2 2 3 3 4 5 5 5 6 6 7 7 8 8 8 8 9 9

Segundo grupo

1 1 1 1 1 2 2 2 2 5 6 8 8 8 8 9 9 9 9 9

Tercer grupo

4 5 5 5 5 5 5 5 5 5 6 7 7 7 7 7 7 7 7 7

La gran diferencia entonces entre el segundo y tercer grupo es la dispersión de las notas (este tema lo trataremos en otro capítulo), en el segundo grupo los valores están repartidos hacia los extremos y eso significa que hay grandes diferencias de aprovechamiento entre los alumnos; en cambio en el tercer grupo, las notas se concentran en los valores medios (entre 5 y 7) sin que haya mucha diferencia en las notas; en dos palabras esto significa que la distribución en el segundo grupo es muy heterogénea y en el tercero es homogénea.

Si este análisis no se hace con los valores de la Mediana, ésta no dará un valor agregado a la variable analizada ya que como se ve en los ejemplos un mismo valor (5.5) de la Mediana para los tres grupos no discrimina la situación de cada grupo.

Moda

Esta medida indica que existe una mayor cantidad – no necesariamente la mayoría – de elementos de la muestra con este valor.

No es posible aplicar métodos matemáticos para estudiar la Moda, además, si hay varios valores (más de dos) de igual “peso” (es decir tienen la misma frecuencia), la importancia de la Moda como medida descriptiva pasa a ser nula ya que no puede utilizarse un valor sobre otro como representante de la muestra.

Puede darse el caso de tener dos valores modales, pero si hay más se dice que no existe Moda; en este caso se deben tomar en cuenta algunas consideraciones en el estudio de la variable. Al encontrar un valor Modal, se debe tener cuidado en su interpretación ya que este puede no ser representativo en realidad.

Utilizando el ejemplo que hemos venido trabajando

1 1 3 5 9 11

En este caso el único valor que se repite es el uno, por tanto, sería el valor Modal, pero ¿consideraría usted que este dato (uno) es un buen representante de la muestra estudiada?,

es decir los valores que toma la variable ¿tienden a estar alrededor del uno? Parecería que no.

RELACIÓN ENTRE LAS TRES MEDIDAS DE TENDENCIA CENTRAL

En cuanto a la relación de las tres medidas de tendencia central, es más frecuente que, entre ellas, la Media y Mediana sean valores cercanos entre sí (no es que deba ocurrir), la Moda no necesariamente, en el ejemplo estudiado ocurre esto:

$$\text{Media } (\bar{x}) = 5$$

$$\text{Mediana (Md)} = 4$$

$$\text{Moda (Mo)} = 1$$

Respecto a lo que manifestaba líneas atrás sobre la posible “miopía” de las medidas de tendencia central, con el siguiente ejemplo trataré de demostrar esto.

Supongamos que en un centro educativo se dan las siguientes condiciones al mismo tiempo en dos paralelos distintos (asunto que es muy probable que ocurra):

Materia: Química

Nivel: segundo de bachillerato

Paralelos: A y B

Profesor: el mismo para ambos grupos

Promedio de la materia en ambos paralelos: 6.52 (se califica sobre 10 puntos)

Condiciones de aprendizaje para los dos grupos: horarios similares, igual número de estudiantes, mismo número de horas por semana, iguales condiciones ambientales dentro y fuera del aula, los estudiantes son antiguos para la institución.

Tipo de variable: cuantitativa continua

¿Cuál sería entonces la conclusión lógica? Pues no otra que tomar las mismas medidas pedagógicas para mejorar el promedio en ambos paralelos y reforzar los conocimientos.

Visto así el presente caso, el psicólogo educativo o cualquier pedagogo habrán tomado una decisión de proponer métodos de aprendizaje y refuerzo, dinámicas y acercamientos con las mismas características para los dos grupos.

¿Es esto correcto? Al parecer en principio sí, pero veamos lo siguiente:

Notas (en orden ascendente) de los paralelos A y B en la materia señalada:

PARALELO A				PARALELO B			
1	2	9	10	5	5	6	8
1	3	9	10	5	5	6	8
2	4	9	10	5	6	6	8
2	7	9	10	5	6	6	10
2	8	9	10	5	6	6	10
			10				10

Si calculamos el promedio de cada grupo nos da exactamente 6.52381 ¡para ambos casos! Y en cuanto a la Mediana y Moda los valores son:

Paralelo A: Mediana: 9; Moda: 10

Paralelo B: Mediana: 6; Moda: 6

Luego de conocer los datos de cada grupo, ¿tomaría usted entonces las mismas medidas pedagógicas y psicológicas para estos grupos? Si su respuesta es afirmativa le sugiero se mantenga expectante ya que más adelante volveré con el ejemplo para analizar con otros valores estadísticos que puedan dar luces sobre una decisión final.

Sin embargo de tener mucha información respecto a la situación de los grupos estudiados, la “miopía” que aún existe y no nos permite ver mejor las cosas está dada por una razón básica: no conocemos otras medidas que permitan quitar el velo y así ofrecer mejores conclusiones y recomendaciones.

Por lo pronto al estudiar cualquier variable se recomienda hacerlo no solo con los valores de las medidas de tendencia central, si no dando un vistazo general a todos los datos ordenándolos de menor a mayor; en este caso el gráfico de las notas de los dos paralelos se vería así:

PARALELO A															
1	1	2	2	2	2	3	4	7	8	9	9	9	9	9	10

PARALELO B															
5	5	5	5	5	5	5	6	6	6	6	6	6	6	8	10

Fíjese la “repartición” de los números, es notorio que en el paralelo “B” hay un “peso” o distribución hacia notas mayores a cinco y en el paralelo “A” hay un “equilibrio” en cuanto a la repartición (distribución) de notas.

Para iniciar el desarrollo de los ejercicios, paso a indicar la forma de calcular las medidas de tendencia central en Excel en lo que se conoce como datos simples, en este caso desarrollaré el proceso con los datos del paralelo A del ejemplo.

Excel calcula los valores de las medidas de tendencia central sin importar ni el orden de los datos ni la forma de presentación de los mismos, es decir los datos pueden presentarse en una sola columna o fila o en una matriz.

Se debe aclarar también que hay varias formas de realizar los cálculos, aquí presentaré algunas de ellas para ir familiarizándonos con los procesos.

PARALELO A			
1	2	9	10
1	3	9	10
2	4	9	10
2	7	9	10
2	8	9	10
			10

Colóquese en alguna celda vacía y en el menú principal al final de la pestaña “Inicio” encontrará el símbolo “ Σ Autosuma” y una flecha hacia abajo, allí hay varias funciones, escoja “Promedio” resalte todos los datos y aplaste la tecla “enter”. Habrá encontrado el valor: 6,52380952.

Para el cálculo de la Mediana escoja otra celda vacía y haga “click” en f_x que se encuentra sobre la letra “D” correspondiente a los nombres de las columnas, el cuadro de diálogo se presenta en la Figura 5:

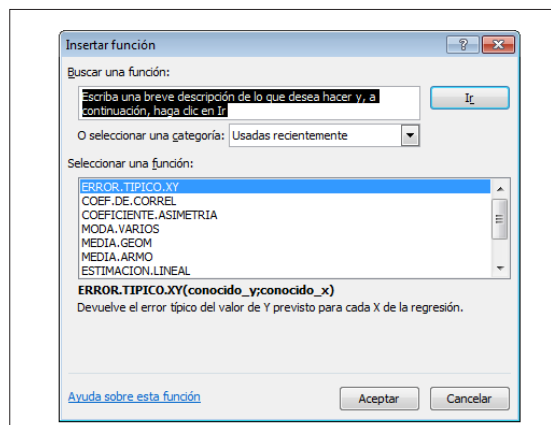


FIGURA 5: CUADRO DE DIÁLOGO PARA BUSCAR LA FUNCIÓN MEDIANA

NOTA: Lo que se vea en el cuadro de diálogo que se presente no tendrá por qué coincidir con lo que he presentado en la Figura 5

Haga click en el campo donde dice “Usadas recientemente” y cambie a la categoría “Todo”, allí busque la función: “Mediana” y acepte, la Figura 6 a continuación deberá ser lo que encuentre luego de esta acción.

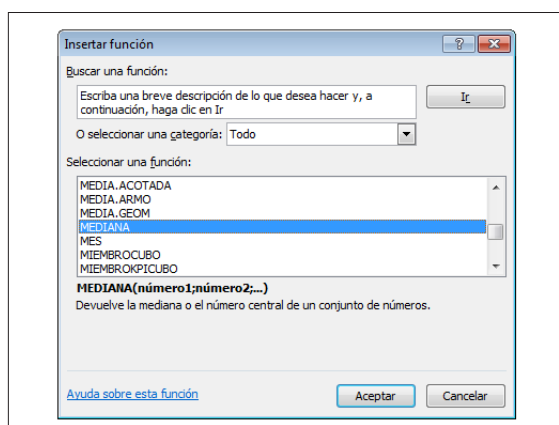


FIGURA 6: CUADRO DE DIÁLOGO DONDE SE APRECIA LA FUNCIÓN MEDIANA PARA SU CÁLCULO

Luego de aceptar (mediante un click en “Aceptar” o dando “enter”), encontrará un cuadro de diálogo como el que se presenta en la Figura 7:

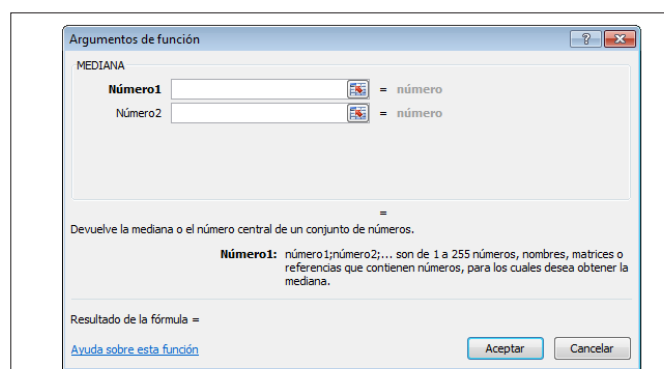


FIGURA 7: CUADRO DE DIÁLOGO QUE PERMITE ESCOGER LOS VALORES PARA CALCULAR LA MEDIANA

Resalte los datos del ejercicio y acepte, habrá obtenido 9

Para el cálculo de la Moda siga los mismos pasos anteriores buscando la función “Moda” (hay varias opciones, escoja solo la que dice “Moda”), habrá encontrado el valor 10.

EJERCICIOS DE APLICACIÓN DEL CAPÍTULO

Ejemplo 1

Los siguientes datos se refieren al tiempo (en segundos) que un grupo de personas se demoró en reaccionar a un estímulo externo. Determine la variable, encuentre las tres medidas de tendencia central, observe, comente, concluya y recomiende.

0,1	0,84	1,34	1,93	2,14	3,2	4,22	5,56
0,19	0,91	1,36	1,93	2,27	3,46	4,62	6,04
0,25	1,07	1,62	2,05	2,5	3,58	4,9	7,05
0,44	1,12	1,8	2,08	2,75	3,71	5,1	7,25
0,5	1,29	1,91	2,11	3,15	3,72	5,39	9,56

Antes de seguir adelante con el ejercicio, quisiera aclarar lo que a mi juicio debe tomarse en cuenta al analizar un caso, y es el tema de: Observar, Comentar, Concluir y Recomendar.

Estos cuatro momentos del análisis se diferencian por el nivel de complejidad, a saber:

Observar: es el estado en el que el investigador da una rápida mirada a los datos, podría decirse que “entramos fríos al análisis” y trata de darse una idea de la situación de la variable, en este punto debe tener ya una idea del comportamiento de la variable.

Comentar: es lo que yo llamaría un nivel primario, ya que luego de observar podemos decir ciertas cosas muy generales sobre los datos en función de la variable de estudio, ya hemos “calentado” para seguir con el proceso.

Concluir: luego de procesar los datos, encontrar los estadísticos correspondientes, realizar gráficos, comparar con parámetros (si los hay específicamente o por conocimiento previo de la variable) y estudiar los resultados; podríamos decir que esta etapa significa

estar ya en plena actividad dentro del análisis, podemos dar nuestro “veredicto” del problema estudiado.

Recomendar: Esta última etapa es la más importante y seguramente puede asegurarse es la razón de ser del estudio, ya que en ella se cumple con el objetivo general de la investigación, es decir, contestar el para qué se investigó. Aquí el investigador debe dar su punto de vista, por ejemplo, para solucionar el problema, dar una guía para otros casos o establecer condiciones para un futuro.

Tanto la conclusión como la recomendación deben sustentarse en los resultados encontrados

Volvamos al caso entonces.

¿Por qué es importante determinar la variable?

Recordemos que la variable es el objeto de estudio y para poder concluir o recomendar es fundamental que tengamos claro a qué se refieren los datos.

La variable en este caso sería: “reacción ante un estímulo externo medido en segundos” es decir lo que interesa medir es el tiempo de respuesta y con ello se espera concluir algo sobre el grupo de estudio y obviamente recomendar en base a los resultados.

¿Qué puede decir observando los datos? ¿Puede ya comentar algo?

Los datos ya están ordenados de menor a mayor aunque para cálculos en Excel no hace falta que esto ocurra.

Observaciones directas:

Número de datos: 40

Valor mínimo: 0.1 segundos

Valor máximo: 9.56 segundos

Primeros cálculos:

Media: 2.8753

Mediana: 2.125

Moda: 1.93

Comentario: las personas reaccionan de manera distinta ante situaciones similares.

Primera propuesta de conclusión dada por alguna persona: “el grupo es muy lento para reaccionar”, esto no significa que usted esté de acuerdo con lo expresado.

En la conclusión usted puede decir lo que a su juicio le parezca, pero cualquiera que esta sea, debe tener un sustento numérico del cual apoyarse, en este caso basado en su forma de pensar ya que no hay parámetros.

En este caso a esta persona le parece que el grupo es lento y puede apoyarse diciendo que dado el valor del promedio de reacción ante el estímulo externo (2.87 segundos), este es un tiempo muy largo, además de existir en el grupo muchos datos (son 16 de 40) que indican ser superiores a las medidas de tendencia central.

Segunda propuesta de conclusión: “el grupo reacciona en un tiempo normal”

En este caso a la persona le parece que entre 1.93 segundos (Moda, valor más bajo) y 2.87 segundos (Media valor más alto), es normal que se reaccione dado que el tiempo es corto y no hay mucha diferencia en las medidas representativas.

Tercera propuesta de conclusión: “el grupo en general reacciona de manera muy rápida ante el estímulo externo”.

En este caso la persona considera que como tan solo son 2 segundos y centésimas (tomando en cuenta Media o Mediana) ese tiempo significa mucha rapidez de reacción; además hay 15 personas que reaccionaron en menos de 1.93 segundos que es el valor más bajo en tiempo de reacción.

¿Cuál es la conclusión válida? Puede decirse que las tres tienen sus razones justificadas o ninguna de ellas (y de pronto puede haber otras); ante eso no se puede discutir dado que cada persona tendrá una forma de ver las cosas estemos o no de acuerdo.

¿Cuál es el problema en este ejercicio? ¡Faltan datos!, pues sí, ya que no sabemos por ejemplo de qué estímulo externo estamos hablando o las edades de esas personas o las características del grupo estudiado, entre otros; por tanto, cualquier conclusión es discutible.

Respecto a la recomendación, esto depende de la conclusión a la que haya llegado el investigador.

En el caso de las tres conclusiones podría decirse lo siguiente:

Recomendación 1: “realizar ejercicios de motricidad con el grupo para estimular la capacidad de reacción”

Recomendación 2: “no hacer nada con el grupo ya que los tiempos son lo esperado”

Recomendación 3: “ayudar a los miembros del grupo a que tengan mayor control de sus reacciones”.

Lo que debe ocurrir entonces es que no haya contradicción entre la conclusión y la recomendación ya que la segunda depende del criterio emitido en la primera.

Pero los análisis generalmente no se hacen con los datos tal cual nos llegan, especialmente en Psicología y procesos pedagógicos, he recomendado siempre que a los datos de la variable se los trabaje haciendo grupos. ¿Por qué? Porque en el comportamiento humano muchas veces se establecen diferencias que deben tratarse entre personas con una misma afinidad o característica lo más cercana posible.

Por ejemplo, en el caso que estamos tratando, el tiempo de reacción entre el más “rápido” (0.1 segundos) y el más “lento” (9.56 segundos) es demasiado amplio, ya que, sin importar (por lo pronto) el tipo de estímulo, la diferencia es significativa.

Este hecho de que existan todas estas conclusiones y recomendaciones (y tal vez haya más) sigue dándose por la miopía de la que he hablado ya que no tenemos otros elementos que nos permitan ver mejor; insisto, esto se solucionará con el estudio de otras medidas descriptivas más adelante.

Para un mejor estudio y especialmente en variables que impliquen comportamiento humano se recomienda hacer los análisis de dos formas:

1. Con los datos simples como se hizo en el ejemplo anterior y
2. Agrupando los datos en base a algún criterio técnico

El hacer grupos o intervalos ayudará para realizar un mejor análisis ya que se podrá “desmenuzar” a la muestra en subgrupos de estudio que tendrán características más afines entre sí y por tanto las conclusiones y especialmente recomendaciones serán más precisas y acordes a las características específicas de los grupos establecidos.

En el caso del ejemplo, para hacer grupos podemos aglutinar a las personas cuya diferencia en tiempo de reacción sea más cercana entre sí ya que esto homogeneiza a personas con características similares en lo que se refiere a impacto emocional; el proceso (en Excel) sería de la siguiente manera:

1. Se colocan los datos en columna, no importa si están ordenados o no.
2. Debe establecerse con claridad la variable a estudiar y colocarla en la celda superior (no deje espacios entre esto y los datos).
3. Colóquese sobre la celda donde nombró a la variable (por ejemplo A1)
4. Haga “click” en la ficha “Insertar” del menú principal
5. Escoja “Tabla dinámica”
6. En el cuadro de diálogo que aparece en la Figura 8 debe encontrar que Excel le indica (en fondo negro) el rango de datos con el que va a trabajar, por ejemplo: Hoja1!\$A\$1:\$A\$41; se recomienda que se acepte permitiendo que Excel cree una nueva hoja, acepte y siga al siguiente paso.

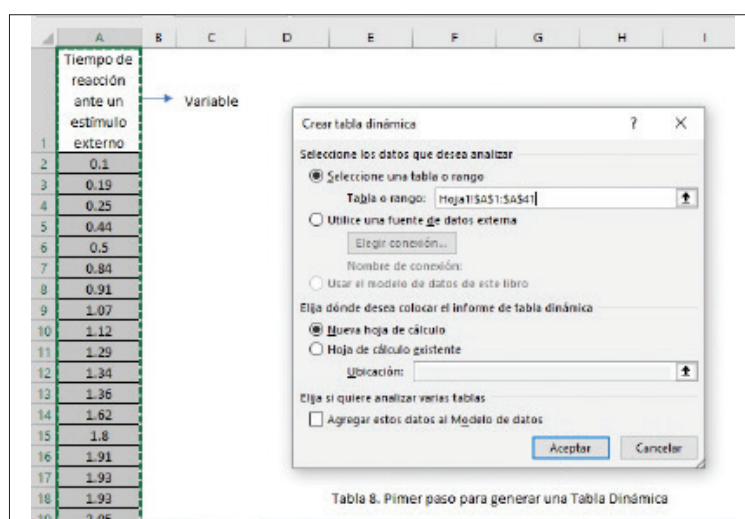


FIGURA 8: PRIMER PASO PARA GENERAR UNA TABLA DINÁMICA

7. Excel habrá creado una nueva hoja (Figura 9) en la cual podrá trabajar con una tabla dinámica.

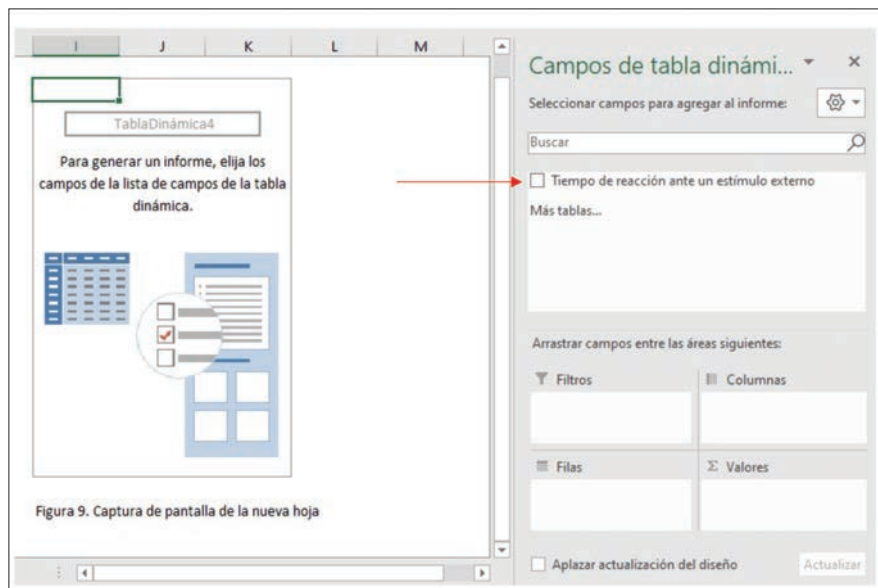


FIGURA 9: CAPTURA DE PANTALLA DE LA NUEVA HOJA

8. Hacia la derecha de dicha hoja (ver Figura 9) encontrará el nombre que dio a la variable, arrástrela hacia donde dice: “Etiquetas de fila” y suéltela allí, luego haga la misma acción pero arrastrando la variable hacia el campo: “ Σ Valores”, según se indica en la Figura 10.

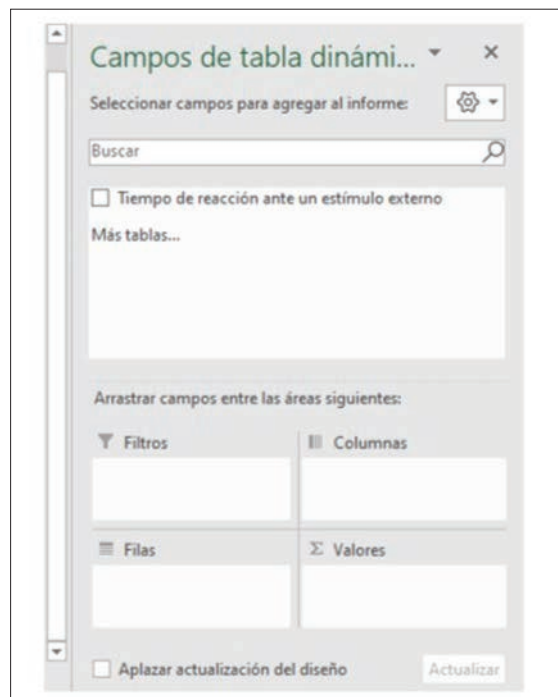


FIGURA 10: SEGUNDO PASO PARA GENERAR UNA TABLA DINÁMICA

Al hacer esto, notará que a la izquierda de la hoja se colocaron los datos de menor a mayor según se muestra en la Figura 11.

Etiquetas de fila	Suma de Tiempo de reacción ante un estímulo externo
0,1	0,1
0,19	0,19
0,25	0,25
0,44	0,44
0,5	0,5
0,84	0,84
0,91	0,91
1,07	1,07
1,12	1,12
1,29	1,29
1,34	1,34
1,36	1,36
1,62	1,62
1,8	1,8

FIGURA 11: DATOS ORDENADOS LUEGO DEL SEGUNDO PASO

9. Dependiendo de la configuración interna del Excel, al arrastrar la variable al campo “ Σ Valores” aparecerá “suma de (y el nombre de la variable)”, hay que cambiar esto para que en su lugar aparezca “Cuenta de (y el nombre de la variable)”.
- Para hacer el cambio haga “click” sobre “suma de...” y en el cuadro de diálogo que sale busque “configuración de campo de valor...” y en el nuevo cuadro escoja “cuenta” y acepte todo esto se ha representado en la Figura 12.

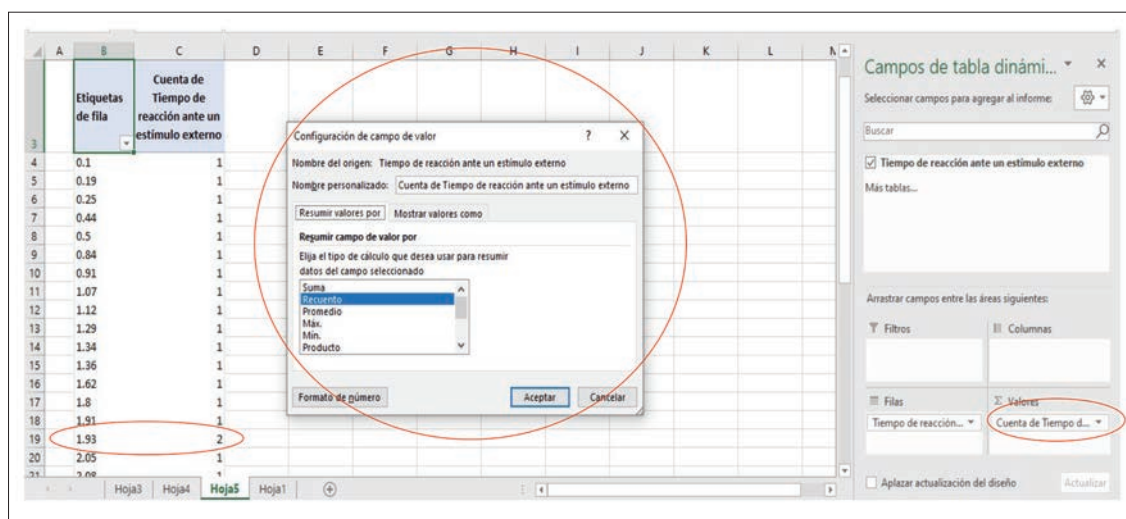


FIGURA 12: TERCER PASO PARA GENERAR UNA TABLA DINÁMICA

Con esta acción, aparecerán los datos ordenados de menor a mayor y para cada valor se ha establecido la frecuencia con la que aparece cada uno, en este ejemplo, solo el valor 1.93 le aparecerá con frecuencia “2” (ver Figura 12)

10. Haga “clik” derecho sobre cualquiera de los valores de la variable (columna izquierda) y en el cuadro de diálogo que aparece escoja “Agrupar” según se representa en la Figura 13.

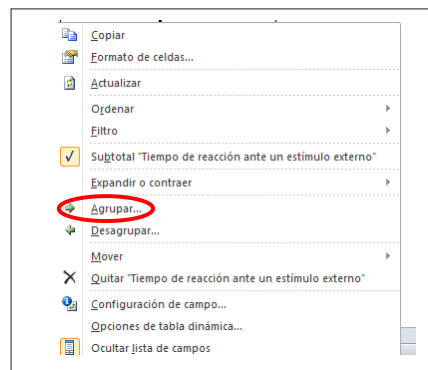


FIGURA 13: CUARTO PASO PARA GENERAR UNA TABLA DINÁMICA

11. Luego de esto aparecerá un cuadro de diálogo igual al se presenta en la Figura 14:

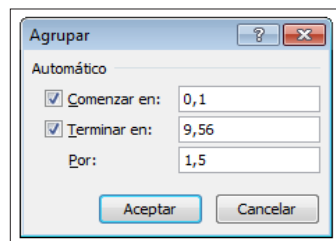


FIGURA 14: CUADRO DE DIÁLOGO PARA ESTABLECER LOS INTERVALOS

12. En “Por” puede escoger lo que se conoce como “ancho del intervalo” o “amplitud”, esto le ayudará a obtener grupos más afines en cuanto al tiempo de reacción.

Por ejemplo, si escoge 1.5 el resultado será igual al que se indica en la Figura 15:

Etiquetas de fila	Cuenta de Tiempo de reacción ante un estímulo externo
0,1-1,6	12
1,6-3,1	12
3,1-4,6	7
4,6-6,1	6
6,1-7,6	2
9,1-10,6	1
Total general	40

FIGURA 15: DATOS AGRUPADOS CON AMPLITUD 1.5

A los valores que están a la izquierda de la tabla se los conoce como límites de cada intervalo y obviamente cada uno tiene dos: límite inferior y límite superior, por ejemplo, el límite inferior del tercer intervalo es 3.1 y el límite superior del mismo es 4.6. Fíjese también que el límite superior de un intervalo puede coincidir con el inferior del siguiente, cuando esto ocurre se dice que se ha realizado un cuadro de datos agrupados con límites coincidentes; esto no siempre va a suceder y normalmente será cuando la variable se haya medido con

números decimales; pero cuando la variable es discreta, Excel hará intervalos cuyos límites no coincidan de un grupo (intervalo) a otro.

En ese caso se debe determinar lo siguiente: si un elemento coincide con un límite, ese valor pertenece al intervalo del límite superior que lo contenga, por ejemplo, si en este ejercicio hubiese un valor de 1.6, ese dato se toma en cuenta para el primer intervalo y ya no para el segundo.

En una primera observación se puede notar que la concentración de personas (frecuencia) en tiempo de reacción, es mayor hacia valores bajos de la variable que corresponde a los dos primeros intervalos [0.1 a 3.1] segundos.

Como ejemplo adicional se puede hacer un cambio en la amplitud en este caso de 2 segundos, en la Figura 16 se observa el resultado obtenido:

Tiempo de reacción por grupos	Número de personas que pertenecen a cada grupo
0,1-2,1	19
2,1-4,1	11
4,1-6,1	7
6,1-8,1	2
8,1-10,1	1
Total general	40

FIGURA 16: DATOS AGRUPADOS CON AMPLITUD 2

Con este cambio se puede observar que ya solo hay un grupo con mayor número de personas, por tanto, se intuye que la moda será un valor entre 0.1 y 2.1.

Con estos resultados el investigador puede establecer otras conclusiones y recomendaciones que serían más puntuales para cada grupo.

Como complemento a esta parte analítica se recomienda graficar (en datos agrupados) los resultados obtenidos ya que esto suele dar más luces para sustentar las conclusiones y recomendaciones; en todo caso observando el cuadro y recordando las conclusiones propuestas sobre este ejemplo ¿la tercera observación y conclusión dadas para los datos simples parecen ser las más acertadas no?

CÁLCULO DE LAS MEDIDAS DE TENDENCIA CENTRAL EN DATOS AGRUPADOS

Para calcular los valores de las Medidas de Tendencia Central debe completarse un cuadro (cuadro base) con la siguiente información en varias columnas:

Límites inferiores de cada intervalo “*L. inf.*”

Límites superiores de cada intervalo “*L. sup.*”

Punto medio (pm) de cada intervalo (también llamado marca de clase)

Frecuencia simple “*f*” de cada grupo

Frecuencia Acumulada “*fa*” de cada intervalo (representa la suma de frecuencias de un intervalo a otro hasta llegar al total)

Frecuencia relativa “*fr*” de cada grupo es el valor porcentual que representa cada una de las frecuencias simples (se encuentra dividiendo cada valor de “*f*” para el total de datos “*n*”

Frecuencia relativa acumulada “*fra*” tiene la misma idea de la frecuencia acumulada

El cuadro completo se vería tal como se muestra en la Figura 17 con el ejemplo desarrollado con amplitud 1.5

Variable		Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
L. Inf.	L. Sup.					
0,1	1,6	0,85	12	12	30,00%	30,00%
1,6	3,1	2,35	12	24	30,00%	60,00%
3,1	4,6	3,85	7	31	17,50%	77,50%
4,6	6,1	5,35	6	37	15,00%	92,50%
6,1	7,6	6,85	2	39	5,00%	97,50%
7,6	10,6	8,35	1	40	2,50%	100,00%

FIGURA 17: CUADRO BASE PARA CALCULAR MEDIDAS DE TENDENCIA CENTRAL EN DATOS AGRUPADOS

Con esto desarrollado ya se pueden realizar los cálculos para cada medida de tendencia central, así:

Los valores de Media, Mediana y Moda se pueden calcular con las siguientes fórmulas luego de agrupar los datos:

Media

$$\text{Media: } x = \frac{\sum f * pm}{n}$$

f: frecuencia simple de cada intervalo

pm: punto medio de cada intervalo

n: tamaño de la muestra

Mediana

$$\text{Mediana: } Md = L + \frac{\frac{n}{2} - faA}{f} * Am$$

Para realizar los cálculos se sugiere resaltar el intervalo en el cual se va a aplicar la fórmula.

El intervalo Mediana es aquel que garantiza haber acumulado al menos la mitad de los datos, esto viene dado por el concepto y que en la fórmula está determinado por $n/2$

L: Límite inferior del intervalo Mediana

$n/2$: mitad del total de datos

faA: frecuencia acumulada Anterior al intervalo Mediana

f: frecuencia simple del intervalo Mediana

Am : amplitud del intervalo

Moda

$$\text{Moda: } Li + \frac{fi + 1}{fi - 1 + fi + 1} * Am$$

Li : límite del intervalo Modal

$fi-1$: frecuencia simple del intervalo anterior al intervalo Modal

$fi+1$: frecuencia simple del intervalo siguiente al intervalo Modal

Am : amplitud del intervalo

El intervalo Modal es aquel que tiene la frecuencia simple más alta

Recomendación: para el caso de la Moda también se la puede obtener determinando el punto medio del intervalo donde se concentra la mayor cantidad de elementos (frecuencia simple), en todo caso será necesario verificar dos cosas:

1. Si es bimodal se sugeriría trabajar con el valor más cercano a los valores de Media y Mediana
2. Si no es bimodal y el valor del punto medio está muy alejado de los valores de Media y Mediana, se sugeriría cambiar la amplitud de los intervalos y verificar si el valor de la Moda es algo más lógico y cercano a la tendencia central.

Lo dicho anteriormente no significa que se deba ignorar el valor modal distinto a media y mediana, la sugerencia es para verificar si al cambiar de criterio en cuanto a la amplitud, los resultados de las tres medidas de tendencia central se acercan más; caso contrario el análisis debe hacerse según los resultados obtenidos.

Como en el ejemplo hay dos intervalos con igual concentración de datos, habrá entonces dos Modas, a saber: 0.85 y 2.35 (promedio obtenido con los límites de cada grupo), en este caso sería preferible trabajar con el segundo valor de la Moda, ya que está más cercano a los otros de tendencia central.

Los cálculos, utilizando las fórmulas planteadas, para las tres medidas de tendencia central con datos agrupados (tomando en cuenta la amplitud de 1.5) se ilustran a continuación en la Figura 18:

Cálculo de la Media						
Variable	Pm	f	fa	fr	fra	$f * Pm$
L. Inf. L. Sup.		(frecuencia simple)	(frecuencia acumulada)	(frecuencia relativa)	(frecuencia relativa acumulada)	
0,1 1,6	0,85	12	12	30,00%	30,00%	10,2
1,6 3,1	2,35	12	24	30,00%	60,00%	28,2
3,1 4,6	3,85	7	31	17,50%	77,50%	26,95
4,6 6,1	5,35	6	37	15,00%	92,50%	32,1
6,1 7,6	6,85	2	39	5,00%	97,50%	13,7
7,6 10,6	8,35	1	40	2,50%	100,00%	8,35
Sumatoria de $f * Pm$						119,5
División de la sumatoria						2,9875
Valor de la Media (promedio)						

FIGURA 18: CÁLCULO DEL VALOR DE LA MEDIA

Se creó una columna “Cálculo de la Media” para poder aplicar la fórmula.

Para el cálculo de la Mediana se debe establecer qué intervalo garantiza haber acumulado el 50% de los datos (concepto de Mediana), se recomienda entonces resaltar aquel grupo que garantice esto, en este caso en el segundo intervalo se encuentran 24 datos que es el primer grupo que ya garantiza tener al menos el 50% de los datos, por ello se lo resalta para mayor facilidad en la aplicación de la fórmula según se aprecia en la Figura 19.

Cálculo de la Mediana						
Variable		Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
L. Inf.	L. Sup.					
0,1	1,6	0,85	12	12	30,00%	30,00%
1,6	3,1	2,35	12	24	30,00%	60,00%
3,1	4,6	3,85	7	31	17,50%	77,50%
4,6	6,1	5,35	6	37	15,00%	92,50%
6,1	7,6	6,85	2	39	5,00%	97,50%
7,6	10,6	8,35	1	40	2,50%	100,00%

FIGURA 19: IDENTIFICACIÓN DEL INTERVALO MEDIANA PARA SU CÁLCULO

Hecho esto se reconoce cada uno de los valores que se utilizarán en la formula, para el ejemplo tenemos:

$$\text{Mediana: } Md = L + \frac{\frac{n}{2} - fa_A}{f} * Am$$

L : Límite inferior del intervalo Mediana: 1.6

$n/2$: 20

fa_A : Frecuencia acumulada anterior al intervalo señalado: 12

f : frecuencia del intervalo Mediana: 12

Am : amplitud del intervalo (es la misma para todos): 1.5

Con la identificación de estos datos, el cálculo de la Mediana al aplicar la fórmula se verá reflejado según lo que aparece en la Figura 20

Cálculo de la Mediana						
Variable		Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
L. Inf.	L. Sup.					
0,1	1,6	0,85	12	12	30,00%	30,00%
1,6	3,1	2,35	12	24	30,00%	60,00%
3,1	4,6	3,85	7	31	17,50%	77,50%
4,6	6,1	5,35	6	37	15,00%	92,50%
6,1	7,6	6,85	2	39	5,00%	97,50%
7,6	10,6	8,35	1	40	2,50%	100,00%

Aplicando la fórmula

2,6

Valor de la Mediana

FIGURA 20: CÁLCULO DEL VALOR DE LA MEDIANA

Para calcular la Moda en este caso se lo hará de dos maneras, utilizando la fórmula dada y según la recomendación antes indicada (esta última forma será la utilizada en este libro)

$$\text{Moda: } Li + \frac{fi + 1}{fi - 1 + fi + 1} * Am$$

Utilizando la fórmula:

Al igual que se hizo con la mediana, luego de identificar los valores correspondientes a: $Li: 1.6$; $fi + 1: 7$; $fi - 1: 12$ y $Am: 1.5$; el resultado se verá como se presenta en la Figura 21

Variable		Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)	Cálculo de la Moda
L. Inf.	L. Sup.						
0,1	1,6	0,85	12	12	30,00%	30,00%	Se resalta el intervalo modal (intervalo con la frecuencia simple más alta (en este caso hay dos intervalos de igual frecuencia por lo tanto se escoge aquel que esté más cercano a los valores de la Media y Mediana))
1,6	3,1	2,35	12	24	30,00%	60,00%	
3,1	4,6	3,85	7	31	17,50%	77,50%	
4,6	6,1	5,35	6	37	15,00%	92,50%	
6,1	7,6	6,85	2	39	5,00%	97,50%	
7,6	10,6	8,35	1	40	2,50%	100,00%	

Aplicando la fórmula

2,153

Valor de la Moda

Figura 21. Cálculo del valor de la Moda

FIGURA 21: CÁLCULO DEL VALOR DE LA MODA

Aplicando la recomendación se determinarían dos modas: 0.85 y 2.35 correspondientes a los puntos medios de los dos intervalos de mayor frecuencia; de ellas se escogería la de 2.35 ya que es la más cercana a la Media y Mediana y también a la obtenida con la fórmula.

Se debe aclarar que cuando se crean intervalos o grupos, se “pierde” información detallada y por ello los resultados encontrados en datos simples, no coinciden con los de los datos agrupados; y aunque suele ocurrir que la variación no es mayor, esto dependerá del mayor o menor número de grupos que se haga.

Una de las maneras de conocer el número de grupos que deberían aceptarse es obteniendo la raíz cuadrada de “n” (número de datos) y se suele recomendar que el número de intervalos no sea menor a cinco ni mayor de doce; esto no está “grabado en piedra” y dependerá de la variable y de las circunstancias específicas de ésta y (a mi juicio) obviamente del criterio y conocimiento que la persona tenga sobre la variable a investigar.

Si se opta por este proceso, la amplitud de los intervalos a escoger en Excel estaría dada por la siguiente fórmula:

$$Am = \frac{R}{\sqrt{n}}$$

En donde:

R: rango (diferencia entre el mayor y menor de los datos)

\sqrt{n} : raíz cuadrada del número de datos

Si aplicáramos esta fórmula para el ejercicio desarrollado, tendríamos lo siguiente:

$$Am = \frac{9.56 - 0.1}{\sqrt{40}}$$

$$Am = 1.49$$

Que en este caso resulta ser prácticamente la amplitud escogida con la que se desarrolló el ejercicio

GRÁFICO DE RESULTADOS

También debemos tomar en cuenta que en muchas ocasiones los resultados numéricos suelen ser difíciles de interpretar, es por ello que la presentación gráfica es un elemento de gran ayuda especialmente para aquellas personas que les es difícil visualizar numéricamente el comportamiento de la variable estudiada.

Tal vez entre los gráficos más utilizados está el de barras (columnas según el Excel), también tenemos el de pastel (circular para el Excel) que he utilizado para este ejemplo cuya utilidad radica especialmente cuando se quiere realizar un análisis porcentual de la variable, respecto al gráfico de líneas éste se usa para dar una idea del comportamiento de la variable en términos de continuidad.

Cada uno tiene su lógica de representar datos y dependerá de la variable y del investigador para que el uso de uno u otro permita ser más claro en la explicación de los resultados. Lo que sí es imprescindible aclarar es que un gráfico no demuestra nada, es decir cualquier explicación debe hacerse a través del análisis numérico, el gráfico seguramente ayudará a un mejor entendimiento.

Para el ejemplo se presenta un gráfico de pastel que debe leerse según las manecillas del reloj a partir de las 12h00, en la Figura 22 el primer grupo va de 0.1 a 1.6.

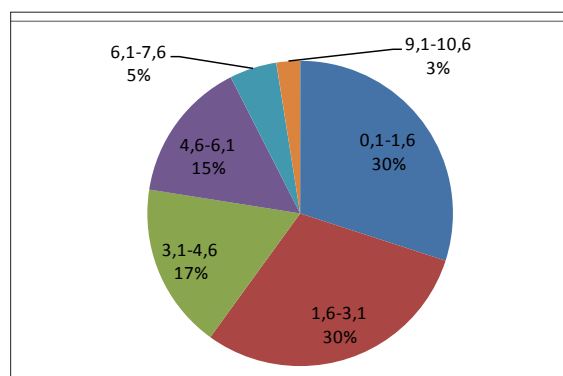


FIGURA 22: GRÁFICO DE PASTEL DEL EJEMPLO DESARROLLADO

Es bastante claro que hay dos grupos (1ero y 2do) con igual “peso” en cuanto al porcentaje y sumados los dos tendríamos un 60%, esto indica que el tiempo de reacción en la mayoría de las personas está dentro de los primeros 3.1 segundos.

Esto es lo único que se puede decir ya que no tenemos valores referenciales para comparar y por tanto opinar.

Ejemplo 2

A 81 estudiantes de segundo de básica se les aplicó el test de atención d2 de Rolf Brickenkamp y los resultados indicados en la tabla inferior corresponden al valor TOTAL

referente a la efectividad total en la prueba. Si se espera que los valores medios en esta prueba estén entre 330 y 350, qué opinión daría del grupo; esta prueba tiene como valores teóricos mínimo y máximo respectivamente: 0 y 660.

Para este ejercicio ya tenemos valores referenciales (parámetros) que nos permitirán expresar conclusiones y recomendaciones acordes a ellos.

208	260	286	312	299	325	338	364	351
351	338	403	364	403	273	351	364	312
338	273	286	377	299	286	247	429	286
325	325	364	234	325	377	364	377	299
299	273	338	312	286	312	338	221	325
338	403	234	312	429	377	234	273	351
377	273	299	390	312	429	312	364	286
325	325	390	299	325	260	403	390	234
429	286	338	338	312	299	221	247	312

Los valores de las tres medidas de tendencia central en datos simples serían los siguientes:

Media: 322.75

Mediana: 325

Moda: 312

En cuanto a los datos agrupados, el cuadro sería el que se presenta en la figura 23, tomado en cuenta que la diferencia que se ha determinado para pertenecer a un grupo u otro es de treinta puntos, el experto en este tipo de prueba puede disentir sobre esto y obviamente proponer otra división.

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i> (frecuencia simple)	<i>fa</i> (frecuencia acumulada)	<i>fr</i> (frecuencia relativa)	<i>fra</i> (frecuencia relativa acumulada)
208	237	222,5	7	7	8,64%	8,64%
238	267	252,5	4	11	4,94%	13,58%
268	297	282,5	12	23	14,81%	28,40%
298	327	312,5	24	47	29,63%	58,02%
328	357	342,5	12	59	14,81%	72,84%
358	387	372,5	11	70	13,58%	86,42%
388	417	402,5	7	77	8,64%	95,06%
418	447	432,5	4	81	4,94%	100,00%

FIGURA 23: CUADRO DE DATOS AGRUPADOS DEL EJERCICIO

Y los resultados para las medidas de tendencia central son:

Media: 323.61

Mediana: 319.87

Moda: 312.5

Gráfico de barras en datos agrupados, Figura 24

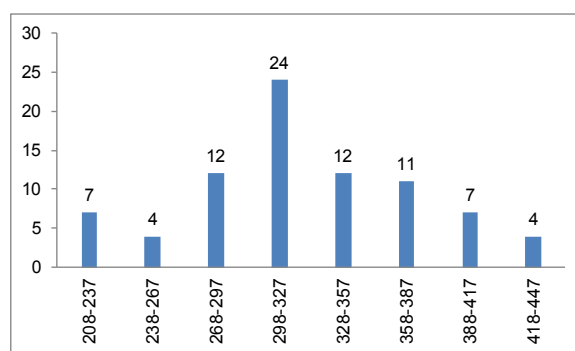


FIGURA 24: GRÁFICO DE BARRAS DEL EJEMPLO 2

El gráfico de barras nos dice que hay una buena concentración de estudiantes de segundo de básica con valores entre 298 y 327 en atención, además se puede ver que hay más estudiantes en niveles más altos de la variable estudiada a partir de este grupo; sin embargo y tomando en cuenta el valor máximo sobre el que se mide la variable, a estos estudiantes les falta mucho por alcanzar ese valor.

Cuando se trabaja con datos agrupados **no** se puede determinar con exactitud cuántos estudiantes habrán obtenido determinado puntaje, por ejemplo en el cuarto intervalo [298, 327] se sabe que hay 24 estudiantes, pero no se puede determinar cuántos habrán obtenido un puntaje de 300.

Con estos cálculos y los datos dados en el ejercicio, ¿a qué conclusión podría llegar? Y en base a esta ¿cuál o cuáles serían sus recomendaciones?

Ejemplo 3

Los valores referentes a los salarios de los jefes departamentales de varias empresas del mismo sector económico se detallan a continuación. De acuerdo a una de las encuestas salariales del mercado, los salarios de este sector se encuentran alrededor de \$750, ¿cuál sería su opinión, conclusión y recomendación al gerente general si en su organización el valor del promedio de este grupo en particular es de \$800?

390	520	640	750	820	898
390	520	647	750	840	900
400	570	650	759	847	917
400	570	658	760	860	920
450	580	677	767	870	967
465	585	690	770	870	990
490	600	700	778	879	999
510	600	720	797	880	1020
510	610	720	800	887	1040
515	640	750	810	890	1140

Agrupando los datos con una amplitud de 120 puntos se obtiene el siguiente cuadro de datos agrupados (Figura 25):

L. Inf.	L. Sup.	Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
390	509	449,5	7	7	11,67%	11,67%
510	629	569,5	12	19	20,00%	31,67%
630	749	689,5	10	29	16,67%	48,33%
750	869	809,5	15	44	25,00%	73,33%
870	989	929,5	11	55	18,33%	91,67%
990	1109	1049,5	4	59	6,67%	98,33%
1110	1229	1169,5	1	60	1,67%	100,00%

FIGURA 25: CUADRO DE DATOS AGRUPADOS EJEMPLO 3

Gráfico de línea en datos agrupados (Figura 26)

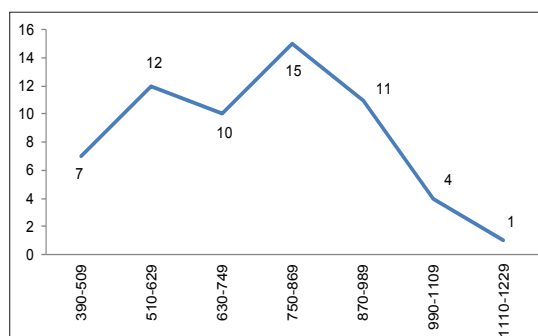


FIGURA 26: GRÁFICO DE LÍNEA DEL EJEMPLO 3

Resultados de las medidas de tendencia central en datos simples y agrupados (Tabla 3)

	Datos simples	Datos agrupados
Media	722,36	743,5
Mediana	750	758
Moda	750	809,5

TABLA 3. RESULTADOS REFERENTES AL EJEMPLO 3

Fíjese que los valores obtenidos para las tres medidas de tendencia central son distintos (especialmente en la Moda) según se calcule de manera simple o agrupada, como se había indicado previamente, esto dependerá del número de grupos que se formen, pero debe primar un criterio técnico respecto a la amplitud (o diferencia) de cada intervalo.

Cabe recalcar entonces que en la resolución de ejercicios hay dos cosas a tomar en cuenta:

1. Cálculos numéricos
2. Análisis a partir de dichos cálculos

El primer punto tiene la importancia de que, si se hacen mal, cualquier análisis posterior será vano, la ventaja es que en Excel (o cualquier otro procesador) este tema se eliminará o al menos se minimizará y no tendrá consecuencias.

La segunda parte es la más importante ya que si se limita a encontrar resultados, estadísticamente no se ha hecho nada y solo se ha realizado un proceso aritmético; la Estadística (al menos a mi juicio) es el complemento de estos dos pasos que permitirán al investigador proponer soluciones prácticas en base a los resultados y condiciones en las que se desarrolla la variable estudiada, es por ello que en cada ejercicio se pide concluir y recomendar aunque en algunas ocasiones (por ejemplo primer ejercicio resuelto) no se tienen datos (parámetros) para poder establecer la situación específica de la muestra a estudiar.

Ejemplo 4

En una institución educativa se tomaron pruebas de ingreso en varias áreas de conocimiento, el cuadro siguiente representa los resultados en Matemáticas (valor máximo 100 puntos) de todos los estudiantes que se inscribieron en dicha institución. Aplicando cualquier técnica de muestreo, escoja dos muestras distintas y compare los resultados; la condición para aprobar es tener una nota superior a 70 puntos; establezca conclusiones.

61	36	79	74	93	75	84	44	93	91
91	80	26	99	77	75	83	89	51	73
54	75	54	63	67	86	43	97	22	93
50	53	93	33	54	50	78	39	35	49
95	50	82	26	44	80	31	81	64	19
64	71	97	72	18	24	70	93	32	60
97	84	72	66	67	50	71	88	38	72
81	99	77	55	92	40	50	41	86	86
39	55	88	48	73	92	35	90	40	79
63	84	75	22	95	75	55	76	75	49
22	27	84	39	32	80	29	57	19	47
53	33	65	76	25	86	58	88	20	81
76	68	28	37	53	65	57	66	90	56
70	55	42	81	79	40	86	35	26	51
65	39	34	28	50	63	89	21	36	79
60	58	91	27	35	97	19	78	57	68
63	52	78	82	76	88	78	65	68	75
57	69	28	91	25	34	35	71	28	60
88	50	92	42	39	29	50	50	81	18
65	66	43	99	96	26	37	46	86	84
66	57	45	53	98	37	49	77	55	39
59	91	80	20	97	62	52	73	62	58
99	71	63	19	38	21	25	49	45	43
55	44	95	42	99	20	72	89	82	42
73	92	88	64	85	61	98	99	61	24
62	38	29	20	77	45	94	84	83	22
49	89	62	66	44	94	83	44	65	33
93	54	33	80	82	59	33	61	41	26
57	19	35	91	27	87	18	63	29	53
96	30	76	64	26	97	22	43	24	34

Se han escogido 60 datos para cada muestra, el cuadro siguiente representa el primer grupo

18	50	78
19	50	80
22	55	82
26	57	82
26	57	83
26	61	84
28	61	85
30	62	88
34	63	89
39	63	91
39	64	92
41	64	93
43	65	93
43	65	94
44	66	95
44	73	95
45	73	96
47	75	96
49	75	97
50	76	99

En la Figura 27 se aprecia el cuadro de datos agrupados de esta muestra.

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i> (frecuencia simple)	<i>fa</i> (frecuencia acumulada)	<i>fr</i> (frecuencia relativa)	<i>fra</i> (frecuencia relativa acumulada)
18	27	22,5	6	6	10,00%	10,00%
28	37	32,5	3	9	5,00%	15,00%
38	47	42,5	9	18	15,00%	30,00%
48	57	52,5	7	25	11,67%	41,67%
58	67	62,5	10	35	16,67%	58,33%
68	77	72,5	5	40	8,33%	66,67%
78	87	82,5	7	47	11,67%	78,33%
88	97	92,5	12	59	20,00%	98,33%
98	107	102,5	1	60	1,67%	100,00%

FIGURA 27: CUADRO DE DATOS AGRUPADOS EJEMPLO 4, PRIMER GRUPO

En la Tabla 4 se presentan los resultados de las medidas de tendencia central tanto en datos simples como agrupados

	Datos Simples	Datos agrupados
Media	63	62,67
Mediana	63,5	58,8
Moda	50	97

TABLA 4. RESULTADOS REFERENTES AL EJEMPLO 4, PRIMER GRUPO

Gráficamente los resultados del primer grupo se verían según la Figura 28

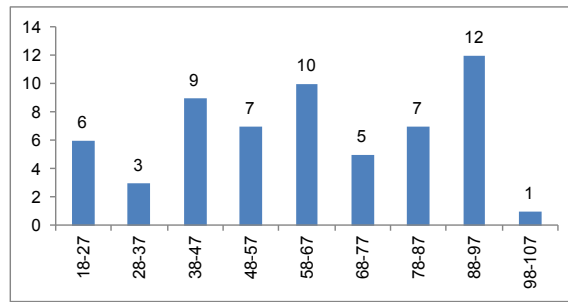


FIGURA 28: GRÁFICO DE LÍNEA DEL EJEMPLO 4, PRIMER GRUPO

Se escogieron también 60 datos para el *segundo grupo*

18	43	70
18	43	71
20	43	72
22	44	75
24	44	75
24	44	75
25	50	76
26	50	81
28	53	86
32	55	86
33	55	86
33	57	88
34	58	91
34	58	95
35	61	95
38	61	97
39	64	97
40	64	98
42	65	99
42	69	99

En la Figura 29 se presentan los valores de cada columna del cuadro de datos agrupados de este grupo

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i> (frecuencia simple)	<i>fa</i> (frecuencia acumulada)	<i>fr</i> (frecuencia relativa)	<i>fra</i> (frecuencia relativa acumulada)
18	27	22,5	8	8	13,33%	13,33%
28	37	32,5	7	15	11,67%	25,00%
38	47	42,5	11	26	18,33%	43,33%
48	57	52,5	6	32	10,00%	53,33%
58	67	62,5	7	39	11,67%	65,00%
68	77	72,5	8	47	13,33%	78,33%
78	87	82,5	4	51	6,67%	85,00%
88	97	92,5	6	57	10,00%	95,00%
98	107	102,5	3	60	5,00%	100,00%

FIGURA 29: CUADRO DE DATOS AGRUPADOS EJEMPLO 4, SEGUNDO GRUPO

La Tabla 5 resume los valores de las medidas de tendencia central del segundo grupo

	Datos Simples	Datos agrupados
Media	56,67	56,67
Mediana	55,00	48,76
Moda	86,00	42,50

TABLA 5. RESULTADOS REFERENTES AL EJEMPLO4, SEGUNDO GRUPO

El gráfico de barras correspondiente a este grupo se puede apreciar en la Figura 30

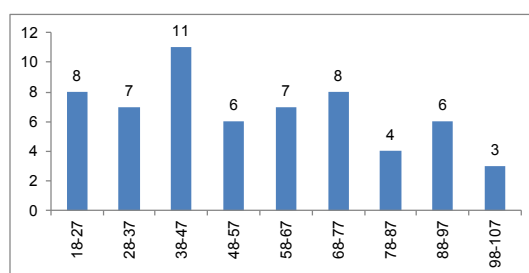


FIGURA 30: GRÁFICO DE BARRAS DEL EJEMPLO 4, SEGUNDO GRUPO

Comparación de resultados (Tabla 6)

	GRUPO 1		GRUPO 2	
	Datos Simples	Datos agrupados	Datos Simples	Datos agrupados
Media	63	62,67	56,67	56,67
Mediana	63,5	58,8	55	48,76
Moda	50	97	86	42,5

TABLA 6: CUADRO DE RESULTADOS COMPARATIVOS

Como se puede notar los resultados son distintos y habrá que hacer un análisis para determinar si existen diferencias significativas entre los resultados obtenidos.

Por lo pronto es muy notoria la diferencia que existe dentro de cada grupo en los valores de la moda, como se indicara en líneas anteriores, esto suele depender de la amplitud que se haya tomado para agrupar los datos y también habrá que revisar si los valores obtenidos son realmente representativos.

En ambos casos se establece que los valores de media y mediana en datos simples son muy cercanos, hay que recordar también que cuando se agrupan datos se “pierde información”.

En todo caso la conclusión es obvia: los resultados son muy bajos y en ninguna de las dos muestras se logran puntajes medios correspondientes al mínimo requerido y la proporción de estudiantes que superan dicho puntaje es muy pequeña.

En este caso y para determinar la proporción de cumplimiento respecto a la nota mínima requerida se debe obtener el porcentaje de aspirantes que lo lograron, estos porcentajes se deben analizar en datos simples dado que en agrupados se ha perdido información sobre

quienes exactamente obtuvieron puntajes de 70 o más, para cada grupo los resultados son los siguientes:

Grupo 1: 41.66%

Grupo 2: 33.33%

Estos porcentajes se ven también reflejados si se comparan las medidas de tendencia central de ambos grupos siendo así que el primer grupo tiene valores medios más altos, lo cual confirmaría que el grupo de esa muestra fue mejor.

Veamos ahora qué valores se obtienen analizando a toda la población, esto se aprecia en la Figura 31.

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i> (frecuencia simple)	<i>fa</i> (frecuencia acumulada)	<i>fr</i> (frecuencia relativa)	<i>fra</i> (frecuencia relativa acumulada)
18	27	22,5	34	34	11,33%	11,33%
28	37	32,5	31	65	10,33%	21,67%
38	47	42,5	32	97	10,67%	32,33%
48	57	52,5	41	138	13,67%	46,00%
58	67	62,5	39	177	13,00%	59,00%
68	77	72,5	35	212	11,67%	70,67%
78	87	82,5	39	251	13,00%	83,67%
88	97	92,5	41	292	13,67%	97,33%
98	107	102,5	8	300	2,67%	100,00%

FIGURA 31: CUADRO DE DATOS AGRUPADOS DE LA POBLACIÓN

En base al cuadro anterior, se obtienen los resultados indicados en la Tabla 7

	Datos simples	Datos agrupados
Media	60,22	60,3
Mediana	61,5	61,08
Moda	50	52,5 / 92,5

Bimodal

TABLA 7. VALORES DE LAS MEDIDAS DE TENDENCIA CENTRAL EN DATOS AGRUPADOS

Al parecer los valores del primer grupo están más cercanos a los resultados de la población, en todo caso tampoco hay mucha diferencia con los parámetros poblacionales y la conclusión será la misma ya dicha para ambos grupos.

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

1. Los siguientes datos corresponden a los resultados de una prueba de mecanografía a 100 personas de quienes se requería conocer su capacidad de escritura instrumental. Determinar la variable y encontrar las tres medidas de tendencia central tanto en datos simples como agrupados (se sugiere usar una amplitud de 5 puntos) y determinar

sus conclusiones sabiendo que se espera resultados medios de 45 con un mínimo de 20 y máximo de 65

22	29	32	36	38	40	43	46	48	51
23	30	33	37	39	40	43	46	48	51
23	31	33	37	39	40	43	46	48	52
23	31	34	37	39	40	44	46	48	52
26	31	34	37	39	41	44	46	49	53
26	31	34	37	39	41	44	46	49	54
26	31	34	37	40	41	45	47	49	55
28	32	35	38	40	41	45	47	49	57
29	32	35	38	40	41	46	47	50	58
29	32	35	38	40	42	46	47	51	58

2. Se aplicó – a un grupo de niños entre 6 y 7 años – el test ABC de Lorenzo Filho que mide el nivel de madurez (comprende 8 áreas, entre ellas coordinación auditiva, fonética, atención concentrada y otros) siendo la puntuación máxima 24 puntos. Los valores son los siguientes:

MADUREZ					
12	15	16	17	18	19
12	15	16	17	18	19
12	16	16	17	18	19
13	16	16	17	18	19
14	16	17	17	18	20
15	16	17	17	18	20
15	16	17	17	18	20
15	16	17	17	18	21
15	16	17	17	19	21
15	16	17	18	19	21

Determine: la variable, las medidas de tendencia central en datos simples y agrupados. Si al aplicar esta prueba se espera que los valores centrales estén entre 18 y 20, qué conclusiones obtiene y qué recomendaría.

3. Un estudio realizado en una población rural del Ecuador determinó que, de la muestra obtenida de los adultos, estos presentaban problemas de desarrollo psicomotriz. Para avalar este estudio se presentaron los siguientes datos:

16	27	34	40	45	48	51
18	28	34	40	46	48	51
19	28	35	41	46	48	52
19	29	35	42	46	48	56
20	31	35	42	47	49	56
24	31	35	42	47	49	56
25	32	37	43	47	49	57
26	32	38	45	47	50	57
26	33	38	45	47	50	57
27	33	39	45	48	50	58

Las pruebas realizadas se miden sobre 60 puntos y se supone que un puntaje inferior a 35 es preocupante. Determine: la variable y las medidas de tendencia central en datos simples y agrupados e indique si está o no de acuerdo con la conclusión del estudio realizado.

4. Los resultados de la evaluación de desempeño en una organización se indican a continuación

51	63	68	72	90	81	84	88	92	96
99	65	69	72	98	81	78	89	78	97
56	95	69	75	86	83	86	89	93	98
56	92	71	93	94	68	87	69	77	72
57	86	83	77	80	84	87	90	96	65
86	85	80	94	96	75	76	48	54	98
65	73	79	66	84	96	96	87	88	85
76	68	79	94	83	90	68	68	87	79
96	90	60	85	80	70	94	60	84	66
80	76	86							

Si el valor de la Mediana en esta evaluación se espera sea al menos de 85, qué opinión daría usted de estos resultados. Para un mejor análisis, debe obtener las medidas de tendencia central en datos simples y agrupados.

5. Los datos a continuación se han obtenido de una prueba que mide la motricidad (medida sobre 75) en niños entre 2 y 4 años. Con ellos encuentre las medidas de tendencia central tanto en datos simples como agrupados y responda las preguntas conociendo que los valores normales para esta prueba fluctúan entre 48 y 50 puntos: a) si un niño tiene valores inferiores a 45 ¿se debe referir a un especialista? b) si un niño tiene valores superiores a 58 ¿es recomendable hacer un seguimiento para potencializar aún más su capacidad? ¿Qué opinión le merecen los datos?, ¿qué conclusión obtiene?, ¿qué recomendaría para aquellos casos que presenten dificultades según sus cálculos?

35	45	47	49	52	56
40	46	47	49	52	56
42	46	47	49	53	57
42	46	48	50	53	57
42	46	48	50	53	57
42	46	48	50	53	58
43	46	48	51	54	59
43	47	48	51	54	59
44	47	48	51	56	60
44	47	49	51	56	60

6. Según el test de Zung que mide depresión y ansiedad, se tomaron pruebas a personas de un centro de salud y se obtuvieron los siguientes datos. Se conoce que este test tiene como valor Mínimo: 45 y como valor máximo: 95. Para el ejemplo se presentan los resultados obtenidos en ansiedad.

ANSIEDAD									
45	46	49	52	56	58	60	70	75	80
45	46	50	54	56	60	62	70	75	82
45	46	50	55	56	60	62	70	75	85
45	48	50	55	58	60	65	75	75	85
45	48	50	55	58	60	65	75	76	90
45	49	52	56	58	60	65	75	76	90

Determine la variable, si usted considera que es necesario formar grupos, hágalo; en cualquier caso determine las medidas de tendencia central tanto en datos simples como en agrupados, elabore un gráfico y explique sus conclusiones y recomendaciones.

7. En el test AMPE de inteligencia multifactorial se mide, entre otras cosas, el razonamiento numérico (valor máximo 130 puntos y valores inferiores a 95 indican deficiencia en esta variable). Los siguientes valores se obtuvieron de un grupo de 90 estudiantes de un colegio del valle. Obtenga las medidas de tendencia central en datos simples y agrupados. Concluya y recomiende

RAZONAMIENTO NUMÉRICO								
75	106	120	90	98	110	125	80	95
80	109	120	112	97	115	120	112	112
95	110	120	100	125	110	115	120	112
80	110	90	100	114	100	95	115	115
112	96	85	102	110	98	115	114	115
85	90	90	125	120	120	115	116	125
90	83	120	120	85	95	95	120	80
95	120	125	110	95	114	100	95	110
120	115	125	105	90	115	110	90	120
103	120	130	95	120	120	120	125	90

8. Los datos que se dan a continuación corresponden al tiempo (en minutos) que se demoraron los integrantes de un grupo de estudiantes en resolver un test psicológico. Encuentre las tres medidas de tendencia central. Elabore un cuadro de datos agrupados con intervalos de 0.8 minutos. Si el tiempo medio para resolver este test oscila entre 20 y 24 minutos, ¿cómo calificaría usted a este grupo? ¿Por qué?

19,0	19,9	20,7	21,1	21,8	22,7	23,1	23,8	24,1	24,3
19,5	20,1	20,8	21,2	21,9	22,8	23,3	23,8	24,1	25,0
19,5	20,3	20,9	21,3	22,0	22,8	23,5	23,8	24,2	25,0
19,7	20,7	20,9	21,5	22,2	22,8	23,6	23,9	24,2	25,1
19,8	20,7	20,9	21,6	22,5	22,9	23,7	23,9	24,2	25,3

9. Se encontraron los siguientes datos respecto a la puntuación de una prueba de rapidez mental medida sobre 25 puntos a 90 candidatos para el puesto de ventas en una empresa comercial, se supone que mientras más alto es el puntaje, mejor capacidad demuestra.

4	18	4	12	6	20	15	15	17
9	5	18	15	15	14	6	9	14
18	15	10	20	20	18	13	14	15
15	19	6	14	14	19	14	18	8
19	10	20	15	18	16	20	9	15
16	18	6	1	13	13	18	15	17
4	5	19	3	7	6	19	16	16
10	10	10	10	14	7	16	14	9
18	6	18	8	12	4	17	18	10
7	20	16	14	15	13	8	4	15

Si lo esperado es que los resultados de la prueba tengan puntajes medios entre 11 y 13.5, cuántas personas están fuera de este intervalo. Se supone que una diferencia de 2.5

puntos es importante. Encuentre las medidas de tendencia central tanto en datos simples como agrupados, elabore un gráfico (el que según su criterio sea más adecuado) e indique una conclusión y la recomendación correspondiente.

10. Los datos siguientes corresponden a los resultados de una investigación sobre conocimientos de Literatura Universal (medida sobre 100 puntos) aplicada a un grupo de estudiantes de 2do curso de bachillerato de un determinado centro educativo de nivel medio. Encuentre las medidas de tendencia central en datos simples y agrupados conociendo que 10 puntos hacen diferencia en este tipo de conocimiento, elabore un gráfico de barras y establezca una conclusión y una recomendación.

3,75	8,75	18,75	24,00	36,12
3,75	9,00	19,00	25,00	36,50
3,75	9,12	20,38	25,00	36,70
4,75	9,88	20,38	26,00	46,00
5,50	10,88	20,62	27,12	46,38
5,75	12,38	21,88	29,38	53,88
6,00	12,88	22,50	31,00	57,50
6,12	14,25	22,62	33,38	59,12
8,25	15,50	23,50	35,00	64,75
8,62	15,88	23,60	35,25	80,50

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1

Variable: capacidad de escritura instrumental

Medidas de tendencia central en datos simples (Tabla 8)

Media	40,13
Mediana	40
Moda	40

TABLA 8. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL DEL EJERCICIO 1, DATOS SIMPLES

Elaboración del cuadro de datos agrupados (Figura 32)

L. Inf.	L. Sup.	Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
22	26	24	7	7	7,00%	7,00%
27	31	29	10	17	10,00%	17,00%
32	36	34	14	31	14,00%	31,00%
37	41	39	28	59	28,00%	59,00%
42	46	44	17	76	17,00%	76,00%
47	51	49	16	92	16,00%	92,00%
52	56	54	5	97	5,00%	97,00%
57	61	59	3	100	3,00%	100,00%
			Media			
			40,05			
			Mediana			
			40,39			
			Moda			
			39			

FIGURA 32: RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL EJERCICIO 1 EN DATOS AGRUPADOS

Tanto en datos simples como en agrupados los valores encontrados están alejados del valor medio esperado (45), por lo tanto, se puede concluir que el grupo está un poco bajo en su capacidad de escritura instrumental.

La recomendación sería tomar en cuenta para una segunda etapa de selección (por ejemplo) a quienes hayan obtenido un puntaje igual o superior a 47, es decir a las 24 personas que se encuentran en los tres últimos intervalos.

Ejercicio 3

Variable: desarrollo motriz en adultos de una población rural del Ecuador
Medidas de tendencia central en datos simples (Tabla 9)

Media	40,21
Mediana	42
Moda	48

TABLA 9. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL DEL EJERCICIO 3, DATOS SIMPLES

Elaboración del cuadro de datos agrupados (Figura 33)

L. Inf.	L. Sup.	Pm	f (frecuencia simple)	fa (frecuencia acumulada)	fr (frecuencia relativa)	fra (frecuencia relativa acumulada)
16	23	19,5	5	5	7,14%	7,14%
24	31	27,5	11	16	15,71%	22,86%
32	39	35,5	14	30	20,00%	42,86%
40	47	43,5	19	49	27,14%	70,00%
48	55	51,5	14	63	20,00%	90,00%
56	63	59,5	7	70	10,00%	100,00%
			Media		40,87	
			Mediana		42,11	
			Moda		43,5	

FIGURA 33: RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL EJERCICIO 3 EN DATOS AGRUPADOS

Con la condición dada en el ejercicio: “las pruebas se miden sobre 60 puntos y se supone que un puntaje inferior a 35 es preocupante”; se puede determinar que la conclusión a la que llegaron en este estudio sí tiene fundamentos, veamos:

1. Los valores de tendencia central están muy cerca del valor establecido como preocupante.
2. Según el cuadro de datos agrupados, el 42.85% de los participantes tiene valores inferiores a 39 y si 35 representa un valor muy crítico, el porcentaje resulta ser muy elevado.
3. En cuanto a los datos simples se determina con exactitud que hay 26 personas que no lograron superar el valor crítico y éstas representan un 37.14% que obviamente es un alto porcentaje.

Ejercicio 5

Variable: nivel de motricidad en niños entre 2 y 4 años

Medidas de tendencia central en datos simples (Tabla 10)

Media	49,55
Mediana	49
Moda	46

TABLA 10. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL DEL EJERCICIO 5, DATOS SIMPLES

Elaboración del cuadro de datos agrupados, se estableció una amplitud de cuatro puntos (Figura 34)

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
35	38	36,5	1	1	1,67%	1,67%
39	42	40,5	5	6	8,33%	10,00%
43	46	44,5	11	17	18,33%	28,33%
47	50	48,5	19	36	31,67%	60,00%
51	54	52,5	12	48	20,00%	80,00%
55	58	56,5	8	56	13,33%	93,33%
59	62	60,5	4	60	6,67%	100,00%

Media	49,57
Mediana	49,74
Moda	48,5

FIGURA 34: RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL EJERCICIO 5 EN DATOS AGRUPADOS

Observación: de acuerdo a la información dada, los valores medios del grupo estudiado coinciden con lo esperado dado que se encuentran entre 48 y 50.

En cuanto a las preguntas a y b será el criterio de cada persona; se recomienda en clase hacer un análisis con la guía del profesor.

Conclusión: en general el grupo de niños supera las expectativas (alrededor del 72%) y según los valores de las medidas de tendencia central, este grupo está dentro de los valores normales, pero se puede ver que hay 17 niños que obtuvieron valores inferiores a lo considerado normal, esto representa un 28.33% de los datos y es un valor bastante alto, hay que recordar que se está tratando de motricidad en niños de 2 a 4 años, por tanto este resultado es preocupante.

Sobre lo anterior también se debe anotar algo: el grupo de estudio son niños entre 2 y 4 años y desde ese punto de vista, aunque el 28.33% sea un valor alto también se puede argumentar que dado la edad del grupo estudiado, no hay mucho por qué preocuparse dada su temprana edad; esos criterios que en sí son contradictorios, pueden servir de base para una discusión en términos profesionales y educativos de análisis.

La recomendación obvia será hacer un análisis y seguimiento a los 17 niños que presentan dificultades de motricidad pero sin ejercer mucha presión tampoco sobre ellos, precisamente por su edad muy temprana.

Ejercicio 7

Variable: análisis de Razonamiento Numérico en un grupo de estudiantes de un colegio del valle.

Opcional: análisis de Razonamiento Numérico, con la aplicación del test AMPE que mide inteligencia multifactorial, en un grupo de estudiantes de un colegio del valle

Medidas de tendencia central en datos simples (Tabla 11)

Media	106,67
Mediana	110
Moda	120

TABLA 11. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL DEL EJERCICIO 7, DATOS SIMPLES

Elaboración del cuadro de datos agrupados; ¿cuál es la amplitud utilizada para el desarrollo de este ejercicio? (Figura 35)

L. Inf.	L. Sup.	Pm	f	fa	fr	fra
75	84	79,5	6	6	6,67%	6,67%
85	94	89,5	11	17	12,22%	18,89%
95	104	99,5	19	36	21,11%	40,00%
105	114	109,5	19	55	21,11%	61,11%
115	124	119,5	27	82	30,00%	91,11%
125	134	129,5	8	90	8,89%	100,00%

Media	107,72
Mediana	109,74
Moda (bimodal)	99,5 109,5

FIGURA 35: RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL EJERCICIO 7 EN DATOS AGRUPADOS

Observación: las medidas de tendencia central indican estar muy por encima del valor crítico de 95, pero también están a una buena distancia del valor máximo de 130.

Conclusión: si ordenamos los datos, se podrá comprobar que hay 17 alumnos con un nivel de 120, siete con nivel de 125 y uno que llegó al máximo; esto representa el 27.77% de estudiantes con niveles muy elevados de razonamiento numérico lo cual es muy bueno tanto para los alumnos como para la institución. Este grupo de estudiantes en general tiene un buen nivel de razonamiento numérico lo cual podría indicar un alto potencial para estudios en el área de la Matemática y afines.

Pero del otro lado es preocupante que otros 17 alumnos se encuentren por debajo del valor crítico de 95 y otros 9 estén justo en el límite; es decir hay marcadas diferencias entre estos estudiantes en referencia a la variable estudiada.

Recomendación: trabajar con el grupo de estudiantes con valores bajos realizando ejercicios que les ayude a mejorar esta capacidad.

Ejercicio 9

Variable: rapidez mental en candidatos al puesto de ventas de una empresa comercial
Medidas de tendencia central en datos simples (Tabla 12)

Media	12,92
Mediana	14
Moda	15

TABLA 12. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL DEL EJERCICIO 9, DATOS SIMPLES

Elaboración del cuadro de datos agrupados, Figura 36

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
1	3,5	2,25	2	2	2,22%	2,22%
3,5	6	4,75	7	9	7,78%	10,00%
6	8,5	7,25	12	21	13,33%	23,33%
8,5	11	9,75	11	32	12,22%	35,56%
11	13,5	12,25	6	38	6,67%	42,22%
13,5	16	14,75	21	59	23,33%	65,56%
16	18,5	17,25	20	79	22,22%	87,78%
18,5	21	19,75	11	90	12,22%	100,00%

Media	13,08
Mediana	14,33
Moda	14,75

FIGURA 36: RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL EJERCICIO 9 EN DATOS AGRUPADOS

Si lo esperado es que los resultados de la prueba tengan puntajes medios entre 11 y 13.5, cuántas personas están fuera de este intervalo

Observación: para contestar la pregunta y si nos fijamos en el cuadro de datos agrupados, solo hay 6 personas que se encuentran dentro de dicho intervalo, por lo tanto, 84 personas se encuentran fuera del intervalo.

Esta sería la respuesta a la pregunta, pero lo importante es determinar cuántos están **por debajo** de los valores esperados ya que si de lo que se trata es de determinar rapidez mental, quienes tengan puntajes superiores a lo esperado, serán idóneos para el puesto.

Conclusión: en este caso vemos que, sin tomar en cuenta quienes están dentro del intervalo esperado, hay 52 personas (esto es el 57.77%) que logran valores superiores lo cual puede indicar algunas cosas, entre ellas que la prueba no está midiendo correctamente el nivel de rapidez mental (validez del instrumento) o que quienes tomaron la prueba tienen esta facultad muy bien desarrollada; de todas maneras algunas personas de ese grupo seguro podrán contar con el trabajo.

De otro lado y por el tipo de variable se supondría que quienes obtuvieron puntajes inferiores a lo esperado (esto es el 42.22%) no serán seleccionados y quedarán fuera del concurso.

Recomendación: al parecer la prueba no es muy discriminatoria y se sugiere tomar en cuenta otras variables para una posible selección de este tipo de personal.

En la Figura 37 se resalta los cuatro primeros grupos con valores inferiores indicando que quienes pertenezcan a ellos no podrán seguir en el proceso de selección por haber obtenido valores bajos en la variable estudiada.

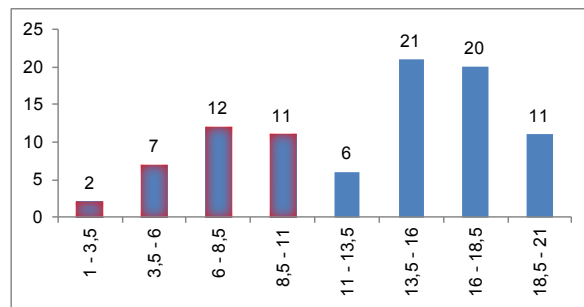


FIGURA 37: GRÁFICO EXPLICATIVO DE LOS RESULTADOS DEL EJERCICIO 9

CAPÍTULO 3:

MEDIDAS DE FORMA

Luego de establecer el estudio de las medidas de tendencia central ya se puede incursionar en otro tipo de medidas que ayudarán a estudiar con más profundidad cualquier variable en base a la distribución de frecuencias que presenten. Estas medidas llamadas de forma tienen dos visiones de estudio que permiten entenderlas mejor, a saber:

1. Perspectiva Analítica
2. Perspectiva Gráfica

En cualquiera de las dos formas de estudio, se establecen dos temas que son complementarios entre sí, estos son:

1. Estudio de la Simetría - Asimetría (Sesgos)

Tanto en la distribución simétrica como en la asimétrica se necesita establecer una comparación entre las medidas de tendencia central para su determinación.

2. Estudio de la Curtosis

A su vez el estudio de la curtosis hace referencia al distanciamiento entre los extremos dentro de una distribución de frecuencias indicando la concentración de datos alrededor de la Media, su valor está dado por otros distintos a las medidas de tendencia central.

En cuanto a los análisis analíticos y gráficos y como se dijera en páginas anteriores, es de suma importancia que quien realice la investigación sepa “leer” bien los resultados en ambos casos.

DISTRIBUCIÓN DE FRECUENCIAS

¿Qué es una distribución de frecuencias? Pues se refiere a las variadas formas gráficas que pueden presentarse de acuerdo a la cantidad de elementos que se encuentren en los distintos intervalos (datos agrupados). Por lo general se clasifican en dos: distribuciones simétricas y distribuciones asimétricas (sesgadas)

Una curva es **simétrica** si, al plegarla por la mitad, ambos lados coinciden. Si una curva no es simétrica, es **sesgada** (Pagano, 2011, p. 60).

Empezaré haciendo el estudio de la distribución no simétrica ya que la curva simétrica conocida como Distribución Normal requiere de otros conceptos que se verán más adelante.

ASIMETRÍAS (SESGOS)

El estudio de la simetría y asimetría da información basada en la relación numérica (aritmética – posicional) de las medidas de tendencia central y determinada gráficamente por la distribución de frecuencias (el gráfico de barras es muy útil para su cualificación).

En cuanto a las asimetrías, se debe comparar numéricamente entre las medidas de tendencia central cuál es la menor, la intermedia y la mayor de ellas y así se establece una subclasificación de las asimetrías que es la siguiente:

1. Asimetría Negativa (sesgo negativo)
2. Asimetría Positiva (sesgo positivo)

Pero en esta comparación de valores pueden darse en general dos casos:

1. Que las tres medidas de tendencia central sean iguales (o se las pueda considerar así) en ese caso estaríamos hablando de Simetría, y
2. Que esos valores no concuerden con la clasificación para determinar el sesgo como positivo o negativo, en ese caso se dice simplemente que son asimétricos (o simplemente que están sesgados) sin determinar un tipo específico.

Todas ellas se refieren y dan cuenta de un determinado tipo de distribución (distribución gráfica de frecuencias) que toma la variable estudiada, esto tiene mucha aplicación en cualquier tipo de análisis y en Psicología y Educación es fundamental su cálculo e interpretación especialmente en estudios grupales.

Revisemos numéricamente cada una de ellas.

Asimetría Negativa

Forma Analítica: cuando las tres medidas de tendencia central tienen la siguiente relación numérica:

$$\bar{X} < Md < Mo$$

Es decir la Media es la de menor valor, luego será la Mediana y el valor más alto será el de la Moda.

Esto significa que la tendencia del grupo estudiado es a obtener valores altos de la variable; recuerde que en un eje horizontal, los valores se expresan como en la recta numérica, es decir de manera ascendente de izquierda a derecha, este tipo de asimetría se presenta en la Figura 38.

Forma Gráfica

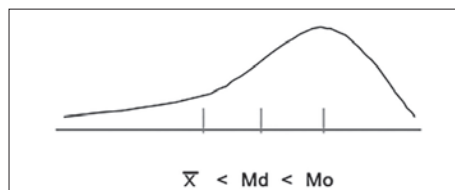


FIGURA 38: DISTRIBUCIÓN DE LA ASIMETRÍA NEGATIVA

La interpretación del gráfico se debe hacer desde dos visiones:

1. Los valores de las Medidas de Tendencia Central se ubican en la recta y es por ello que se observa que la Media está a la izquierda (valor menor) y la Moda a la derecha (valor mayor)
2. La “cresta” de la curva está sobre el valor de la Moda, es decir la mayor concentración de datos está hacia valores más altos (hay más elementos que han logrado valores altos en la variable estudiada).

Si nos fijamos en el gráfico, la “cresta” ubica a la moda; un sistema nemotécnico que da resultado es fijarse hacia dónde se prolonga la “cola” de la curva, para este tipo de distribución la cola va hacia la izquierda, es por ello que se dice que el sesgo será negativo, esto en referencia al estudio de la recta numérica (los valores negativos están a la izquierda del origen).

Asimetría Positiva

Forma Analítica: cuando las tres medidas de tendencia central tienen la siguiente relación numérica:

$$Mo < Md < \bar{X}$$

Es decir la Moda es la de menor valor, luego será la Mediana y el valor más alto será el de la Media.

Esto significa que la tendencia del grupo estudiado es a obtener valores bajos de la variable. Gráficamente se establece que hay más personas que tienen valores más hacia la izquierda (valores menores) de la medición, como se presenta en la figura 39.

Forma Gráfica

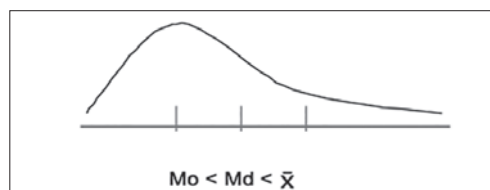


FIGURA 39: DISTRIBUCIÓN DE LA ASIMETRÍA POSITIVA

La interpretación del gráfico es similar al caso anterior:

1. Los valores de las Medidas de Tendencia Central se ubican en la recta y es por ello que se observa que la Moda está a la izquierda (valor menor) y la Media a la derecha (valor mayor)
2. La “cresta” de la curva está sobre el valor de la Moda, es decir la mayor concentración de datos está hacia valores más bajos.

Al igual que el caso anterior, se puede notar que la “cola” de la curva tiende hacia la derecha que en términos de recta numérica indica valores positivos, por ello el tipo de sesgo se dice positivo.

Como podrá notarse, en cualquier caso, la “cresta” de la curva siempre indica la posición de la Moda

Dentro de este estudio, encontramos también una relación muy especial entre las tres medidas de tendencia central y es cuando éstas teóricamente son iguales, (aunque en la práctica esto es muy difícil que se dé), podemos aceptar como hecho que si son muy cercanas entre sí, esta distribución se denomina Normal o Simétrica y por tanto la curva no estará “cargada” a ninguno de los extremos, en cuanto al estudio específico de este tipo de distribución, se abordará con mayor profundidad en un capítulo posterior.

Esto significa que la relación numérica de las tres medidas de tendencia central será así:

$$M_o = Md = \bar{X}$$

Por lo tanto, la “cresta” de la curva estará en el centro según se puede observar en la Figura 40:

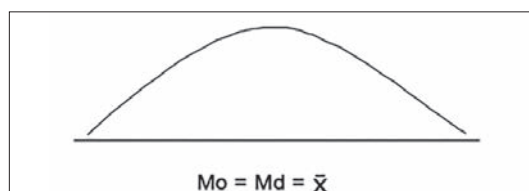


FIGURA40: GRÁFICO DE LA DISTRIBUCIÓN SIMÉTRICA

Lo desarrollado anteriormente se basa en la relación numérica de las Medidas de Tendencia Central y su correspondiente gráfico, pero para determinar de manera más estricta si los valores obtenidos en una variable presentan uno de los tres casos, se debe calcular un coeficiente llamado Coeficiente de Asimetría (c.as.).

El valor y signo que tome este coeficiente determinan si la distribución de los datos se puede identificar como simétrica o asimétrica y en este último caso como positiva o negativa.

En rigor si el coeficiente de asimetría es negativo, la distribución tendrá un sesgo negativo caso contrario el sesgo será positivo y si el valor es cero la distribución no será sesgada, es decir el coeficiente de asimetría puede ser mayor a cero, menor a cero o cero, esto último indicaría una distribución simétrica.

Respecto a lo indicado en el párrafo anterior, he enfatizado que **en rigor** esto debe ser así, pero dado que en la práctica será casi imposible que alguna vez el coeficiente de asimetría sea cero, se considera lo siguiente:

Algunos autores sugieren que si el coeficiente de asimetría toma valores dentro del intervalo $[-0.5 ; 0.5]$ entonces se considera que no hay sesgo y por tanto la distribución no estaría sesgada estrictamente, pero esto es una propuesta como otras pueden ser que el intervalo sea $[-0.2 ; 0.2]$, es decir más exigente dado que se acerca más a cero. Por ejemplo, si en un caso el coeficiente de asimetría ($c.as.$) = -0.35 , entonces aunque el signo indique sesgo negativo se asumirá que no existe sesgo dado que el valor de -0.35 está dentro del intervalo determinado y por tanto se puede considerar que la distribución es simétrica. De todas maneras, esto siempre quedará a criterio del investigador.

El valor de este coeficiente en datos simples, se lo obtiene de Excel con la función *Coeficiente.asimetria*.

Para encontrar el valor del coeficiente de asimetría en datos agrupados podemos aplicar la siguiente fórmula:

$$c.as. = \frac{3*(x - Md)}{s}$$

En donde

x : Media

Md : Mediana

s : Desviación estándar

Fíjese que para que el coeficiente de asimetría sea cero (datos agrupados) los valores de la Media y Mediana deben ser exactamente iguales que es lo que en teoría debería ocurrir para considerar simétrica a la distribución.

De todas maneras, siempre será recomendable hacer un análisis y verificación muy juiciosos sobre esto, dado que las medidas de tendencia central pueden dar una idea de algún tipo de sesgo, pero el coeficiente de asimetría no asegure esto.

Por ejemplo si tuviésemos un caso como el siguiente, ¿qué hacer?:

$$\bar{x} = 7.25 \quad Md = 7.00 \quad Mo = 6.69 \quad c. as. = -0.55$$

Según la relación de las medidas de tendencia central el sesgo se calificaría como positivo, pero al mismo tiempo los valores son muy cercanos y puede considerarse como una distribución simétrica; al mismo tiempo el coeficiente de asimetría por signo y valor nos indica un sesgo negativo, pero como el intervalo indicado anteriormente $[-0.5 ; 0.5]$ no debe tomarse necesariamente con rigidez, se puede decir que se confirmaría la simetría; la decisión entonces estará en el investigador, de todas maneras habrá un respaldo numérico que justifique cualquier camino que se siga al respecto.

CURTOSIS

En lo referente a la Curtosis habíamos dicho que esta distribución tiene que ver con el distanciamiento que puede existir entre los valores extremos de la variable, numéricamente esto depende, por fórmula, de la desviación estándar (medida que la analizaremos más adelante), por tanto, en este punto solo vamos a determinar las distintas expresiones gráficas de la curtosis.

Por definición, la curtosis (también llamada apuntamiento) determina cuán puntiaguda o achatada se presenta la curva simétrica e indica si los datos se encuentran más cerca o distantes de la Media, existen entonces tres tipos de curtosis: Leptocúrtica, Mesocúrtica y Platicúrtica y sus respectivos gráficos son los que se muestran en la Figura 41:

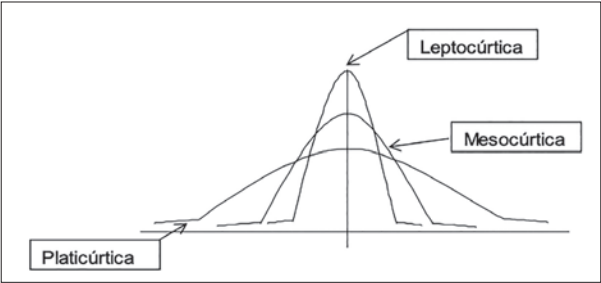


FIGURA 41: GRÁFICAS DE LOS TRES TIPOS DE CURTOSIS

Tanto las Asimetrías como la Curtosis tiene sus respectivos Coeficientes: Coeficiente de Asimetría ($c.a.s$) y Coeficiente de Curtosis (cu), los valores numéricos que puede adoptar el coeficiente de Asimetría resumen de la siguiente manera según la Figura 42:

Para las Asimetrías

	Valor	Interpretación	Gráfico	Significado
Coeficiente de Asimetría (Ca)	> 0	Asimetría Positiva		Los valores de la variable tienden a puntuaciones bajas
	$= 0$	Distribución Normal		Los valores se distribuyen simétricamente alrededor de las medidas de tendencia central
	< 0	Asimetría Negativa		Los valores de la variable tienden a puntuaciones altas

FIGURA 42: CARACTERÍSTICAS DE LAS DISTRIBUCIONES ASIMÉTRICAS Y DE LA SIMETRÍA

La Figura 43 que se presenta a continuación resume los valores numéricos e interpretación en cuanto a la curtosis.

Para la Curtosis

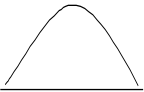


	Valor	Interpretación	Gráfico	Significado
Coeficiente de Curtosis (Cu)	> 0	Leptocúrtica		Gran concentración de datos cerca a la Media
	= 0	Mesocúrtica		Los valores se distribuyen simétricamente alrededor de la Media
	< 0	Platicúrtica		Los valores de la variable tienden a estar más alejados de la Media

FIGURA 43: CARACTERÍSTICAS DE LA DISTRIBUCIÓN DE LOS DATOS SEGÚN ESTÉN MÁS CERCA O LEJOS DE LA MEDIA

Revisemos por ejemplo el 10mo ejercicio propuesto referente a los resultados de una investigación sobre conocimientos de Literatura Universal en un grupo de estudiantes de 2do curso de bachillerato; los valores de las tres medidas de tendencia central y los coeficientes de asimetría y curtosis se calcularán (en esta ocasión) en los datos simples.

3,75	8,75	18,75	24,00	36,12
3,75	9,00	19,00	25,00	36,50
3,75	9,12	20,38	25,00	36,70
4,75	9,88	20,38	26,00	46,00
5,50	10,88	20,62	27,12	46,38
5,75	12,38	21,88	29,38	53,88
6,00	12,88	22,50	31,00	57,50
6,12	14,25	22,62	33,38	59,12
8,25	15,50	23,50	35,00	64,75
8,62	15,88	23,60	35,25	80,50

Media	23,93
Mediana	21,25
Moda	3,75
Coef. Asimetría	1,212
Curtosis	1,390

TABLA 13. MEDIDAS DE TENDENCIA CENTRAL Y COEFICIENTES DE ASIMETRÍA Y CURTOSIS

Según los resultados tanto la Asimetría como la Curtosis tienen signos positivos por lo ello la conclusión sería la siguiente: los datos tienen un sesgo positivo y se puede decir que este grupo de estudiantes tiene una tendencia hacia valores bajos en conocimientos de Literatura Universal; en cuanto a la curtosis podemos decir que hay gran concentración de datos alrededor de la Media, lo cual supone una homogeneidad sobre ese valor que puede calificarse como buena o mala según sobre cuánto se mida la variable, en este caso la variable se mide sobre 100 puntos y esto no da una buena imagen sobre la tendencia del grupo a valores cercanos a 23.93.

En los procesos referentes a comportamiento humano estas dos medidas tienen mucha importancia debido a que la forma en que se distribuyan los datos determinará que las conclusiones del grupo a estudiar tengan mayor o menor impacto en el análisis.

Pero es importante hacer un análisis más profundo especialmente en lo que se refiere al valor del coeficiente de asimetría ya que si bien es cierto el dato numérico habla sobre una aglutinación de datos en valores bajos es importante que el análisis vaya más allá.

La intencionalidad de los procesos estadísticos numéricos no debe quedar en encontrar resultados mediante fórmulas, el estudio de la variable debe también “determinar” si el comportamiento de la variable puede calificarse mediante un juicio de valor, por ejemplo como bueno o malo, en base a los resultados obtenidos y la ética de quien los analice.

En el caso del ejemplo estudiado recordemos que la variable es: “conocimiento de Literatura Universal en un grupo de estudiantes de 2do curso de bachillerato de un determinado centro educativo de nivel medio”.

El ejercicio sí nos revela más información sobre la variable y es que esta prueba se mide sobre 100 puntos por tanto tenemos al menos un parámetro para hacer un mejor análisis.

Las medidas de tendencia central indican valores entre 3 y 23 (en este ejemplo la moda es un valor que distorsiona mucho y en realidad no es muy representativa dado que el valor de 3.75 se repite solo en tres casos, por tanto habrá que hacer un análisis tomando en cuenta las otras dos medidas) esto significa que los valores centrales a analizar serán 21.25 y 23.93; esto nos da ya una idea del grupo y podemos decir que los conocimientos en esta materia son en general muy bajos (recuerde que el valor máximo es 100), con este dato el analista ya puede hacer sugerencias.

En cuanto al coeficiente de asimetría (1.39), éste nos confirma que la tendencia del grupo en este tema hace ver que los conocimientos son muy malos según la prueba aplicada, el signo nos dice que la asimetría es positiva, es decir hay mayor concentración de alumnos con resultados bajos.

Este dato prenderá las alarmas sobre los conocimientos de los alumnos, pero también puede servir para revisar si la prueba aplicada está acorde al nivel de los estudiantes ya que las preguntas pueden exigir conocimientos que para un estudiante de 2do de bachillerato sean muy elevadas.

Si se trabaja con datos agrupados, los resultados serían los indicados en la Figura 44:

<i>L. Inf</i>	<i>L. Sup</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>
3	13	8	17	17
13	23	18	11	28
23	33	28	9	37
33	43	38	6	43
43	53	48	2	45
53	63	58	3	48
63	73	68	1	49
73	83	78	1	50

Media	24,6
Mediana	20,27
Moda	8
Coef. Asimetría	0,73

FIGURA 44: VALORES DE TENDENCIA CENTRAL Y ASIMETRÍA EN DATOS AGRUPADOS

En la Figura 45 se establece de manera gráfica la distribución de los datos.

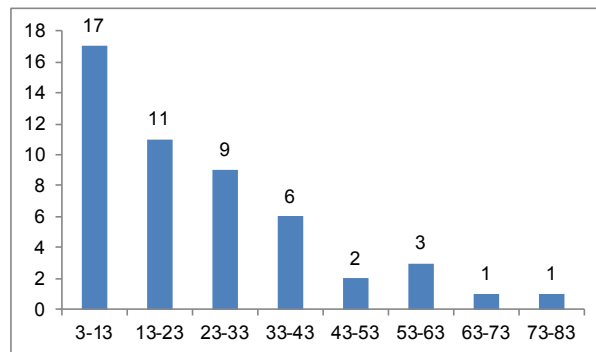


FIGURA 45: GRÁFICO DE BARRAS DEL EJERCICIO

Analíticamente (resultados numéricos) la moda es el valor más bajo y la media el más alto, por tanto, se confirma que existe una asimetría positiva, es decir hay más resultados bajos en la variable estudiada, esto se comprueba con el signo y valor del coeficiente de asimetría y también con el gráfico en el cual se observa una mayor concentración de datos en valores hacia el primer grupo y por tanto la “cola” está hacia la derecha.

Tanto en datos simples como en agrupados se determina el tipo de distribución asimétrica positiva y se comprueba con el gráfico.

Con este sencillo ejemplo he querido determinar que las medidas estadísticas que se encuentren en la resolución de problemas deben “ayudarse” entre sí para que el analista pueda tener una visión más amplia de la situación específica en cuanto al comportamiento de la variable a estudiar.

En cuanto al tema referente a la curva normal, es decir cuando es simétrica, lo trataremos más adelante y a mayor profundidad por la importancia de este específico tipo de distribución y su aplicación en los procesos relacionados en Psicología y Educación y también porque para su estudio es necesario conocer sobre otras medidas complementarias a las de tendencia central que las estudiaremos en el siguiente capítulo.

CAPÍTULO 4:

MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Lo que dijera al iniciar el estudio de las medidas de tendencia central sobre que éstas son algo “miopes” para el análisis estadístico, hemos podido comprobar cuando se trata de concluir sobre temas como los presentados en el segundo ejercicio de los ejemplos del capítulo anterior (“a 81 estudiantes de segundo de básica se les aplicó el test de atención d2 de Rolf Brickenkamp...”), en ese ejercicio se preguntaba lo siguiente: ¿a qué conclusión podría llegar? y con base en ésta ¿cuál o cuáles serían sus recomendaciones? Se determinó entonces que los valores de las medidas de tendencia central son:

Media: 322.75

Mediana: 325

Moda: 312

Pero supongo que le habrá sido difícil decir algo más sobre el grupo estudiado y peor aún haber contestado con total certeza las preguntas formuladas, ya que aunque se establecieron valores referenciales, no sabemos bien cómo se distribuyen esos valores y, por tanto, no podríamos respaldar numéricamente si, por ejemplo, el comportamiento del grupo en esta variable es homogéneo o no, es decir, si existe mucha o poca dispersión en los resultados obtenidos en el grupo y las consecuencias de esto.

Las medidas llamadas de dispersión miden exactamente eso, la homogeneidad o heterogeneidad de los datos lo que nos permitirá conocer con más detalle la verdadera situación del grupo, su distribución general y su distribución comparada con las medidas de tendencia central (especialmente la Media).

Estos nuevos valores incorporados al análisis de datos quitarán aún más la venda que en realidad tienen las medidas de tendencia central si se manejan solas y por lo tanto, el investigador será capaz de proponer más y mejores recomendaciones.

Las medidas de dispersión – también llamadas de variabilidad – expresan exactamente eso, la variación en el comportamiento de los datos dentro de un grupo de valores a estudiar, “Por lo regular, se emplean tres medidas de variabilidad en las ciencias del comportamiento: el rango, la desviación estándar y la varianza” (Pagano, 2011, p. 79). Hago un comentario respecto a lo indicado por este autor, en realidad esas son las medidas de dispersión tradicionales que se estudian en los distintos cursos de Estadística, pero hay otro valor que también mide variabilidad en una muestra y se le conoce como coeficiente de variación que lo revisaremos junto con las otras mencionadas.

Revisemos entonces las definiciones de cada una de ellas y sus características:

Rango: mide la diferencia que existe entre el mayor y el menor de los datos de la muestra a estudiar, es por tanto la medida más simple de calcular, fórmula:

$$R = M - m$$

M : dato Mayor

m : dato menor

Como se puede entender esta medida establece tan solo la lejanía a la que se encuentran los extremos, pero no indica cómo se encuentra el resto de los valores de la muestra.

He recomendado siempre que respecto a la interpretación del rango hay que tener un poco de cuidado si sólo se utiliza el resultado numérico de la resta de los valores extremos, porque puede interpretarse mal, a continuación, un ejemplo sobre esto.

Supongamos que tenemos dos grupos diferentes de personas y en ellos se ha determinado que el rango en cuanto a la edad es el mismo: veinte años; ¿la conclusión inmediata sería que están igualmente dispersos? La respuesta es: “no necesariamente”, ya que puede ocurrir lo siguiente:

	Grupo 1	Grupo 2
Mayor	60	11
Menor	50	1

TABLA 14. RANGO DE EDAD ENTRE DOS GRUPOS DE PERSONAS

Si bien es cierto la diferencia en ambos casos es de 10 años, las características en cuanto a dispersión no son las mismas, ya que las diferencias reales de vida, experiencia, conocimientos y otras tantas son muy distintas para cada grupo ya que 10 años de diferencia entre personas de 50 y 60 no es tan significativa como los mismos 10 años entre personas del segundo grupo, que en realidad no tienen nada en común.

Desviación Estándar: mide la variabilidad (lejanía o cercanía) de los datos alrededor de la media. Si se tomaran varias muestras de la misma población, la desviación estándar será por lo general más estable que el rango, es por ello que se la utiliza con mayor frecuencia en procesos de investigación especialmente para determinar si existen diferencias significativas entre ellas.

Para el cálculo de la desviación estándar de datos simples no presentaré fórmula porque puede calcularse directamente en Excel utilizando la función *desvest*, por ejemplo:

+DESVEST(B1:E10).

La fórmula para datos agrupados es la siguiente:

$$\sqrt{\frac{\sum f * (pm - x)^2}{n - 1}}$$

En donde:

f : frecuencia del intervalo

\underline{pm} : punto medio del intervalo

\bar{X} : promedio de la muestra

n : tamaño de la muestra

En cuanto a la **Varianza** debo indicar que es poco usada en la Estadística Descriptiva ya que esta medida de dispersión es el cuadrado de la desviación estándar y por el hecho de tener unidades cuadradas tiende a confundir su aplicación. El uso de la varianza es más claro y frecuente en la Estadística Inferencial es por estas razones, que tan solo menciono a esta medida de variabilidad y no haremos cálculos ni interpretaciones por estar fuera del alcance de los objetivos de esta sección.

Como dijera unas líneas atrás, la otra medida de dispersión que lastimosamente no se utiliza con la frecuencia requerida es el Coeficiente de variación (c.v.), esta media se encuentra dividiendo el valor de la desviación estándar entre la media de la muestra y para una mejor interpretación se la traduce a porcentaje (multiplicando por 100):

El coeficiente de variación es una medida adimensional que permite determinar la variabilidad de la muestra aunque es más utilizada en la comparación de dos muestras distintas no necesariamente de la misma variable.

En cuanto a su interpretación hay algunas propuestas de sendos autores que difieren unas de otras, a continuación, se presentan (sin ningún tipo de orden) las opciones antes indicadas:

Primera propuesta

COEFICIENTE DE VARIACIÓN	PRECISIÓN
Hasta 10%	Buena
De 11% a 20%	Aceptable
Más de 20%	No confiable

TABLA 15. PRIMERA PROPUESTA PARA INTERPRETAR EL COEFICIENTE DE VARIACIÓN

Segunda propuesta

COEFICIENTE DE VARIACIÓN	PRECISIÓN
< 10%	Poca dispersión
[10% - 33%]	Aceptable
[34% - 50%]	Alta dispersión
>50%	Muy alta

TABLA 16. SEGUNDA PROPUESTA PARA INTERPRETAR EL COEFICIENTE DE VARIACIÓN

Tercera propuesta

En pruebas de laboratorio	
COEFICIENTE DE VARIACIÓN	PRECISIÓN
[0% - 10%]	Bueno
[10% - 15%]	Aceptable
> 15%	Desechable

TABLA 17. TERCERA PROPUESTA PARA INTERPRETAR EL COEFICIENTE DE VARIACIÓN

Cuarta propuesta

En pruebas de campo	
COEFICIENTE DE VARIACIÓN	PRECISIÓN
[0% - 15%]	Bueno
[15% - 25%]	Aceptable
> 25%	Desechable

TABLA 18. CUARTA PROPUESTA PARA INTERPRETAR EL COEFICIENTE DE VARIACIÓN

Como se puede observar los criterios difieren y seguramente habrá otros, por lo tanto, dependerá del investigador aceptar alguno de ellos; la recomendación para esto será que depende de la experiencia y conocimiento que se tenga sobre la variable estudiada y en todo caso lo que parece debe ser tomado en cuenta es, si debe hacerse una diferencia en el caso de que la investigación se haga en el campo o en un laboratorio.

Para iniciar la aplicación y explicación concreta de las medidas de dispersión debo recordar que cuando iniciamos el estudio de las medidas de tendencia central hacía hincapié en las características de cada una de ellas, especialmente entre Media y Mediana, en esa ocasión ponía ejemplos que permitían notar que, de grupos distintos con diferentes valores, el resultado de estas medidas podía ser igual pero las particularidades se notarían con otro tipo de medidas, y éstas son las de dispersión.

Propongo el siguiente ejemplo para hacer notar los errores que se pueden cometer al analizar una variable solo con los datos de las medidas de tendencia central y la diferencia al hacerlo con ayuda de las medidas de dispersión.

Grupo 1

2,5	2,5	15	20
22,5	2,5	20	20
20	5	20	2,5
5	5	22,5	12,5
20	20	2,5	5
2,5	2,5	2,5	20
20	20	20	2,5
2,5	5	20	20
22,5	2,5	20	22,5
2,5	5	20	5

Grupo 2

17,5	12,5	7,5	2,5
5	20	5	17,5
17,5	5	7,5	5
20	9	10	5
2,5	12,5	20	17,5
7,5	15	12,5	20
10	17,5	17,5	10
12,5	5	20	12,5
20	11	10	10
30	5	12,5	5

Para iniciar la comparación en este ejemplo haré referencia inicialmente solo a la Media, el valor de ésta para ambos casos es exactamente 12.0625, pongamos estos valores dentro de un contexto como el siguiente: se ha evaluado la capacidad de concentración en dos grupos de niños entre 6 y 7 años cuyas condiciones socio económicas son iguales.

Si tomamos el valor de la Media como única medida de análisis podríamos concluir que los dos grupos tienen las mismas características y si es que se debe realizar alguna acción de mejora o seguimiento, esa acción deberá ser la misma en ambos grupos debido a sus similares condiciones y además porque su promedio es exactamente el mismo. ¿Esta conclusión y su recomendación se ajustan a la realidad de los grupos?

Analicemos un poco más al detalle los datos y al calcular las otras medidas de tendencia central encontramos lo siguiente:

	Grupo 1	Grupo 2
Mediana	13,75	11,75
Moda	20	5

TABLA 19. OTROS VALORES DE TENDENCIA CENTRAL

Por lo pronto y con estos nuevos datos podemos estar claros ya que la utilización de la Media como única medida de análisis es un error. ¿Qué nos dice la Mediana? La interpretación inicial de esta medida nos indica que en el primer grupo el 50% de los datos tienen

valores inferiores a 13.75 y el otro 50% superó ese valor; y respecto al segundo grupo el 50% de los niños está por debajo de valores a 11.75 y el resto sobrepasa el valor señalado; por lo pronto, ¿cuál grupo consideraría usted que está mejor?

Respecto a la Moda es clara la diferencia, si comparáramos los grupos solo en base a este valor, la observación sería evidente e inmediata: en el primer grupo hay un conjunto de niños (no necesariamente la mayoría) que tiene mejores niveles de concentración y alguien podría asegurar que por tanto el primer grupo es mejor y más aún si se determina que de los 40 niños, en el primer grupo 15 tienen esta característica y en el segundo solo 8 representan la Moda.

De todos modos, hasta este momento del análisis todavía podríamos tener discrepancia en la evaluación general de los grupos.

Calculemos entonces las medidas de dispersión:

	Grupo 1	Grupo 2
Rango	20	27,5
Desviación estándar	8,53	6,33

TABLA 20. MEDIDAS DE DISPERSIÓN

Estos valores ya nos dicen algo más sobre el comportamiento de la variable en cada grupo; como se podrá apreciar en el segundo grupo hay mayor diferencia entre el mayor y menor de los datos (rango), ¿esto nos dice que hay mayor diferencia entre la capacidad de atención de los niños del segundo grupo que de los del primero? Pues de cierto modo sí, pero solo nos indica que esa diferencia es mayor comparando a los niños de mayor y menor capacidad de cada grupo y no nos dice más sobre el resto de ellos.

En cuanto a la desviación estándar, estos valores hacen referencia a la dispersión de los datos alrededor de la Media y no solo sobre los extremos, por los resultados se puede ver que en el segundo grupo hay menos dispersión que en el primero, por lo tanto según esta medida el grupo 2 es más homogéneo, en otras palabras y aplicando la definición de desviación estándar, en el segundo grupo hay mayor cantidad de niños que se encuentran más cercanos respecto de la Media que en el primero, pero no deja de haber dispersión.

Como hemos podido notar según se calculaban las distintas medidas tanto de tendencia central como de dispersión, no solo que se encontraban más elementos de juicio respecto a los dos grupos, si no que de pronto esas apreciaciones iban cambiando según los particulares resultados e interpretaciones.

Se presenta entonces un resumen de los resultados obtenidos durante este proceso:

	Grupo 1	Grupo 2
Media	12,06	12,06
Mediana	13,75	11,75
Moda	20	5
Rango	20	27,5
Desviación estándar	8,53	6,33

TABLA 21. VALORES DE TENDENCIA CENTRAL Y DISPERSIÓN

Aquí algunas puntualizaciones:

i. En cuanto a las medidas de tendencia central

1. Como ya se dijera que puede ocurrir al estudiar las medidas de tendencia central, aquí se determina que los valores de Media y Mediana son cercanos entre sí cosa que no ocurre con la Moda
2. El valor de la Moda sí es preocupante en el segundo grupo ya que representa el 20% de la muestra.
3. Según la relación numérica de las medidas de tendencia central en el primer grupo se puede establecer que: $\bar{X} < Md < Mo$, por lo tanto podemos decir que el primer grupo tiene un sesgo negativo, no así en el segundo en el cual se establece que la relación numérica es así: $Mo < Md < \bar{X}$ es decir tiene un sesgo positivo.
4. Si calculamos el coeficiente de asimetría tenemos lo siguiente para cada grupo

	Grupo 1	Grupo 2
Coefficiente de Asimetría	-0,0385	0,5145

TABLA 22. COEFICIENTE DE ASIMETRÍA DE CADA GRUPO

Esto confirma lo indicado en el numeral anterior respecto al tipo de sesgo para cada grupo, por lo tanto, gráficamente el comportamiento de la variable se presentaría, para cada grupo, según aparece en la Figura 46:

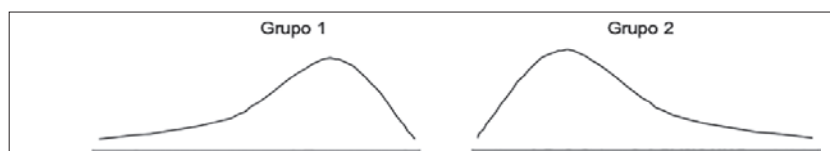


FIGURA 46. GRÁFICA DE LA DISTRIBUCIÓN DE FRECUENCIAS DE CADA GRUPO

Se deduce entonces que la tendencia en el primer grupo es hacia valores más altos de la variable y que al contrario en el segundo grupo la tendencia es hacia valores bajos.

¿Qué nos dice esta primera parte? ¿El primer grupo de niños tiene mejor capacidad de concentración que el segundo? Pues al parecer esa sí sería la conclusión a la que deba llegarse.

ii. En cuanto a las medidas de dispersión

Según el rango hay menor diferencia entre los extremos del primer grupo que de los extremos del segundo, esto se da básicamente porque en el segundo grupo hay **un** niño que obtuvo 30 puntos en la prueba, esto se deberá tomar en cuenta para la conclusión y recomendación final, (cuando en un grupo de datos ocurre que hay un valor que está “fuera de tono” se dice que existe un “valor inusitado”, ante esto, el investigador deberá determinar – en base a un análisis de la distribución de los valores, si eliminar o no dicho dato).

Este análisis de la distribución se ejemplifica en la Figura 47:

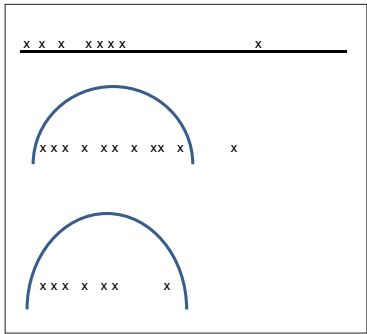


FIGURA 47. IDENTIFICACIÓN GRÁFICA DE UN VALOR PARA SABER SI ES FORTUITO O NO

Para el primer caso el valor extremo debe averiguarse si es inusitado o no, en el segundo caso se puede decir que sí es un valor inusitado, pero en el último caso, el valor debe considerarse como una observación importante debido a que está dentro de la curva esperada.

1. La desviación estándar dice que el segundo grupo es más homogéneo en la distribución de los datos alrededor de la Media, esto ayuda para establecer que este grupo seguramente es más fácil de manejar que el primero.

De todas maneras, los valores tanto del rango como de la desviación son altos, lo cual indica que los dos grupos están significativamente dispersos; cabe señalar que en problemas relacionados con las ciencias del comportamiento, es muy normal que el investigador encuentre que los datos están dispersos, esto se debe a las grandes diferencias que pueden existir en un grupo humano por más afinidades sociales que existan en la población.

2. Si calculamos la curtosis de cada grupo tenemos

	Grupo 1	Grupo 2
Curtosis	-1,974	-1,339

TABLA 23. COEFICIENTE DE CURTOSIS PARA CADA GRUPO

Ambos valores indican una distribución similar de los datos en cuanto a la distancia (dispersión) de los extremos tomando una curva de forma Platicúrtica. Esto confirma lo dicho en el numeral anterior.

En cuanto al valor del Coeficiente de variación tenemos lo siguiente:

	Grupo 1	Grupo 2
Media	12,06	12,06
Desviación estándar	8,53	6,33
Copeficiente de variación	70,73%	52,49%

TABLA 24. MEDIDAS DE DISPERSIÓN Y COEFICIENTE DE VARIACIÓN PARA AMBOS GRUPOS

Es notorio que en ambos grupos la dispersión es muy alta (cualquiera sea la interpretación escogida según páginas atrás) y, haciendo la comparación entre ambos, el primer grupo es mucho más disperso que el segundo; observación que confirma lo anotado con anterioridad.

A manera de revisión también desarrollemos el ejercicio con datos agrupados. Se decidió hacer el ejercicio con una amplitud de 2.5 puntos, si usted cambia la amplitud tan solo a 3, los resultados son muy distintos; es por ello la importancia de decidir objetiva y argumentativamente el valor de la amplitud en cada variable por analizar, la Figura 48 a continuación representa el cuadro de datos agrupados con la amplitud descogida.

Grupo 1

Capacidad de concentración niños de 6 a 7 años		pm	f	fa
$L. Inf.$	$L. Sup.$			
2,5	5	3,75	12	12
5	7,5	6,25	7	19
12,5	15	13,75	1	20
15	17,5	16,25	1	21
20	22,5	21,25	19	40

FIGURA 48. DATOS AGRUPADOS CON AMPLITUD 2,5

Media	13,063
Mediana	15
Moda	21,25
Rango	20
Desviación estándar	8,243
Coeficiente de Asimetría	-0,705

TABLA 25. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN EN DATOS AGRUPADOS

La Figura 49 es la representación gráfica de los datos agrupados del grupo 1

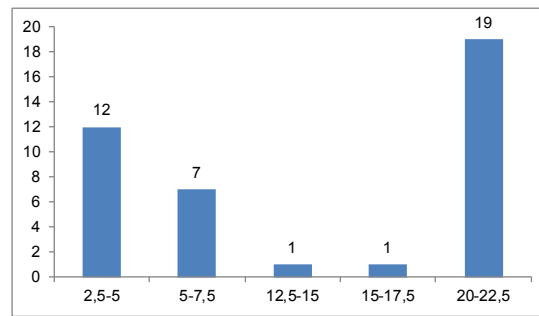


FIGURA 49: GRÁFICO SEGÚN AGRUPACIÓN DE 2,5 PUNTOS

Como puede verse, al agrupar los datos los resultados de las medidas de tendencia central y dispersión han sufrido cambios es decir se ha perdido información, analicemos esto:

- Respecto a los valores de las medidas de tendencia central y dispersión la variación de los resultados no es significativa
- El coeficiente de asimetría sí tiene un cambio significativo en lo referente a su valor absoluto ya que el signo nos sigue indicando un sesgo negativo.

Grupo 2

Para el segundo grupo se utilizó la misma amplitud según aparece en la Figura 50:

Capacidad de concentración niños de 6 a 7 años		<i>pm</i>	<i>f</i>	<i>fa</i>
<i>L. Inf.</i>	<i>L. Sup.</i>			
2,5	5	3,75	2	2
5	7,5	6,25	8	10
7,5	10	8,75	4	14
10	12,5	11,25	6	20
12,5	15	13,75	6	26
15	17,5	16,25	1	27
17,5	20	18,75	6	33
20	22,5	21,25	6	39
27,5	30	28,75	1	40

FIGURA 50: DATOS AGRUPADOS CON AMPLITUD 2,5

La Tabla 26 resume los cálculos descriptivos de este grupo:

Media	13,19
Mediana	12,5
Moda	6,25
Rango	27,5
Desviación estándar	6,19
Coefficiente de Asimetría	0,33

TABLA 26. RESULTADOS DE LAS MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN EN DATOS AGRUPADOS

La Figura 51 indica gráficamente la distribución de los datos del segundo grupo:

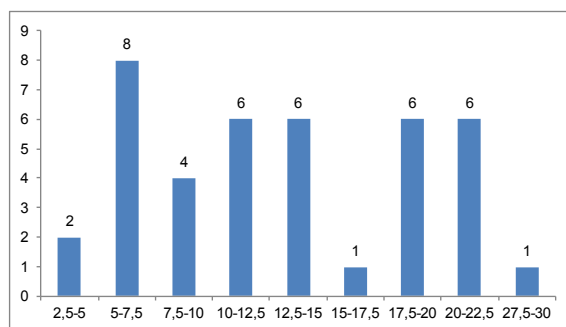


FIGURA 51. GRÁFICO SEGÚN LA AGRUPACIÓN DE 2,5 PUNTOS

Al igual que lo ocurrido con el grupo 1 se pierde información y exactitud al agrupar los datos. He decidido mantener la misma amplitud que en el primer grupo para una menor distorsión en la comparación de los grupos.

Para este grupo los valores tanto de medidas de tendencia central como de dispersión y el coeficiente de asimetría se mantienen más estables que con el grupo anterior y las diferencias no son significativas. En cuanto al gráfico, éste no nos da luces para determinar una concentración de datos a los extremos y tampoco podemos decir que su comportamiento tiene una definida tendencia en cuanto a agrupación se refiere.

En referencia a la variabilidad de los datos en cada grupo encontramos lo siguiente:

	Grupo 1	Grupo 2
Media	13,06	13,19
Desviación estándar	8,24	6,19
Coeficiente de variación	0,631	0,469

TABLA 27. COMPARACIÓN DE LAS MEDIDAS DE DISPERSIÓN ENTRE LOS DOS GRUPOS

Los valores de dispersión (*c.v.*) en datos agrupados también son elevados y en general indican mucha variabilidad (aunque en ambos casos ésta haya disminuido su valor).

Todo lo indicado anteriormente son observaciones de los resultados de cada grupo estudiado y se han realizado algunas comparaciones entre los grupos, pero no he llegado a ninguna conclusión y menos aún he propuesto recomendaciones; esto lo dejo al lector que según su experiencia pueda hacerlo.

Lo que sí recomiendo es que cuando se haga el análisis de uno o más grupos en forma comparativa o no, se realicen todos estos cálculos para tener un panorama más amplio con el cual poder decidir con mayor objetividad.

¿Recuerda usted el ejercicio de los dos paralelos A y B en la materia de Química segundo año de bachillerato y otras condiciones más cuando se estudiaba la relación entre las tres

medidas de tendencia central? En ese momento se preguntaba si tomaría usted entonces las mismas medidas pedagógicas y psicológicas para estos grupos.

En ese ejemplo el promedio para ambos grupos fue exactamente igual (6.5238) y se preguntaba si dado este caso las acciones a tomar deberían ser las mismas.

En esas circunstancias la respuesta lógica es que sí dado que no se conocen otras medidas y – lastimosamente – es un ejemplo de lo que en realidad pasa en muchas circunstancias parecidas especialmente en las instituciones educativas ¡de cualquier nivel!

Pero si se hiciera el análisis tomando en cuenta tan solo las medidas hasta ahora estudiadas, seguramente la decisión no será la misma, veamos lo que ocurre en cada caso:

	PARALELO A	PARALELO B
Promedio	6,524	6,524
Mediana	9	6
Moda	10	6
Rango	9	5
Desviación Estándar	3,655	1,750

TABLA 28. COMPARACIÓN DE LAS MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN ENTRE LOS DOS PARALELOS

De acuerdo a los resultados de las medidas de dispersión, al paralelo “A” se lo puede calificar como heterogéneo y al paralelo “B” como Homogéneo.

En cuanto a las medidas de forma y dispersión:

	PARALELO A	PARALELO B
Medidas de forma	Sesgo negativo	Distribución Simétrica
Coef. Asimetría	-0,493	1,280
Coef. Variación	0,560	0,273

TABLA 29. COMPARACIÓN ENTRE LAS MEDIDAS DE FORMA Y EL COEFICIENTE DE VARIACIÓN ENTRE LOS DOS PARALELOS

Tomando en cuenta las medidas de tendencia central y de dispersión; se puede observar que la distribución de las notas en el paralelo B es homogénea; esto significa, por ejemplo, que no hay grandes diferencias de conocimientos y por tanto concluir que ese paralelo tiene el mismo nivel de conocimientos pero lastimosamente alrededor de 6 que es bajo, también se puede ver que el 50% de las notas en este paralelo es superior solo a 6 puntos, en cambio en el paralelo A el 50% supera el 9.

Desde otro punto de análisis utilizando el tipo de distribución de las notas, el paralelo A presenta un sesgo negativo lo que refleja una tendencia a notas más altas, esto por la influencia de la mitad de las notas mayores a 9; pero también es un grupo heterogéneo, con esto se puede concluir que será más difícil trabajar con este grupo debido a las extremas diferencias

y seguramente la decisión será tomar medidas de ayuda tan solo al 50% bajo; en cambio en el paralelo B habrá que trabajar con la gran mayoría.

Fíjese también que, según los coeficientes de variación, estos confirman la mayor homogeneidad del paralelo B respecto al otro pero esto no significa que estén mejor dado que lo homogéneo del grupo se aglutina alrededor de un bajo promedio.

Si formamos grupos con dos puntos de amplitud, obtendremos un cuadro como el de la Figura 52:

Notas paralelo A	Nro. De alumnos por grupo	Notas paralelo B	Nro. De alumnos por grupo
1 - 2	6	5 - 6	15
3 - 4	2	7 - 8	3
7 - 8	2	9 - 10	3
9 - 10	11		
Total general	21		21

FIGURA 52: AGRUPACIÓN EN CADA PARALELO CON AMPLITUD DE 2 PUNTOS

En forma gráfica se verá según lo representado en la Figura 53:

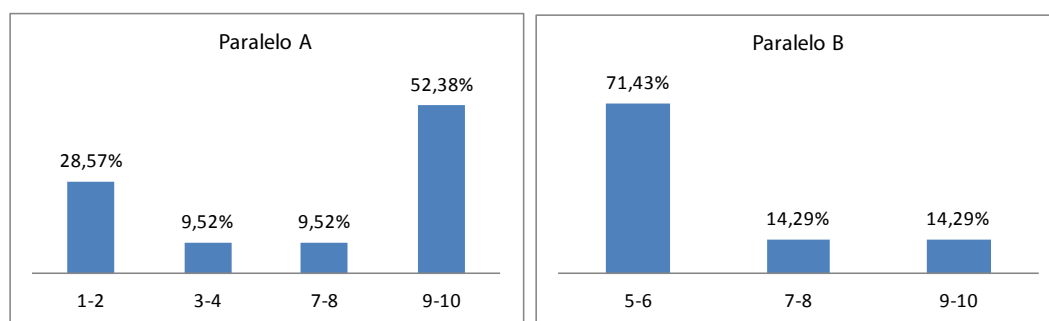


FIGURA 53: COMPARACIÓN DE LOS GRÁFICOS SEGÚN DISTRIBUCIÓN DE CADA GRUPO

Ahora el análisis que se puede realizar es mucho más completo y el gráfico, aunque no decida nada como tal, tendrá algo que “decir” ¿no le parece?

EJERCICIOS DE APLICACIÓN DEL CAPÍTULO

Ejemplo 1

Los siguientes datos corresponden a los resultados de un test de motricidad (medido sobre 150) tomado al personal de taller en una empresa de fabricación de juguetes; determinar la variable, encontrar las medidas de tendencia central y de dispersión, establecer el tipo de sesgo comparando la relación entre las medidas de tendencia central y comprobando este resultado con el coeficiente de asimetría, establezca la variabilidad de los datos, según su criterio ¿cómo calificaría a la distribución de esta muestra?

136	137	137	137	137	138	138	139	139	140
140	140	141	141	141	141	141	142	142	142
142	142	142	142	143	143	144	144	144	144
145	145	146	147	147	147	147	147	148	149
149	149	149	150	150					

Solución (datos simples)

Variable: Niveles de motricidad del personal de taller en una empresa de juguetes.

Media	142,98
Mediana	142
Moda	142
Desviación estándar	3,98
Rango	14
Coefficiente de asimetría	0,14
Coefficiente de Variación	2,78%

TABLA 30. VALORES DESCRIPTIVOS DE LA VARIABLE A ESTUDIAR

Relación entre las medidas de tendencia central:

$$Mo = Md < \bar{x}$$

Por lo tanto y de manera estricta no podemos determinar un tipo de sesgo, pero se puede considerar objetivamente que los valores de las medidas de tendencia central son prácticamente iguales, por tanto es válido decir que la distribución de los datos es simétrica, lo confirma esto el valor del coeficiente de asimetría.

Haciendo un análisis de la distribución de los datos se puede determinar que el grupo estudiado tiene una distribución no sesgada (c.a. 0.14) y que las medidas de tendencia central indican que sus resultados son altos dado que el puntaje máximo es de 150, por tanto, se puede establecer que el grupo presenta un buen nivel de desarrollo en motricidad.

En referencia a las medidas de dispersión, tanto el rango como la desviación estándar indican homogeneidad dado que el valor de la desviación estándar (3.98) comparado con los datos de la muestra es muy pequeño, lo mismo ocurre con el rango ya que 14 puntos de diferencia entre el menor y mayor no es significativo para el caso de esta variable.

En cuanto al coeficiente de variación (2.78%) el valor es mínimo y por tanto se confirma la homogeneidad en la muestra estudiada; esto ya se esperaba según lo descrito en el análisis de las medidas de dispersión.

La Figura 54 presenta la solución (datos agrupados)

Valores de motricidad del personal de una empresa de juguetes		<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
<i>L. Inf.</i>	<i>L. Sup.</i>					
136	138	137	7	7	15,56%	15,56%
139	141	140	10	17	22,22%	37,78%
142	144	143	13	30	28,89%	66,67%
145	147	146	8	38	17,78%	84,44%
148	150	149	7	45	15,56%	100,00%

FIGURA 54: TABLA DE DATOS AGRUPADOS CON AMPLITUD 3

Media	142,87
Mediana	143,27
Moda	143
Desviación estándar	3,89
Rango	14
Coefficiente de asimetría	-0,31
Coefficiente de variación	2,72%

TABLA 31. VALORES DESCRIPTIVOS SEGÚN DATOS AGRUPADOS

La Figura 55 indica la distribución gráfica de los datos.

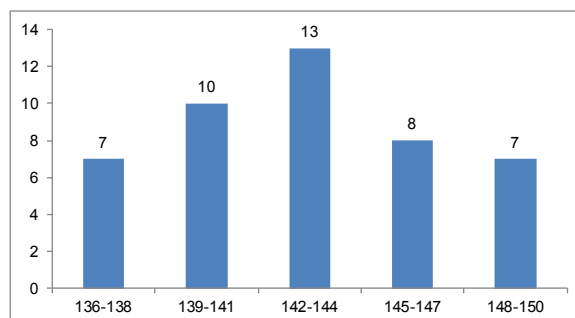


FIGURA 55: GRÁFICO DE BARRAS DE LOS DATOS AGRUPADOS

La relación entre las medidas de atención central es la siguiente:

$$\bar{X} < Mo < Md$$

Estos valores no cumplen estrictamente una de las dos formas de distribución estudiadas, pero dado que los valores son muy cercanos también sugieren una distribución simétrica confirmada a su vez por el coeficiente de asimetría y por el gráfico.

La diferencia entre el desarrollo con datos simples y el de agrupados radica básicamente en el signo del coeficiente de asimetría, pero esto no hace diferencia en la aseveración de que en ambos casos la distribución es simétrica debido a la condición establecida en páginas anteriores.

En cuanto a la “calificación” de la distribución, se supone que es lo esperado ya que teóricamente las variables referentes a comportamiento humano suelen presentar simetría

en su distribución de datos; en este caso además de haber presentado una simetría, los valores medios están muy cerca del valor máximo esperado, lo cual confirma que los niveles de motricidad son elevados en el grupo. Cabe anotar sin embargo que en análisis grupales normalmente se encontrarán individuos con valores que pueden considerarse preocupantes y a quienes habrá que realizar seguimientos.

Lo indicado para las medidas de dispersión y el coeficiente de variación en los datos simples, también se aplica para el caso de datos agrupados.

Ejemplo 2

Se encontraron los siguientes datos respecto en los resultados obtenidos en una prueba psicológica que mide capacidad de abstracción, aplicada a 60 personas entre 20 y 40 años de edad. Esta prueba tiene como valores mínimo y máximo 4 y 45 respectivamente. Determinar la variable, encontrar las medidas de tendencia central y de dispersión, establecer el tipo de sesgo comparando la relación entre las medidas de tendencia central y comprobando este resultado con el coeficiente de asimetría, según su criterio ¿cómo calificaría a la distribución de estos datos?

4	19	18	10	5	10	6	16	18	20
9	16	15	4	10	18	19	15	4	6
10	7	19	20	10	6	10	5	18	4
5	12	23	25	14	9	10	7	33	40
28	35	18	9	21	32	19	15	10	35
20	12	25	30	18	35	16	20	28	10

Solución (datos simples)

Variable: capacidad de abstracción

Media	16,08
Mediana	15,5
Moda	10
Rango	36
Desv. Est.	9,17
Coef. Asime.	0,73

TABLA 32. VALORES DESCRIPTIVOS EN DATOS SIMPLES

Solución (datos agrupados) según se establece en la Figura 56:

Capacidad de abstracción	<i>pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
4	10	7	24	24	40,00%
11	17	14	9	33	15,00%
18	24	21	16	49	26,67%
25	31	28	5	54	8,33%
32	38	35	5	59	8,33%
39	45	42	1	60	1,67%

FIGURA 56: TABLA DE DATOS AGRUPADOS CON AMPLITUD 7

Media	16,45
Mediana	15,67
Moda	7
Rango	41
Desv. Estándar	9,72
Coef. Asimetría	0,24

TABLA 33. VALORES DESCRIPTIVOS EN DATOS AGRUPADOS

En forma gráfica la distribución se presenta según lo indicado en la Figura 57:

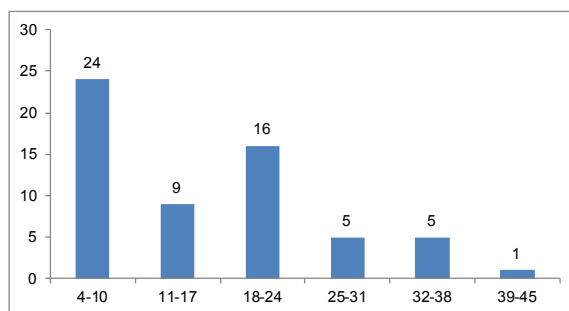


FIGURA 57: GRÁFICO DE BARRAS DE LOS DATOS AGRUPADOS

Revisando los resultados tanto en datos simples como agrupados podemos decir no existen variaciones significativas entre una u otra forma salvo un poco en la Moda.

Comparando los valores de las medidas de tendencia central se puede determinar con claridad lo siguiente:

$$Mo < Md < \bar{X}$$

Esto significa que, en lo referente a la prueba mencionada, la tendencia en este grupo es hacia valores bajos de la variable (asimetría positiva) por tanto en esta ocasión se podría “calificar” como mala a esta distribución de los datos (¿por qué?), esto también se puede observar en el gráfico y lo confirma el signo en el coeficiente de asimetría.

Pero ¿por qué calificarla así? Debemos recordar que el evaluar los resultados de alguna variable depende de algunos factores, entre ellos: la variable en sí, el criterio del investigador y la lógica de la variable; en este caso estamos hablando de una prueba que mide capacidad de

abstracción y si los resultados indican una agrupación hacia valores bajos, lo lógico es decir que la tendencia de los resultados nos dice que el grupo estudiado no tiene un buen desarrollo en esta capacidad y más aún si son personas (20 – 40 años) de las cuales se espera que ya hayan desarrollado a plenitud esta capacidad.

A esto hay que sumar que la variable se mide entre 4 y 45 y si damos un vistazo a los datos simples (se recomienda ordenarlos), podemos ver que el 80% de la muestra ha obtenido valores inferiores a la **mitad del valor máximo**, esto da más luces sobre lo que ocurre con el grupo en esta variable.

Todos estos comentarios son en base a los datos proporcionados en el ejercicio y hasta allí pueden llegar los criterios emitidos ya que, en la práctica, el investigador debe recabar más información como por ejemplo las características del grupo, las circunstancias en las que se desarrolló la prueba, el nivel socio económico de la población y otras tantas características que le permitirán establecer más y mejores juicios sobre la muestra estudiada y por ende sobre la población.

Ejemplo 3

En un análisis que desea determinar el nivel de desarrollo motor en niños de hasta 2 años, se aplicó una prueba a 90 niños obteniéndose los siguientes puntajes.

9	5	9	8	4	1	7	1	5
10	8	5	7	1	2	10	4	4
8	4	8	8	10	7	9	9	1
7	9	8	5	9	10	5	8	6
8	8	7	8	6	8	7	9	8
9	6	6	9	8	2	3	7	10
10	8	4	10	5	3	1	5	2
8	5	8	8	2	1	3	7	1
6	7	6	10	3	1	2	1	3
7	9	9	7	5	3	10	2	2

Esta variable se mide entre 0 y 10 puntos donde 10 es el valor óptimo. Determinar lo siguiente:

1. Medidas de Tendencia Central y Dispersión en datos simples
2. Medidas de Tendencia Central y Dispersión en datos agrupados (amplitud 2)
3. Determine el sesgo tanto en datos simples como agrupados
4. Indique si para usted esta distribución de frecuencias es buena o mala
5. Gráfico. ¿El gráfico confirma lo establecido en el numeral anterior?
6. Establezca lo siguiente: un comentario, una conclusión y una recomendación

Solución datos simples

Variable: Nivel de desarrollo motor en niños de hasta 2 años

Resultados:

Media	6,04
Mediana	7
Moda	8
Rango	9
Desv. Estándar	2,89
Coef. Asimetría	-0,41
Sesgo	Negativo

TABLA 34. VALORES DESCRIPTIVOS EN DATOS SIMPLES

La Figura 58 representa la solución datos agrupados.

Valores de desarrollo motor en niños de 0 a 2 años		<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
<i>L. Inf.</i>	<i>L. Sup.</i>					
1	2	1,5	16	16	17,78%	17,78%
3	4	3,5	11	27	12,22%	30,00%
5	6	5,5	15	42	16,67%	46,67%
7	8	7,5	28	70	31,11%	77,78%
9	10	9,5	20	90	22,22%	100,00%

FIGURA 58: TABLA DE DATOS AGRUPADOS CON AMPLITUD 2

Media	6,06
Mediana	7,21
Moda	7,5
Rango	9
Desv. Estándar	2,81
Coef. Asimetría	-1,24
Sesgo	Negativo

TABLA 35. VALORES DESCRIPTIVOS EN DATOS AGRUPADOS

En forma gráfica se vería según la Figura 59:

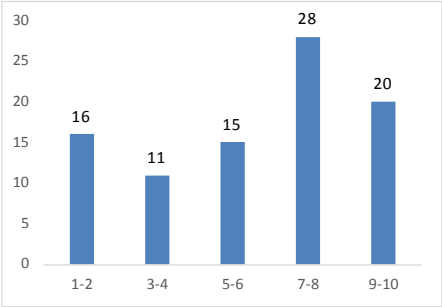


FIGURA 59: GRÁFICO DE BARRAS DE LOS DATOS AGRUPADOS

Al elaborar los distintos cuadros presentados se ha respondido a las preguntas 1 a 3, en cuanto a la pregunta 4 (Indique si para usted esta distribución de frecuencias es buena o mala) revisemos lo siguiente: se trata de estudiar el desarrollo motor en niños de 2 a 4 años, por lo tanto lo ideal sería que los resultados apunten a valores altos de la variable (supongo que todos queremos que nuestros niños tengan los mejores puntajes en esto) pero tampoco estaría mal si la distribución fuese no sesgada (es decir simétrica) dado que es una variable de comportamiento humano.

El resultado tanto en datos simples como agrupados indica que la distribución tiene un sesgo negativo, esto significa que en este grupo de niños hay una tendencia hacia valores altos de desarrollo motor lo cual es bueno, fíjese también que el gráfico (pregunta 5) sí da una idea de esa tendencia.

En cuanto a la pregunta 6 podríamos decir:

Comentario: el grupo de niños tiene niveles altos de desarrollo lo cual es bueno.

Conclusión: los padres y maestros de estos niños han sabido estimular positivamente el desarrollo motor de sus hijos.

Recomendación:

1. Se sugiere a los padres de familia de estos niños mantener actividades que estimulen en sus hijos su desarrollo motor pero que no presionen sobre esto, dado que tienen algunos años para que alcancen su “madurez” en esta variable.
2. Se sugiere a los padres de los niños con valores inferiores a la Media (es la medida de tendencia central más baja), contactar a un profesional o pedir en la institución educativa, que ayude a sus hijos a mejorar los niveles de desarrollo motor.

Ejemplo 4

El siguiente ejemplo trata también de dar una aplicación de la relación entre un valor concreto de la variable, la Media y la Desviación Estándar especialmente para diferenciar a un sujeto específico de la muestra en relación al grupo estudiado. Para ello voy a introducir tres conceptos:

Puntuación directa: es el valor obtenido o atribuido a un determinado elemento de la muestra o población en el estudio de alguna variable. Simbólicamente se expresa con “ x ” o “ x_i ”

Puntuación diferencial es el resultado de restar la puntuación directa con el promedio (Media).

$$P. \text{ dif: } x_i - \bar{X}$$

Puntuación típica es el resultado de dividir la puntuación diferencial entre la desviación estándar. Su resultado es adimensional y representa la distancia (dada en desviaciones estándar) de un valor con respecto a la media aritmética.

$$P. \text{ típica: } z_i = (x_i - \bar{X}) / s$$

Supongamos que se trata de estudiar el nivel de memoria en un grupo de estudiantes; dentro del grupo se escoge un alumno al azar y se establece que dicho alumno ha obtenido un valor de 15, ¿qué podremos decir de su nivel de memoria y especialmente de su “ubicación” respecto al grupo? ¿Será alto, bajo o aceptable? Pues nada ya que no tenemos referencia alguna para comparar.

Debemos conocer al menos el valor del promedio del grupo, digamos que sea 12 entonces podremos calcular su Puntuación Diferencial:

$$\text{P. dif: } x_i - \bar{x}$$

$$\text{P. dif: } 15 - 12 = 3$$

Como es positiva este alumno puede quedar tranquilo ya que se comprueba que tiene un nivel de memoria superior al promedio; pero ¿esto indica que está muy alejado o no respecto a la Media? Supongamos que la Desviación Estándar es muy pequeña (digamos 0.5) entonces este valor de 3 indicará que está muy alejado del promedio y representará un valor alto y puede interpretarse como un estudiante de gran capacidad de memoria; caso contrario, si la Desviación Estándar es grande (digamos 2.5), significa que está muy cerca del promedio y entonces no se puede decir que se destaca por su nivel de retención.

Entonces ¿cómo se puede interpretar esta Puntuación Diferencial para saber si el resultado es alentador o no? Para ello debemos recurrir entonces a la Desviación Estándar ya que de ella depende la verdadera relación de este alumno respecto al grupo.

Para el caso voy a determinar dos situaciones: $S_1 = 2$ y $S_2 = 4$; calculando las Puntuaciones Típicas para cada caso tendremos:

$$\text{P. típica1} = 3 / 2 = 1.5$$

$$\text{P. típica1} = 3 / 4 = 0.75$$

En el primer caso nuestro alumno podrá decir que pocos compañeros le superan en su nivel de memoria, si fuera el segundo caso, no podrá entonces jactarse ante sus condiscípulos.

Veamos otro caso, en dos materias distintas (supongamos Matemáticas y Química) un estudiante obtiene 5 y 6 respectivamente (calificadas sobre 10), es obvio pensar que en Química le fue mejor, pero aumentemos información; la Media del curso en Matemáticas es 4.5 y en Química es 5.5 y calculando las Puntuaciones Diferenciales tendremos:

$$\text{P. dif. (Matemáticas): } 5 - 4.5 = 0.5$$

$$\text{P. dif. (Química): } 6 - 5.5 = 0.5$$

Esto solo nos dice que está por encima del promedio en ambas materias lo cual para sus intereses es bueno, pero ¿cómo estará respecto a sus compañeros?

Es necesario entonces conocer el valor de la desviación estándar para ambas materias, digamos que en Matemáticas es 0.3 y en Química es de 1.2 y calculemos entonces las puntuaciones típicas para cada caso:

$$\text{P. típica (Matemáticas)} = 0.5 / 0.3 = 1.66$$

$$\text{P. típica (Química)} = 0.5 / 1.2 = 0.41$$

La conclusión inicial de que en Química le fue mejor sigue siendo válida, pero comparado con sus compañeros, hay menos estudiantes que le superaron en Matemáticas y por tanto está mejor que muchos.

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

1. Se desea saber si existe una clara diferencia entre sujetos educados en niveles socioeconómicos bajos (B) de los educados en niveles socioeconómicos altos (A), (según Figura 60) en base a un test que mide la capacidad de análisis. Valores mínimo y máximo esperados: [65 ; 110] y valores medios esperados: [90 ; 95]

VARIABLE	n_A	VARIABLE	n_B
70-74	3	70-74	3
75-79	13	75-79	15
80-84	36	80-84	32
85-89	40	85-89	110
90-94	58	90-94	15
95-99	20	95-99	9
100-104	18	100-104	10
105-109	12	105-109	6

FIGURA 60: DATOS YA AGRUPADOS DE LOS GRUPOS A COMPARAR

Determine lo siguiente para cada grupo:

- i. Valor de la amplitud
 - ii. Medidas de Tendencia Central y Dispersión.
 - iii. Determine el sesgo (asimetría), indique si al comparar el sesgo dado por la relación de las Medidas de Tendencia Central y el calculado por fórmula, el criterio de asimetría se mantiene
 - iv. Indique si para usted la distribución de frecuencias se la puede calificar como buena o mala
 - v. Gráfico. ¿El gráfico confirma lo establecido en el numeral anterior?
 - vi. Establezca lo siguiente: un comentario, una conclusión y una recomendación
2. Los datos presentados en la Figura 61 son los resultados de evaluación de una misma materia calificada sobre 100 y en la cual se debe obtener al menos 60 para su aprobación. En una promoción determinada las notas de los estudiantes de dos grupos semejantes se distribuyeron del siguiente modo

Variable	A	B
0-13	10	16
14-27	25	24
28-41	60	50
42-55	140	84
56-69	35	76
70-83	19	30
84 o más	11	20

FIGURA 61: DATOS AGRUPADOS SEGÚN LA DISTRIBUCIÓN EN CADA GRUPO A ESTUDIAR

Para cada grupo determine lo siguiente:

- i. Medidas de Tendencia Central y Dispersión
 - ii. Indique si para usted esta distribución de frecuencias se la puede calificar como buena o mala
 - iii. ¿Aproximadamente qué porcentaje de estudiantes aprobaría el curso?
 - iv. Gráfico. ¿El gráfico confirma lo establecido en el numeral anterior?
 - v. Establezca lo siguiente: un comentario, una conclusión y una recomendación
3. A continuación se proporcionan los valores obtenidos por 100 jefes de personal al aplicárseles un test que mide el autoritarismo (se mide sobre 150 puntos), determine lo siguiente:
- i. Variable
 - ii. Medidas de Tendencia Central y Dispersión en datos simples
 - iii. Medidas de Tendencia Central y Dispersión en datos agrupados (se sugiere amplitud de 16)
 - iv. Determine el sesgo tanto en datos simples como agrupados, compare los resultados
 - v. Indique si para usted esta distribución de frecuencias se la puede calificar como buena o mala
 - vi. Gráfico. ¿El gráfico confirma lo establecido en el numeral anterior?
 - vii. Establezca lo siguiente: un comentario, una conclusión y una recomendación

14	28	42	46	51	55	58	65	84	98
14	28	43	46	51	55	60	65	85	112
15	28	44	46	51	55	60	66	88	114
15	29	44	47	52	56	61	67	88	115
15	29	44	48	52	56	61	67	89	115
16	31	45	49	54	56	61	68	89	116
16	31	45	49	54	57	61	68	90	121
16	31	45	50	54	57	62	68	90	130
16	31	46	50	54	57	64	69	90	130
16	32	46	51	55	57	64	69	91	132

4. Los datos de la Figura 62 corresponden a puntajes de una prueba de atención obtenidos por personas que pretendían trabajar en una empresa del País. Se conoce que el rango de dicha prueba está entre 1 mínimo y 100 máximo y que los valores considerados como esperados están en el intervalo [65 , 80].

PUNTAJE	f	PUNTAJE	f	PUNTAJE	f
25 - 29	1	50 - 54	12	75 - 79	59
30 - 34	1	55 - 59	27	80 - 84	55
35 - 39	3	60 - 64	23	85 - 89	36
40 - 44	10	65 - 69	56	90 - 94	20
45 - 49	3	70 - 74	59	95 - 99	2

FIGURA 62: DATOS YA AGRUPADOS DEL EJERCICIO PLANTEADO

Determinar lo siguiente:

- Medidas de Tendencia Central y Dispersión
 - Indique si para usted esta distribución de frecuencias se la puede calificar como buena o mala
 - ¿Aproximadamente qué porcentaje de personas están dentro del intervalo esperado?
 - ¿Qué porcentaje de personas superan el valor máximo esperado?
 - Gráfico. ¿El gráfico confirma lo establecido en el numeral anterior?
 - ¿Considera que es un grupo con altos niveles de atención?
5. Se presentan en la tabla siguiente los resultados de las notas de ingreso a una determinada facultad, dicha prueba se mide sobre 100 puntos y para ingresar se necesita un valor superior a 75.

51	87	70	95
84	84	95	81
88	85	81	71
71	89	85	85
82	96	81	76
81	90	92	87
88	83	61	64
76	90	74	55
85	88	82	89
76	64	66	71
90	96	52	56
90	66	68	84
91	51	58	89
90	64	60	88
91	95	62	86

Determinar lo siguiente:

- Variable
- Medidas de Tendencia Central y Dispersión en datos simples
- Medidas de Tendencia Central y Dispersión en datos agrupados (elabore el cuadro completo con amplitud 5)
- Determine el tipo de distribución tanto en datos simples como agrupados, compare los resultados
- Indique si para usted esta distribución de frecuencias se la puede calificar como buena o mala

- vi. En datos agrupados, si se ofrece una segunda oportunidad para quienes estén en un grupo inmediatamente inferior a la condición de ingreso se les da una nueva oportunidad, determinar el porcentaje de estudiantes que ganarían este derecho; ¿cuál sería su apreciación de cada sub grupo y del grupo en general?
6. En una investigación se trató de medir la capacidad de adaptación a nuevos ambientes sociales a personas que hayan superado los 45 años de edad. El instrumento valora los resultados en una escala de 60 a 140 puntos, este último valor indicaría total capacidad de adaptación. Los siguientes datos se obtuvieron de un grupo de profesionales entre hombres y mujeres.

81	75	97	112	122	113
86	68	119	77	117	117
116	105	107	87	112	92
86	100	107	82	92	87
118	105	109	112	112	122
91	105	132	102	112	92
96	105	127	107	92	109
101	105	117	102	97	109
126	75	112	92	107	112
114	70	102	102	117	112
112	75	119	112	77	122
115	105	104	87	109	77
116	110	130	106	117	107
116	110	121	107	112	117
102	115	117	112	111	87

- Medidas de Tendencia Central y Dispersión en datos simples
 - Medidas de Tendencia Central y Dispersión en datos agrupados (amplitud 10)
 - Determine el sesgo tanto en datos simples como agrupados, compare los resultados
 - Indique si para usted esta distribución de frecuencias se la puede calificar como buena o mala
 - Indique si en este grupo existe un problema grave o no en cuanto a la capacidad de adaptación
 - Determine la puntuación típica de una persona que haya obtenido 105 en esta investigación
 - Proponga actividades para quienes necesiten algún refuerzo según su criterio
7. Los datos a continuación son el resultado de una prueba que mide los conocimientos básicos en el manejo del paquete Office a profesores de una institución educativa. (valores de medición: 10 a 60). Con ellos encuentre las medidas de tendencia central en datos agrupados (amplitud 5) y responda las preguntas conociendo que se esperan resultados superiores a 40 en tendencia central y al menos un 60% de profesores con resultados superiores a 45. ¿Qué opinión le merecen los datos?, ¿qué conclusión obtiene?, ¿qué recomendaría para los casos que se presenten según sus cálculos?

32	50	56	45	53	17
37	51	56	20	53	50
25	51	16	45	54	30
39	15	57	46	54	14
39	51	21	18	12	51
39	25	57	46	54	22
23	51	18	47	20	53
40	52	57	47	55	20
41	19	57	18	57	54
41	52	58	47	57	11

8. Los datos siguientes corresponden a los resultados de una investigación que trataba de determinar el nivel de sentimiento de seguridad en jóvenes universitarios con una escala cualitativa entre 10 y 50. Construya con estos datos 8 intervalos

44	50	21	14	44	49	47	49	28
15	47	19	36	19	17	31	13	38
11	40	22	50	50	47	45	17	45
33	15	22	34	23	45	24	43	35
35	50	25	45	21	18	24	44	29
50	16	21	23	42	17	30	40	19
42	41	49	50	32	17	28	32	18
17	21	41	41	45	48	14	15	42
19	19	18	41	46	44	19	17	35
47	50	39	35	35	35	48	17	11
29	25	42	13	31	13	43	45	34
33	42	33	49	14	50	30	44	18
21	33	48	45	14	19	22	38	33
38	40	43	36	28	13	36	11	32
16	48	38	13	45	13	16	27	14
18	25	25	45	50	33	22	42	28
39	33	40	30	13	35	49	23	35
42	46	29	33	36	45	37	45	27
19	33	48	19	38	46	25	11	43
44	33	35	30	33	43	11	29	43

Determine tanto en datos simples como agrupados lo siguiente:

- Medidas de Tendencia Central y dispersión
 - Elabore un gráfico (datos agrupados)
 - ¿Según el valor del coeficiente de asimetría lo consideraría como un tipo de distribución bueno?
 - Haga un análisis general calculando todas las medidas descriptivas desarrolladas hasta el momento
9. Los valores siguientes (Figura 63) corresponden a los resultados de una prueba de habilidad para un grupo de personas que desean un trabajo de destreza manual. Se conoce que el valor máximo de esta prueba es 80 y el mínimo 10 y que para conseguir el trabajo se exige un puntaje superior a 50.

VARIABLE	f
13.6 - 22.6	3
22.6 - 31.6	4
	22
	2
	1
	5
	3

FIGURA 63: TABLA PARA COMPLETAR LOS DATOS AGRUPADOS DEL EJERCICIO

- i. Complete el cuadro
 - ii. Determine la amplitud de los intervalos
 - iii. Encuentre las Medidas de Tendencia Central y Dispersión
 - iv. Elabore un gráfico e interprételo
 - v. Cómo considera la distribución de los datos
 - vi. Encuentre la puntuación típica de una persona que obtuvo 35 en dicha prueba y compárela con otra que sacó 25
 - vii. Comente los resultados
10. Complete (si es posible) los datos presentados en la Figura 64. Si no es posible hacerlo, indique la razón. La variable se refiere a notas de Psicología Social valoradas sobre 30 puntos. Determine medidas de tendencia central y dispersión y haga un análisis de la situación de las notas en la materia en cuestión.

VARIABLE	pm	f	fa
12.4 - 15.4		2	
			5
			9
			11
			16

FIGURA 64: TABLA DE DATOS AGRUPADOS PARA COMPLETAR

11. Se encontraron los siguientes datos respecto a los resultados en la resolución de una prueba de conocimientos básicos de Matemáticas aplicada a 30 estudiantes que pretendían ingresar a 8vo. de básica en una institución educativa. Esta prueba se mide sobre 20 puntos y podrán ingresar quienes superen un puntaje de al menos el 80% del total. Realizar el análisis correspondiente en datos agrupados con amplitud 3 puntos y determinar cuántos candidatos aprobarían y el porcentaje que estos representarían.

4	13	17	19
9	15	13	12
18	16	4	5
17	14	4	18
12	18	12	13
15	10	5	19
14	11	20	14
19	18	18	16
14	5	6	17
17	16	6	15
16	6	7	14
15	10	10	16
11	9	17	15
15	18	10	12
20	15	12	15

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1

- Capacidad de análisis de personas con distintos niveles socioeconómicos
- Amplitud: 5 puntos
- Grupo 1

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
70	74	72	3	3	1,50%	1,50%
75	79	77	13	16	6,50%	8,00%
80	84	82	36	52	18,00%	26,00%
85	89	87	40	92	20,00%	46,00%
90	94	92	58	150	29,00%	75,00%
95	99	97	20	170	10,00%	85,00%
100	104	102	18	188	9,00%	94,00%
105	109	107	12	200	6,00%	100,00%

FIGURA 65: TABLA DE DATOS AGRUPADOS CON AMPLITUD 5

Media	90,225
Mediana	90,690
Moda	92
Rango	39
Desv. Estándar	8,112
Coef. Asimetría	-0,057
Coef. Variación	0,090

TABLA 36. RESULTADOS DESCRIPTIVOS

Grupo 2

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
70	74	72	3	3	1,50%	1,50%
75	79	77	15	18	7,50%	9,00%
80	84	82	32	50	16,00%	25,00%
85	89	87	110	160	55,00%	80,00%
90	94	92	15	175	7,50%	87,50%
95	99	97	9	184	4,50%	92,00%
100	104	102	10	194	5,00%	97,00%
105	109	107	6	200	3,00%	100,00%

FIGURA 66: TABLA DE DATOS AGRUPADOS CON AMPLITUD 5. INTERVALO MEDIANA RESALTADO

Media	87,4
Mediana	87,27
Moda	87
Rango	39
Desv. Estándar	6,68
Coef. Asimetría	0,02
Coef. Variación	0,08

TABLA 37. RESULTADOS DESCRIPTIVOS EN DATOS AGRUPADOS

- iv. Grupo 1: Tipo de asimetría: por el signo es negativa, por el valor (-0.057) se puede decir que es simétrica; tomando en cuenta la relación de las tres medidas de tendencia central: $\bar{x} < Md < Mo$, esto confirma el tipo de relación, pero también sus valores son muy cercanos y por tanto se puede decir que existe simetría.
- Grupo 2: Tipo de asimetría: por el signo es positiva y por el valor (0.019) se puede decir que es simétrica; tomando en cuenta la relación de las tres medidas de tendencia central: $Mo < Md < \bar{x}$, esto confirmaría el tipo de relación, pero de igual manera sus valores son muy cercanos, por tanto, se puede decir que existe simetría.
- v. Grupo 1: dado que la variable se refiere a la capacidad de análisis, el tipo de asimetría sí es bueno, dado que indicaría una mayor concentración de individuos hacia valores altos. Grupo 2: dado que la variable se refiere a la capacidad de análisis, el tipo de asimetría es malo, dado que indicaría una mayor concentración de individuos hacia valores bajos. Pero si se toma en cuenta las medidas de tendencia central, en ambos casos se puede decir que no hay una marcada tendencia y por tanto están en las mismas condiciones, pero si es así de todas maneras el primer grupo presenta valores más altos que el segundo.

vi. Grupo 1

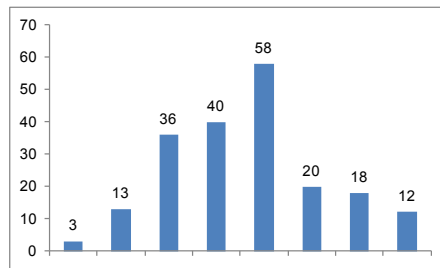


FIGURA 67: GRÁFICO DE BARRAS GRUPO 1

Aunque no es totalmente objetivo, se podría decir que sí confirma el tipo de asimetría, dado que la “cola” parece estar hacia la izquierda

Grupo 2

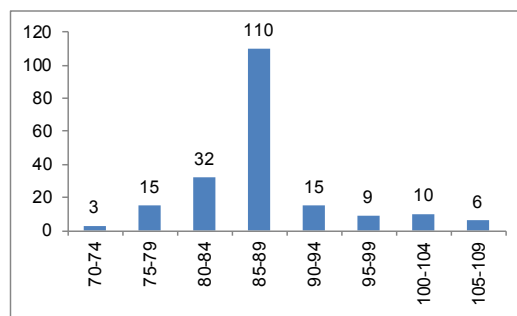


FIGURA 68: GRÁFICO DE BARRAS GRUPO 2

Este grupo presenta gráficamente una marcada distribución simétrica.

vii. Este ítem se deja al lector para su análisis y discusión.

Ejercicio 3

- Variable: medición de autoritarismo en jefes de personal
-

Media	58,17
Mediana	55
Moda	16
Rango	118
Desv. Estándar	27,89
Coef. Asimetría	0,75
Coef. Variación	0,48

TABLA 38. RESULTADOS DESCRIPTIVOS EN DATOS SIMPLES

i.

<i>L. Inf.</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
14	29	21,5	15	15	15,00%	15,00%
30	45	37,5	13	28	13,00%	28,00%
46	61	53,5	39	67	39,00%	67,00%
62	77	69,5	13	80	13,00%	80,00%
78	93	85,5	10	90	10,00%	90,00%
94	109	101,5	1	91	1,00%	91,00%
110	125	117,5	6	97	6,00%	97,00%
126	141	133,5	3	100	3,00%	100,00%

FIGURA 69: TABLA DE DATOS AGRUPADOS CON AMPLITUD 16

Media	58,62
Mediana	55,03
Moda	53,5
Rango	127
Desv. Estándar	27,65
Coef. Asimetría	0,13
Coef. Variación	0,47

TABLA 39. RESULTADOS DESCRIPTIVOS DATOS AGRUPADOS

ii. Tanto en datos simples como agrupados el sesgo es positivo tomando en cuenta las medidas de tendencia central y el valor del coeficiente de asimetría.

Salvo los valores de la moda y coeficiente de asimetría, todos los resultados son similares.

iii. La distribución de los datos presenta un sesgo positivo lo cual indica una tendencia de aglutinamiento a valores bajos de la variable; esto se puede interpretar como bueno dada la característica de la variable.

iv.

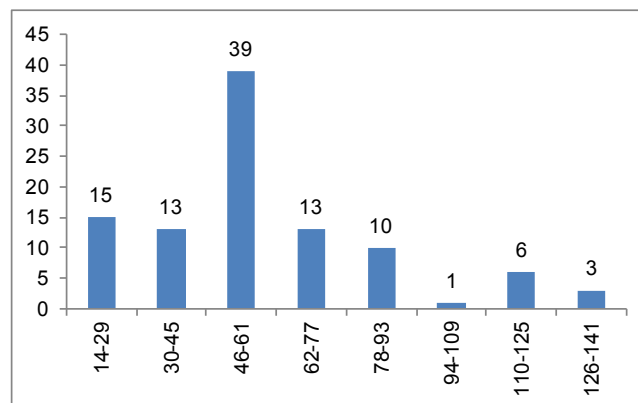


FIGURA 70: GRÁFICO DE BARRAS SEGÚN DATOS AGRUPADOS

El gráfico sí confirma el tipo de distribución ya que se advierte una mayor concentración en valores bajos de la variable.

v. Comentario: los jefes de personal parece ser que sí controlan su carácter.

Conclusión: se podría decir que los jefes de área han desarrollado la habilidad de un manejo democrático del personal a su cargo.

Recomendación: se sugiere a aquellos jefes de personal que tuvieron valores superiores a las medidas de tendencia central, tomar en cuenta los resultados para que pongan más empeño en mejorar este rasgo.

Ejercicio 5

i. Variable: Conocimientos de aspirantes a una facultad

ii.

Media	78,77
Mediana	83,5
Moda	90
Rango	45
Desv. Estándar	12,86
Tipo distribución	As. Negativa
Calificación	BUENA
Coef. Asimetría	-0,70

TABLA 40. RESULTADOS DESCRIPTIVOS EN DATOS SIMPLES

iii.

<i>L. Inf</i>	<i>L. Sup.</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
51	55	53	4	4	6,67%	6,67%
56	60	58	3	7	5,00%	11,67%
61	65	63	5	12	8,33%	20,00%
66	70	68	4	16	6,67%	26,67%
71	75	73	4	20	6,67%	33,33%
76	80	78	3	23	5,00%	38,33%
81	85	83	14	37	23,33%	61,67%
86	90	88	15	52	25,00%	86,67%
91	95	93	8	60	13,33%	100,00%

FIGURA 71: TABLA DE DATOS AGRUPADOS CON AMPLITUD 5

Media	78,75
Mediana	83,5
Moda	88
Rango	44
Desv. Estándar	12,31
Tipo Distribución	As. Negativa
Calificación	Buena
Coef. Asimetría	-1,16
Coef. Variación	0,16

TABLA 41. RESULTADOS DESCRIPTIVOS EN DATOS AGRUPADOS

- iv. El porcentaje de estudiantes que pueden acceder a una nueva oportunidad es de 6.67% es decir quienes están en el intervalo $[71 - 75]$

Ejercicio 7

<i>L. Inf</i>	<i>L. Sup</i>	<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
11	15	13	4	4	6,67%	6,67%
16	20	18	9	13	15,00%	21,67%
21	25	23	5	18	8,33%	30,00%
26	30	28	1	19	1,67%	31,67%
31	35	33	1	20	1,67%	33,33%
36	40	38	5	25	8,33%	41,67%
41	45	43	4	29	6,67%	48,33%
46	50	48	7	36	11,67%	60,00%
51	55	53	15	51	25,00%	85,00%
56	60	58	9	60	15,00%	100,00%

FIGURA 72: DATOS AGRUPADOS SEGÚN INSTRUCCIÓN

Media	40,08
Mediana	46,71
Moda	53
Rango	49
Desv. Estándar	15,87
Tipo Distribución	As. Negativa
Calificación	BUENA

TABLA 42. RESULTADOS DESCRIPTIVOS EN DATOS AGRUPADOS

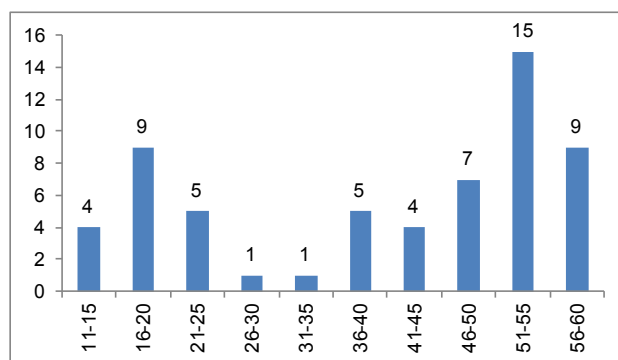


FIGURA 73: GRÁFICO DE BARRAS DE DATOS AGRUPADOS AMPLITUD 5

Estrictamente este grupo de profesores sí cumple con tener valores centrales superiores a 40, pero en cuanto a que al menos el 60% haya obtenido una calificación superior a 45 no cumple, dado que solo el 51.67% lo logra.

Aunque el tipo de asimetría sea negativa y como tendencia esto es bueno, los resultados se pueden considerar bajos, además hay un considerable grupo (18 profesores) con resultados entre 11 y 25 y obviamente se nota una clara diferencia (dispersión) en conocimientos de Office.

La recomendación es obvia para los 18 profesores será necesario exigir tomen un curso para mejorar sus conocimientos de las herramientas de office.

Ejercicio 9

Capacidad en destreza manual		Pm	f	fa	fr	fra
L. Inf.	L. Sup.					
13,6	22,6	18,1	3	3	7,50%	7,50%
22,6	31,6	27,1	4	7	10,00%	17,50%
31,6	40,6	36,1	22	29	55,00%	72,50%
40,6	49,6	45,1	2	31	5,00%	77,50%
49,6	58,6	54,1	1	32	2,50%	80,00%
58,6	67,6	63,1	5	37	12,50%	92,50%
67,6	76,6	72,1	3	40	7,50%	100,00%

FIGURA 74: CUADRO COMPLETO DE DATOS AGRUPADOS CON AMPLITUD 9

Media	40,83
Mediana	36,92
Moda	36,1
Rango	63
Desv. Estándar	14,70
Tipo distribución	SIMÉTRICA
Calificación	MALA

TABLA 43. RESULTADOS DESCRIPTIVOS DE LA VARIABLE

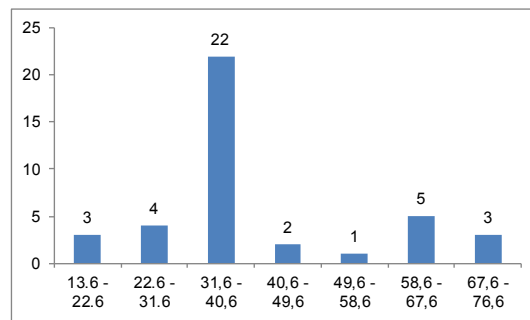


FIGURA 75: GRÁFICO DE BARRAS DATOS GRUPADOS DEL EJERCICIO 9

Puntuaciones típicas de individuos con puntajes de 25 y 35

	Puntuación diferencial	Puntuación típica
Puntaje directo		
25	-15,83	-1,08
35	-5,83	-0,40

TABLA 44. PUNTUACIÓN TÍPICA SEGÚN CONDICIONES DADAS

Estos resultados indican que ambos sujetos están por debajo de la media y que esos 10 puntos de diferencia en verdad representan una diferencia de casi tres veces los niveles de conocimiento entre uno y otro.

Ejercicio 11

Notas de Psicología Social	<i>pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
12,4 - 15,4	13,9	2	2	12,50%	12,50%
15,4 - 18,4	16,9	3	5	18,75%	31,25%
18,4 - 21,4	19,9	4	9	25,00%	56,25%
21,4 - 24,4	22,9	2	11	12,50%	68,75%
24,4 - 27,4	25,9	5	16	31,25%	100,00%

FIGURA 76: CUADRO COMPLETO DE DATOS AGRUPADOS DEL EJERCICIO 11

Media	20,84
Mediana	20,65
Moda	25,9
Rango	15
Desv. Estándar	4,34
Coef. Asimetría	0,04
Coef. Variación	0,21

TABLA 45. VALORES DESCRIPTIVOS DEL EJERCICIO

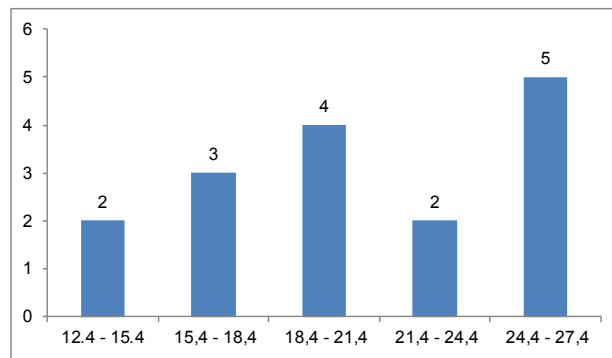


FIGURA 77: GRÁFICO DE BARRAS DEL EJERCICIO CON AMPLITUD 3

Tomando en cuenta los resultados de las tres medidas de tendencia central, se puede notar lo siguiente: los valores entre ellas no permiten determinar que el tipo de distribución sea simétrica y de acuerdo a la relación entre ellas tampoco podemos determinar un tipo de sesgo específico; sin embargo de que el coeficiente de asimetría indique un valor que se puede considerar dentro de lo aceptado para la simetría, pero el gráfico en este caso no corrobora esto y más bien sugeriría un sesgo negativo.

Si nos fijamos en la figura 76, es notorio que los estudiantes se agrupan en valores altos según los resultados (dos últimos intervalos) y estos representan el 43.75% (siete estudiantes de dieciséis), esto ayudaría a sostener que sí existe un sesgo negativo.

CAPÍTULO 5:

MEDIDAS DE POSICIÓN

¿Por qué es importante establecer la posición en la que se encuentra un determinado individuo respecto a un grupo de estudio? En Psicología en Pedagogía y otras ciencias muchas veces se establecen condiciones que deben cumplir los individuos, éstas pueden ser por ejemplo requisitos para ser calificado como instructor, para ingresar a la universidad, para aprobar un curso, para calificar como proveedor, para ingresar o ganar un concurso, para pertenecer a un determinado grupo según alguna condición establecida, para conocer si presenta o no algún trastorno, etc.

En Ciencias de la Educación puede ayudar para saber si un estudiante o grupo tienen problemas de aprendizaje al comparar sus resultados con un valor mínimo, también para determinar si el nivel de rendimiento está dentro de los parámetros establecidos por alguna institución educativa, para establecer condiciones de permanencia en la institución, entre otros.

En psicología estas medidas de posición tienen mucha aplicación en varios ámbitos, así tenemos por ejemplo que para procesos de evaluación de desempeño se puede exigir superar un porcentaje de calificación determinado, para ser considerado apto en una determinada destreza se debe obtener un valor mínimo en un test, para calificar como superada una prueba será necesario haber obtenido cierto valor que garantice tener los requisitos mínimos.

Cada uno de los ejemplos y situaciones descritas establece una partición en una determinada muestra o grupo de estudio que divide a quienes cumplen o no cumplen la condición dada y de eso se encargan las medidas de posición, precisamente de diferenciar dentro del grupo de estudio quiénes cumplen o no un determinado requisito.

En los distintos cuadernillos de los test psicológicos, por ejemplo, se encuentra la explicación y desarrollo estadístico que se siguió para publicar dicha prueba y entre las diferentes tablas que se presentan se encuentra aquella que permitirá comparar el resultado de la aplicación individual de ese test con lo esperado en la teoría y eso determinará la posición alcanzada del individuo en esa prueba específica frente a una población de estudio.

Por tanto, cuando un valor establece el cumplimiento de una condición por parte de un individuo, éste se ubicará en determinada posición respecto al resto del grupo, dicha ubicación especificará un porcentaje de personas que se encuentran **por debajo de él**. He resalado esto, porque la condición fundamental para la interpretación de este tipo de medidas es esa, cuando se encuentra un valor de la variable, ese dato indicará un porcentaje de sujetos que se encuentran por debajo de dicho valor y por tanto permite identificar la ubicación exacta respecto al grupo de estudio.

LAS MEDIDAS DE POSICIÓN

Hasta aquí se ha determinado, para el análisis de una variable, estadísticos de tendencia central, de dispersión y de forma que han permitido establecer criterios sobre el comportamiento de cada variable y situación específica según las condiciones dadas en cada caso.

Las Medidas de Posición permiten establecer lo que se conoce como “una partición de la muestra” en base a una condición dada, esta condición establecerá dos grupos concretos en la variable: aquellos que cumplen la condición y los otros que no; lo importante de estas medidas es la interpretación específica que para cualquier análisis siempre es la misma sin interesar la variable ni los valores sobre los que se midan, esta interpretación indicará siempre el porcentaje de elementos que se encuentran por debajo del valor encontrado (es importante insistir y resaltar esto).

Las Medidas de Posición a estudiar son las siguientes: Percentiles (Centiles), Cuartiles, Deciles y Quintiles; cada una de ellas constituye porciones iguales de la muestra, establecidas en porcentaje, que cumplen una determinada condición y cuya interpretación indica que desde un determinado valor hacia abajo la muestra cumple la condición dada.

Es obvio pensar que los datos estarán ordenados de menor a mayor, aunque para esto en los distintos paquetes estadísticos u hojas de cálculo no es necesario hacerlo.

La diferencia entre una y otra Medida de Posición solo radica en los porcentajes en los cuales se ha dividido la muestra, así:

Los percentiles (P_x) indican divisiones cada una de 1%; en este caso sí es posible calcular percentiles con decimales, esto tiene algunas ventajas en casos en que se quiera hacer cálculos con más precisión

Los cuartiles son particiones de 25%, es decir en la práctica sólo hay tres cuartiles que gráficamente se establecerían según lo establecido en la Figura 78:

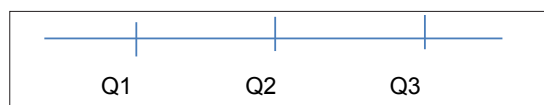


FIGURA 78: REPRESENTACIÓN GRÁFICA DE LOS CUARTILES

Es obvio suponer que lo que sería el Q_4 significa el valor más alto, por ello no tiene sentido calcularlo; de esta forma la muestra queda dividida en cuatro partes cada una ha establecido regiones de 25%, la interpretación sería que el 25% de los datos están por debajo del valor correspondiente al Q_1 ; de esta misma manera el Q_3 indicaría que el 75% (tres partes de 25% cada una) de los datos está por debajo del valor correspondiente; como caso especial está el Q_2 dado que por lógica este valor equivaldrá al de la Mediana dado que ambos parten a la muestra en dos partes iguales correspondientes al 50%, esto significa que la Mediana también es una medida de posición.

En cuanto a los Deciles, estos establecen 9 divisiones en la muestra y por tanto determinarán 10 sectores cada uno de 10% del total; aunque menos usadas tienen la misma interpretación que las otras, es decir y como ejemplo, el decil 7 (D_7) indicará que el 70% de la muestra se encuentra por debajo de ese valor.

Lo mismo ocurre con los Quintiles que representan la quinta parte de la muestra, es decir habrá cuatro Quintiles cada uno de 20% y dividiendo a la muestra en cinco partes, su interpretación es la misma que los anteriores y su aplicación está hoy más difundida en temas de educación, aunque históricamente se aplicó siempre a temas de Economía.

Tanto para percentiles, deciles y quintiles sí se pueden hacer cálculos con decimales, esto tiene la ventaja de poder calcular valores que necesiten más precisión; no ocurre lo mismo con los cuartiles ya que son valores enteros.

A continuación, un par de ejemplos en base a varias condiciones establecidas. Cabe sugerir que para cada análisis en particular se haga un gráfico simple que ayude a determinar qué parte de la muestra cumple la condición dada.

Ejemplo 1

Supongamos que en la materia de Literatura de un grupo de estudiantes se establecen las siguientes condiciones:

- Aquellos alumnos que estén por debajo del primer cuartil deberán tener clases de recuperación
- Quienes superen el tercer cuartil participarán en un intercolegial
- Quienes estén entre el decil 2 y el Quintil 3 elaborarán un ensayo para mejorar su nota
- Se hará un reconocimiento especial a los alumnos que superen el percentil 85

Los datos son los siguientes:

10	7	4	3	5	5	9	6	9	7
3	6	6	10	6	9	6	9	3	9
10	4	8	8	8	7	10	7	4	6
4	9	7	9	4	6	4	8	7	7
6	3	6	5	10	4	7	4	5	6
4	8	3	6	4	8	6	9	9	8
8	5	5	7	9	5	5	9	10	4

Para la primera condición se puede hacer el cálculo del Q_1 o del P_{25} dado que representan la misma posición, esto se indica en la Figura 79.

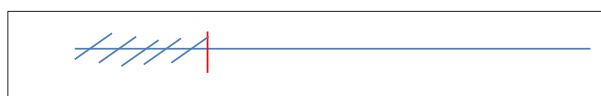


FIGURA 79: GRÁFICO PARA DETERMINAR LA REGIÓN QUE ESTABLECE EL Q_1 O P_{25}

Gráficamente cumplen la condición todos aquellos que se encuentre por debajo del punto en rojo que será el valor a calcular.

En Excel se debe buscar la función “cuartil” y en el cuadro de diálogo se establece el rango de datos (matriz) y el valor del cuartil (para este caso “1”), el resultado es “5” (cinco será ese punto rojo señalado en el gráfico) y la interpretación sería que el 25% de los estudiantes ha obtenido valores inferiores a “5”.

Fórmula en Excel: =+CUARTIL(A1:J7;1)

Resultado de la fórmula: 5, esto se muestra en el cuadro de diálogo según la Figura 80:

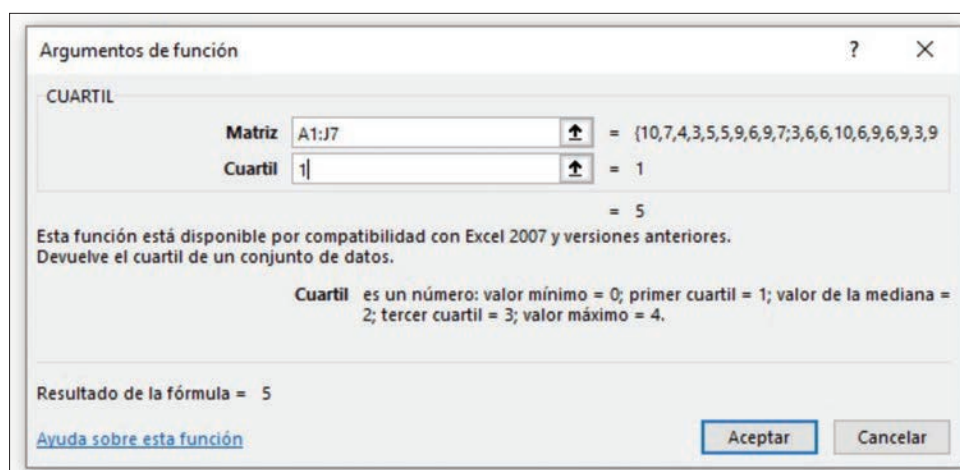


FIGURA 80: CAPTURA DE PANTALLA QUE INDICA EL CÁLCULO DEL PRIMER CUARTIL EN EXCEL

Para la segunda condición la expresión gráfica de la condición sería la siguiente:

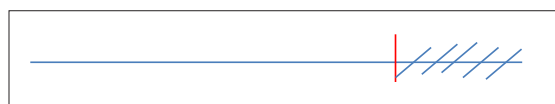


FIGURA 81: GRÁFICO PARA REPRESENTAR LA REGIÓN QUE SE SOLICITA SEGÚN LA CONDICIÓN

Para este caso y según la condición dada: “Quienes superen el tercer cuartil participarán en un intercolegial”

Gráficamente cumplen la condición todos aquellos que se encuentre por encima del punto en rojo que será el valor a calcular, pero debido a la explicación e interpretación teórica el valor que se encuentre siempre nos dirá el porcentaje de individuos que se encuentran por debajo. Por tanto, el resultado habrá que interpretarlo según la condición.

Hay que calcular el tercer cuartil o el percentil 75, con el mismo proceso anterior y encontramos que el valor correspondiente al Q_3 es “8” (este valor es el punto rojo del gráfico) y la interpretación será que el 75% de los estudiantes ha obtenido notas por debajo de “8”; esto significa que participarán en el intercolegial aquellos alumnos que tengan 8 o más (sección rayada del gráfico) dado que ellos superan el valor correspondiente al tercer cuartil.

Fórmula en Excel: `=+CUARTIL(A1:J7;3)`

Resultado de la fórmula: 8

Si esta misma condición se calcula utilizando el percentil 75 (que equivale en porcentaje al tercer cuartil), el resultado lógicamente será el mismo.

Fórmula en Excel: `=+PERCENTIL(A1:J7;0,75)`

Resultado de la fórmula: 8

La Figura 82 indica el cuadro de dialogo al calcular el percentil 75:

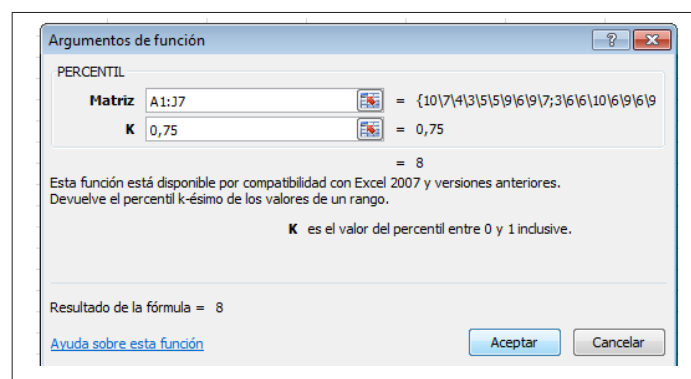


FIGURA 82: CAPTURA DE PANTALLA QUE INDICA EL CÁLCULO DEL TERCER CUARTIL DE EXCEL

Fíjese que para determinar el 75%, en el campo “K” del cuadro de diálogo se debe digitar 0.75 como valor del percentil a calcular.

En el caso de la tercera condición, Excel no calcula ni deciles ni Quintiles, pero usando el concepto de cada una de estas medidas, podemos determinar que el percentil 20 equivalente al decil 2 y el percentil 60 equivalente al Quintil 3 dado que establecen el mismo porcentaje.

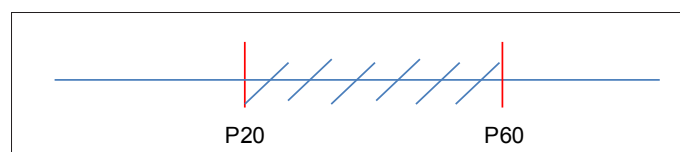


FIGURA 83: POSICIONES DEL SEGUNDO DECIL Y TERCER QUINTIL

Según la condición el gráfico indica que la región de datos señalada es la que cumple la doble condición.

Para el caso de percentiles, se busca esa función y en el cuadro de diálogo se resalta la matriz de los datos y se pide el cálculo del respectivo percentil **en decimal**, para nuestro caso el P_{20} por ejemplo se escribirá 0.2. Por tanto al calcular estos valores Excel arroja los siguientes resultados: $P_{20} = 4$ y $P_{60} = 7$.

La interpretación será que el 30% de los alumnos ha obtenido notas inferiores a 4 y que el 60% de los estudiantes tendrá notas inferiores a 7; esto significa que deberán elaborar un

ensayo aquellos alumnos que hayan obtenido entre 4 y 7 en Literatura. En el gráfico los dos puntos rojos señalan al 4 y 7.

Para la última condición calculamos el percentil 85 y el valor encontrado es 9, esto significa que el 85% de los estudiantes tiene menos de 9 en Literatura y se reconocerá a aquellos que tengan 9 o una nota mayor.



FIGURA 84: POSICIÓN CORRESPONDIENTE AL PERCENTIL 85. LA PARTE RAYADA INDICA A QUIENES SE LES HARÁ EL RECONOCIMIENTO

La región señalada en el gráfico es la que cumple la condición, pero el valor encontrado de 9 indica que el 85% de los estudiantes están por debajo del punto rojo, es decir tienen notas inferiores a 9.

Ejemplo 2

Los datos siguientes corresponden a las puntuaciones obtenidas por 50 personas en una prueba objetiva de análisis de datos que se mide sobre 35. Determine: a) el puntaje correspondiente al percentil 40 b) si se escoge a un grupo de personas en base al decil 7 (D_7) indique el valor mínimo para ser parte de los seleccionados c) el puntaje mínimo para considerarse parte del tercio superior d) qué valores establecerán el Quintil 3.5 y el decil 8.5 e) ¿el valor de la Mediana será igual a cuál Quintil? Compruebe los resultados.

20	8	17	15	12
19	13	5	9	5
11	13	8	9	8
8	14	21	13	11
19	7	14	13	12
15	11	22	10	10
6	7	20	9	11
20	22	11	11	6
15	6	15	9	19
6	16	7	9	20

Solución:

- a) el Percentil 40 arroja un valor de 10.6, esto quiere decir que el 40% de la muestra obtuvo menos de 10.6; el punto rojo indica el valor de 10.6 y la región señalada indica todos aquellos individuos que cumplen la condición.

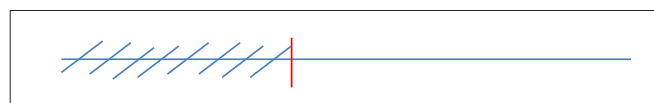


FIGURA 85: LA PARTE RAYADA INDICA LA REGIÓN QUE CUMPLE LA CONDICIÓN

- b) El decil 7 equivale al percentil 70, el valor que se obtiene es 15, es decir que para ser seleccionado se deberá obtener un puntaje igual o mayor a 15; pero la interpretación inicial y conceptual será que el 70% del grupo está por debajo del punto rojo, por tanto, obtuvo menos de 15, para nuestro caso no nos interesa esa porción si no la que está por encima del valor encontrado.

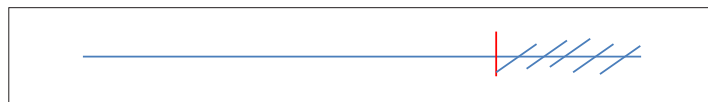


FIGURA 86: LA PARTE RAYADA INDICA LA REGIÓN QUE CUMPLE LA CONDICIÓN

- c) Según la condición a la muestra se la debe dividir en los tres tercios pero nos interesa el superior, como se señala en la figura 87.

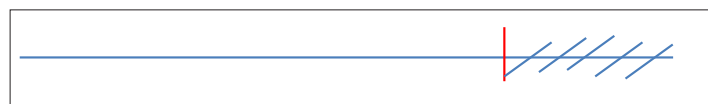


FIGURA 87: GRÁFICO EN DONDE LA LÍNEA ROJA INDICA QUE A PARTIR DE ALLÍ ESTA EL TERCIO SUPERIOR DE LOS DATOS

El punto rojo divide a la muestra en dos partes, la resaltada es aquella porción que cumple con la condición es decir el 33.33% superior, entonces la parte de la izquierda representará el 66.66% de los datos.

Esto significa que se debe calcular el percentil 66.66 ($P_{66.66}$) y ese valor será el que determine a su vez el valor mínimo para pertenecer al tercio superior.

Recuerde que aunque lo que interesa sea el 33.33% superior, las medidas de posición siempre se calculan en base a valores hacia debajo de la condición.

Los cálculos arrojan un valor de 14, por tanto para pertenecer al tercio superior es necesario obtener **14 o más**, o utilizando el concepto se diría que el 66.66% de la muestra obtuvo valores inferiores a 14.

- d) Quintil 3.5 y decil 8.5 (puntos rojos) se ven representados en la figura.



FIGURA 88: POSICIONES CORRESPONDIENTES AL QUINTIL 3,5 Y DECIL 8,5

Si un quintil representa 20% de los datos entonces 0,5 quintil será un 10% de la muestra; por tanto 3.5 quintiles representarán el 70% del total.

Los cálculos de Excel arrojan los siguientes valores en su orden: 15 y 19, la interpretación individual debería ser que el 70% (Quintil 3.5) de las puntuaciones están por debajo de 15 y el 85% está por debajo de 19.

Para la última pregunta se determina que la Mediana será igual al Quintil 2.5 (o al percentil 50, decil 5 y cuartil 2): el resultado es 11 en cada caso.

Si los datos están agrupados, se puede determinar también cada valor según la condición establecida y para ello solo hay que considerar que si la Mediana cumple también la condición de ser una medida de Posición (establece una partición del 50%), su fórmula deberá servir para los cálculos correspondientes.

Con los datos del ejemplo 2 de este apartado, vamos a calcular cada una de las condiciones dadas, pero de manera agrupada.

El cuadro de datos agrupados es el siguiente (se ha establecido una amplitud de 3 puntos)

Prueba objetiva de análisis de datos		<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
<i>L. Inf.</i>	<i>L. Sup.</i>					
5	7	6	9	9	18,00%	18,00%
8	10	9	11	20	22,00%	40,00%
11	13	12	12	32	24,00%	64,00%
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 89: CUADRO DE DATOS AGRUPADOS DEL EJEMPLO 2

La fórmula de la Mediana modificada en cuanto a la nomenclatura sería la siguiente:

$$\text{Percentil: } Pr = L + \frac{nx - faA}{f} * Am$$

En donde *Pr* representa el valor de la variable a calcular

nx representa el número de datos correspondiente al porcentaje de la condición dada

El primer caso pide determinar el Percentil 40, esto significa que el valor a encontrar debe partir a la muestra en dos partes y que la referencia será para este caso el 40% del total; por tanto, debemos calcular el 40 por ciento de 50 (total de datos) y el resultado es 20. En cuanto a la fórmula este valor (20) representa a “*nx*” y se encuentra ubicado en el segundo intervalo (ver tabla), por tanto, ese intervalo nos servirá para aplicar la fórmula al igual que se hace para calcular la mediana.

Prueba objetiva de análisis de datos		<i>Pm</i>	<i>f</i>	<i>fa</i>	<i>fr</i>	<i>fra</i>
<i>L. Inf.</i>	<i>L. Sup.</i>					
5	7	6	9	9	18,00%	18,00%
8	10	9	11	20	22,00%	40,00%
11	13	12	12	32	24,00%	64,00%
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 90: IDENTIFICACIÓN DEL INTERVALO CORRESPONDIENTE AL PERCENTIL 40

$$L = 8 ; nx = 20 ; faA = 9 ; Am = 3 ; f = 11$$

El dibujo es el mismo dado que representa una posición y nada más.



FIGURA 91: IDENTIFICACIÓN GRÁFICA DEL PERCENTIL 40

Aplicando la fórmula el resultado es 11

El Percentil 40 arroja un valor de 11, esto quiere decir que el 40% de la muestra obtuvo menos de 11, si se compara con el resultado encontrado en datos simples la diferencia es mínima.

Para la segunda condición: “escoger a un grupo de personas que supere el decil 7 (D_7)” se debe resaltar en el cuadro el intervalo que asegure haber acumulado el 70% de los datos es decir 35 (70 por ciento de 50), en este caso sería según como se muestra en la figura 92:

Prueba objetiva de análisis		Pm	f	fa	fr	fra
5	7	6	9	9	18,00%	18,00%
8	10	9	11	20	22,00%	40,00%
11	13	12	12	32	24,00%	64,00%
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 92: IDENTIFICACIÓN DEL INTERVALO CORRESPONDIENTE DEL DECIL 7

$$L = 14 \quad ; \quad nx = 35 \quad ; \quad fAa = 32 \quad ; \quad Am = 3 \quad ; \quad f = 7$$

El dibujo es el mismo y señala la región que cumple la condición de superar el D_7 .

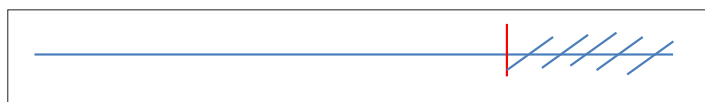


FIGURA 93: IDENTIFICACIÓN GRÁFICA DEL DECIL 7

Aplicando la fórmula el resultado es 15.29

El Percentil 70 equivalente al decil 7, arroja un valor de 15.29, esto quiere decir que el 70% de la muestra obtuvo menos de 15.29, y por tanto para ser parte de los seleccionados debe superar este valor.

La tercera condición es buscar el puntaje mínimo para considerarse parte del tercio superior, en este caso en la tabla se debe buscar el intervalo que cumpla la condición, así:

Prueba objetiva de análisis		Pm	f	fa	fr	fra
5	7	6	9	9	18,00%	18,00%
8	10	9	11	20	22,00%	40,00%
11	13	12	12	32	24,00%	64,00%
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 94: IDENTIFICACIÓN DEL INTERVALO CORRESPONDIENTE AL 66,66% DE LOS DATOS, QUE A SU VEZ DETERMINA AL TERCIO SUPERIOR

Recuerde que por concepto de las medidas de posición no se puede trabajar con el 33.33% superior, por ello se debe buscar en la tabla el 66.66% que esté por debajo del valor a encontrar, esto garantizará que a partir del punto rojo estará el tercio superior.

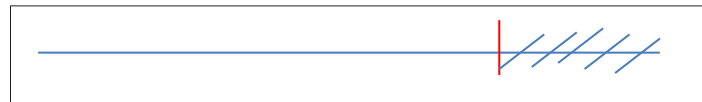


FIGURA 95: GRÁFICA QUE INDICA A LA REGIÓN CORRESPONDIENTE AL TERCIO SUPERIOR

$$L = 14 ; nx = 33^* ; fAa = 32 ; Am = 3 ; f = 7$$

Como el 66.66% de 50 es 33.33, y nx representa número de datos, ^{*} algunos autores consideran que se debe redondear este dato (normalmente al entero superior), pero otros señalan que se debe trabajar con el valor exacto; pero dado que el resultado no se ve muy afectado utilizaré el valor entero de 33.

Al aplicar la fórmula el resultado es 14.43, por tanto, el 66.66% de la muestra ha logrado un puntaje menor de este valor, es decir que para pertenecer al tercio superior, se debe lograr 14.43 o más.

En cuanto al cálculo del Quintil 3.5 y decil 8.5, los intervalos a tomar en cuenta serán los siguientes:

Prueba objetiva de análisis		P_m	f	f_a	f_r	f_{ra}
5	7	6	9	9	18,00%	18,00%
8	10	9	11	20	22,00%	40,00%
11	13	12	12	32	24,00%	64,00%
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 96: IDENTIFICACIÓN DE LOS INTERVALOS CORRESPONDIENTES AL QUINTIL 3,5 Y DECIL 8,5

Intervalo del Quintil 3.5 (70%)

14	16
----	----

Intervalo del Decil 8.5 (85%)

17	19
----	----

Para cada caso se encuentran los siguientes resultados:

Quintil 3.5: 15.29

Decil 8.5: 19.25

Respecto a la pregunta planteada hay que buscar la Mediana y el Quintil que garantice la mitad de los datos, por tanto, será el Quintil 2.5 y dado que el intervalo será el mismo, la tabla se verá así:

Prueba objetiva de análisis	Pm	f	fa	fr	fra	
5	7	6	9	18,00%	18,00%	
8	10	9	11	22,00%	40,00%	
11	13	12	12	24,00%	64,00%	
14	16	15	7	39	14,00%	78,00%
17	19	18	4	43	8,00%	86,00%
20	22	21	7	50	14,00%	100,00%

FIGURA 97: IDENTIFICACIÓN DEL INTERVALO QUE CONTIENE A LA MEDIANA Y AL QUINTIL 2,5

Al realizar los cálculos, se determina un valor de 12.25 para cada una, en este caso sí difiere un poco el valor respecto a lo encontrado en datos simples.

Ejemplo 3

Los datos a continuación se encontraron en una investigación que trataba de medir los niveles de estabilidad emocional en un grupo de trabajadores; este estudio se hizo según el sexo y se ha determinado que valores inferiores a 30 son preocupantes.

MUJERES			HOMBRES		
34	45	10	50	60	69
57	48	35	45	46	62
79	77	38	37	76	55
36	55	35	52	44	24
46	46	47	48	68	44
55	57	71	55	24	51
46	55	49	50	42	66
52	68	33	50	26	73
47	53	50	66	77	26
35	33	24	68	24	46

Determinar para cada caso:

- Medidas de Tendencia Central y Dispersión en datos simples y agrupados (amplitud 10)
- ¿Considera que los valores están dispersos?
- Determinar el tipo de sesgo
- Indique si esta distribución de frecuencias es buena o mala (cualquiera sea su respuesta indique la razón)
- Gráfico. ¿El gráfico confirma el tipo de distribución establecido?
- Si para considerarse sin problemas se establece que valores superiores al percentil 20 no serán tratados, qué nivel de estabilidad emocional garantizaría estar en ese grupo
- Si una persona no supera un valor correspondiente al percentil 15 la decisión será iniciar un proceso urgente de terapia, ¿qué valor determinaría esta situación?

Desarrollo para el grupo de los hombres

i.

	Simple	Agrupados
Media	50,80	51,83
Mediana	50,00	51,27
Moda	50,00	48,50
Rango	53,00	59,00
Desviación estándar	15,84	14,93
Coficiente de asimetría	-0,21	0,11

TABLA 46. RESULTADOS DESCRIPTIVOS PARA EL GRUPO DE HOMBRES

- ii. Los valores sí están dispersos, esto se puede determinar por el rango y la desviación estándar, dado que la variable se refiere a Estabilidad Emocional, una diferencia alrededor de 55 puntos (rango) es mucho en un mismo grupo, la desviación estándar indica que los valores centrales se alejan de la media en cerca de 15 puntos, esto también indica mucha dispersión. (Dado que se ha considerado una amplitud de 10 puntos, la desviación estándar de 15 supera en el 50% este criterio).
- iii. Según las Medidas de Tendencia Central, en datos simples no se puede determinar (no cumple con ninguna de las dos condiciones); según datos agrupados el sesgo sería positivo: $M_o < M_d < \bar{x}$
Según el coeficiente de asimetría, en datos simples el sesgo sería negativo por el signo pero puede considerarse como distribución simétrica debido a que su valor está dentro del intervalo $[-0.5 ; 0.5]$ y en datos agrupados la asimetría sería positiva por el signo pero al igual que en datos simples se puede considerar simétrica a la distribución.
- iv. Ya que se puede considerar una distribución no sesgada y que los valores medios están a 20 puntos por debajo del valor considerado “preocupante” se puede decir que la distribución califica como “buena”.
- v.

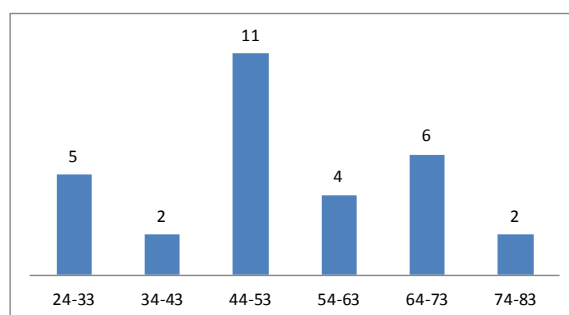


FIGURA 98: GRÁFICA DE LA DISTRIBUCIÓN DE FRECUENCIAS EN LOS HOMBRES

El gráfico NO confirma el tipo de distribución ya que no se nota una simetría según los intervalos establecidos

vi.

	Simples	Agrupados
Percentil 20	41,00	39,00

TABLA 47. VALORES DEL P20 EN DATOS SIMPLES Y AGRUPADOS

Para considerarse parte del grupo que no presenta problemas, se debe obtener un nivel de estabilidad emocional de alrededor de 40 o más.

vii.

	Simples	Agrupados
Percentil 15	29,85	33,00

TABLA 48. VALORES DEL P15 EN DATOS SIMPLES Y AGRUPADOS

Aquellas personas que obtengan un puntaje inferior a 29.85 (datos simples) o 33 (datos agrupados) deberán someterse a un proceso urgente de terapia.

RANGO PERCENTIL (EXCEL)

Una función muy interesante y de gran ayuda en Excel es aquella que permite encontrar lo que se conoce como el “Rango Percentil” (también conocido como Rango Centil), es decir ordena todos los datos de una muestra en función de su posición y establece el porcentaje de los elementos de la muestra que se encuentran por debajo de cada uno de ellos.

Para esto los datos deben presentarse en columnas; voy a tomar el último ejemplo, “niveles de estabilidad emocional en hombres”, para explicar el proceso.

1. Haga “click” en “**Datos**” y dentro del banner escoja “**Análisis de datos**”
2. En el cuadro de diálogo busque la función: “**Jerarquía y Percentil**”

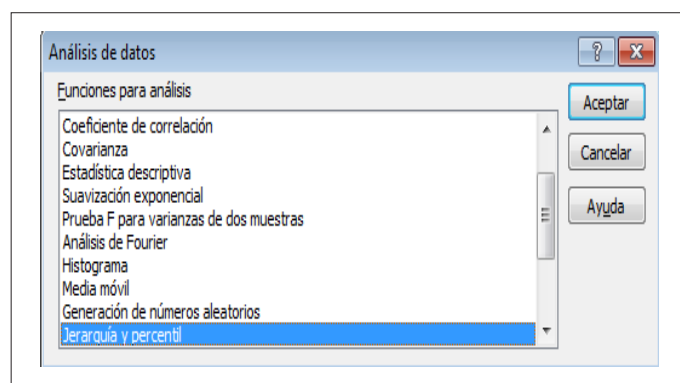


FIGURA 99: PRIMER CUADRO DE DIÁLOGO PARA TRABAJAR CON RANGOS PERCENTILES

3. Al hacer “click” se despliega el siguiente cuadro de diálogo

4.

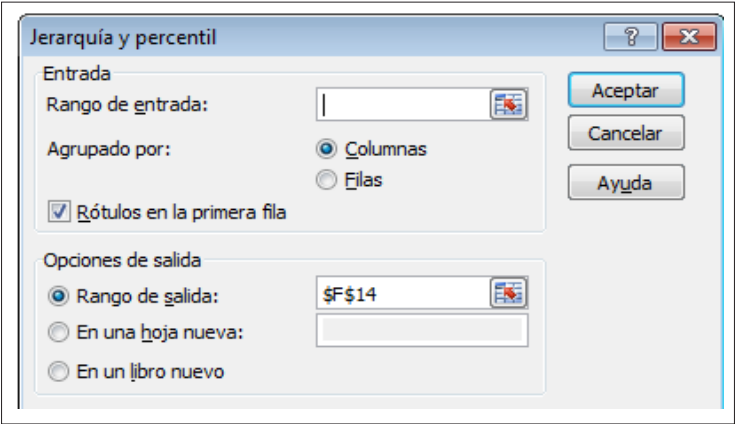


FIGURA 100: SEGUNDO CUADRO DE DIÁLOGO PARA TRABAJAR CON RANGOS PERCENTILES

Estabilidad emocional hombres
50
45
37
52
48
55
50
50
66
68
60
46
76
44
68
24
42
26
77
24
69
62
55
24
44
51
66
73
26
46

5. En “**Rango de entrada**” resalte todos los datos incluido el nombre de la variable (rótulo)
6. Ponga un “*check*” en el casillero “**Rótulos en la primera fila**”
7. En opciones de salida escoja cualquier opción; si desea que el resultado salga en la misma hoja, en “**Rango de salida**” escoja cualquier celda y finalmente acepte
8. El resultado (editado) será el siguiente:

Posición	Estabilidad emocional hombres	Jerarquía	Porcentaje
19	77	1	100,00%
13	76	2	96,50%
28	73	3	93,10%
21	69	4	89,60%
10	68	5	82,70%
15	68	5	82,70%
9	66	7	75,80%
27	66	7	75,80%
22	62	9	72,40%
11	60	10	68,90%
6	55	11	62,00%
23	55	11	62,00%
4	52	13	58,60%
26	51	14	55,10%
1	50	15	44,80%
7	50	15	44,80%
8	50	15	44,80%
5	48	18	41,30%
12	46	19	34,40%
30	46	19	34,40%
2	45	21	31,00%
14	44	22	24,10%
25	44	22	24,10%
17	42	24	20,60%
3	37	25	17,20%
18	26	26	10,30%
29	26	26	10,30%
16	24	28	0,00%
20	24	28	0,00%
24	24	28	0,00%

FIGURA 101: RESULTADO DEL PROCESO REALIZADO EN EXCEL

La columna de la izquierda indica el lugar en el que se encuentra cada dato, por ejemplo, el dato “77” está ubicado en el 19no lugar de la matriz original, si busca el dato “50” verá que ese valor está en los lugares primero, séptimo y octavo de dicha matriz.

Como el valor “77” es el más alto, en la columna “jerarquía” se indica esto y de allí hacia abajo se establece la relación de cada valor respecto a la muestra.

En la columna “Porcentaje” se indica el percentil correspondiente a cada valor de la muestra, así por ejemplo el valor “60” está en la posición décima respecto al grupo (jerarquía) indicando que el 68.9% de los datos se encuentran por debajo de él (columna porcentaje); en otras palabras, el 68.9% de los hombres obtuvo un valor inferior a 60 en estabilidad emocional.

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

1. Como primer ejercicio propuesto, se recomienda desarrollar el caso de las mujeres del tercer ejercicio de ejemplo, tomando en cuenta lo desarrollado para los hombres.
2. Los datos a continuación corresponden a los promedios de los 60 alumnos de noveno año de básica en una institución educativa. Encuentre lo siguiente: i) las medidas de tendencia central y dispersión ii) determine qué tipo de distribución encuentra y cali-fíquela iii) divida al grupo en cuartiles y encuentre los valores correspondientes a cada uno, con ello exprese comentarios iv) si a partir del percentil 80 se exoneran, qué pro-medio debe obtener un estudiante para no rendir exámenes finales v) si un estudiante

tiene promedios inferiores al percentil 30 se le condiciona entonces cuál debe ser el valor mínimo para no ser condicionado vi) si a quienes superen el tercer cuartil (Q_3) se les ofrece una beca, qué promedio debe tener un estudiante para lograrla.

1	5	7
2	8	7
5	5	7
2	5	7
4	6	9
8	6	10
4	6	10
10	8	2
4	5	3
8	3	8
4	5	5
8	3	3
4	6	9
5	6	9
4	6	9
5	4	4
9	5	5
5	3	7
4	6	3
5	6	10

3. En 6 ONG's de la ciudad de Quito, se tomaron datos sobre las prácticas de liderazgo de los directores ejecutivos de sus respectivos proyectos. A continuación, encontrará los datos correspondientes a las evaluaciones obtenidas por los líderes de dichos proyectos.

25	48	68	78	89
27	54	68	78	90
27	54	68	79	90
28	56	68	80	92
28	57	69	80	95
32	58	72	84	96
45	59	74	84	98
45	63	75	88	98
46	65	75	88	99
48	65	75	89	99

Encontrar tanto en datos simples como agrupados (am. 12) lo siguiente: a) MTC, MD, tipo de distribución (calificar e indicar la razón), coeficiente de asimetría (datos simples), gráfico (datos agrupados). EN DATOS SIMPLES determinar lo siguiente: b) si a quienes superen el percentil 75 se les mantendrá en futuros proyectos, cuál será la calificación mínima requerida c) los valores correspondientes a los cuartiles Q_1 y Q_2 d) si a quienes no superen el Quintil 2 no se les tomará en cuenta para otros proyectos, cuál será el puntaje mínimo para no ser parte de este grupo.

4. Los datos del cuadro adjunto representan los valores de niveles de ansiedad medidos a familiares de pacientes de un hospital. Determinar lo siguiente (datos simples y agrupados): MTC., MD., tipo de distribución (calificarla), encuentre en datos simples lo

siguiente: los valores correspondientes a los cuartiles 1 y 3; los valores correspondientes a los percentiles 60, 75 y 50, los valores correspondientes a los quintiles 3 y 4, concluya y recomiende; si se conoce que niveles de ansiedad superiores al percentil 80 ameritan una intervención urgente, qué valor determinaría realizar dicha intervención.

30	42	49	64	53	57
42	48	53	74	57	63
32	43	50	65	53	58
39	62	52	71	56	46
35	58	51	68	53	45
45	36	51	68	54	59
46	38	52	70	54	62
33	43	51	66	53	58
41	47	52	72	56	62
30	42	49	65	53	58

5. Se encontraron los siguientes resultados luego de aplicar una prueba psicológica que mide la capacidad de socialización en jóvenes. a) Determinar las medidas de tendencia central, de dispersión, tipo de distribución (calificar), coeficiente de asimetría b) si cualquier persona que tenga valores inferiores al percentil 35 se considera con problemas por no poder socializar con facilidad, qué valor determinaría el no pertenecer a este grupo c) si a las personas con valores superiores al percentil 65 se las considera con alta capacidad de socialización, qué valores indicarán esta condición d) cuál será el valor de la prueba para que al quinto superior se le considere totalmente normal (se sugiere usar amplitud 5).

8	15	15	16	25	15	23
10	15	24	18	17	16	16
19	13	19	19	35	6	35
7	16	4	5	10	22	18
18	4	14	16	10	15	15
22	18	21	11	20	10	18
21	20	18	9	18	6	25
25	4	20	18	19	27	20
10	16	19	35	16	20	22
10	16	7	15	6	21	16

6. Los valores a continuación se refieren a la capacidad de atención (se mide en el intervalo: [5 ; 30]) de niños entre 7 y 10 años en una institución educativa de la ciudad de Cuenca. Determinar en datos simples y agrupados (amplitud 3): las medidas de tendencia central, de dispersión, tipo de distribución (calificar), coeficiente de asimetría b) se realizará un concurso de deletreo de palabras y para ello los niños deben superar el percentil 80, qué puntaje permitirá a un niño ingresar al concurso c) los niños que tengan niveles de atención inferiores al primer cuartil ingresarán a un programa de refuerzo, qué puntaje hará que un niño requiera de esto, d) si se considera sin problemas los puntajes entre los percentiles 45 y 65, qué nivel de atención se necesitará obtener para estar en este grupo

21	28	13	16	5	19	9	22	25
6	30	16	19	9	21	13	25	28
22	7	15	18	8	21	12	24	27
21	29	14	17	30	19	10	23	25
11	29	14	17	6	19	10	23	25
21	29	14	17	6	19	10	23	26
22	30	16	18	26	21	12	24	28
26	6	23	17	7	20	21	23	21
21	29	14	17	7	20	11	23	26
21	29	15	17	22	20	11	23	26
22	30	16	18	9	21	13	25	28
21	14	13	16	22	19	10	22	25
5	29	15	18	8	20	12	24	27
22	30	15	18	8	20	12	24	27
7	5	13	17	28	19	10	23	25
21	29	15	17	7	20	11	8	11
22	30	15	18	8	21	12	24	27
22	30	16	18	8	21	13	25	28
21	29	14	17	21	20	11	23	26
21	29	15	17	29	20	28	24	27

7. Los datos siguientes se refieren a resultados en experimentos de percepción para docentes y administrativos de una institución de educación superior (medido sobre 25 puntos) cuando se sometieron a pruebas para determinar la exactitud de ciertos hechos. Lo ideal será que los datos centrales deben superar un valor de 18; se ha determinado también que quienes superen el tercer cuartil serán candidatos a ocupar cargos que requieran alta capacidad de observación y a quienes superen el percentil 90 se les ofrecerá cargos administrativos de mayor importancia.

10	12	13	14	15	17
10	12	13	14	15	17
10	12	14	14	16	17
10	12	14	15	16	18
11	12	14	15	16	18
11	13	14	15	16	18
11	13	14	15	16	19
11	13	14	15	16	19
11	13	14	15	16	19
12	13	14	15	17	20

Exprese sus comentarios en base a un análisis descriptivo

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1 (caso mujeres)

i.

Media	48,7	46,5
Mediana	46,5	46,25
Moda	46	No hay moda
Rango	69	69
Desviación estándar	13,42	14,24
Coefficiente de asimetría	1,20	0,02

TABLA 49. VALORES DESCRIPTIVOS DEL PRIMER EJERCICIO

- ii. Los valores sí están dispersos, esto se puede determinar por el rango y la desviación estándar, dado que la variable se refiere a Estabilidad Emocional, una diferencia alrededor de 69 puntos (rango) es mucho en un mismo grupo, la desviación estándar indica que los valores centrales se alejan de la media en cerca de 15 puntos, esto también indica mucha dispersión.
- iii. Según las Medidas de Tendencia Central, en datos simples el tipo de distribución es asimétrica positiva tanto por la relación entre las medidas de tendencia central como por el valor del coeficiente de asimetría; según datos agrupados el sesgo solo se puede establecer con el coeficiente de asimetría (ya que no hay moda) que indica también ser positivo (aunque según este valor se puede considerar simétrica la distribución).
- iv. Si se mantiene el criterio de que el tipo de sesgo es positivo y dado el tipo de variable, esta distribución deberá calificarse como “mala”.
- v.

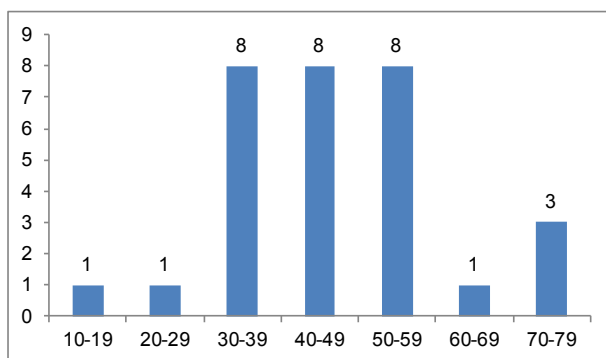


FIGURA 102: GRÁFICO DE LA DISTRIBUCIÓN EN DATOS AGRUPADOS

El gráfico no confirma el tipo de distribución
vi.

	Simples	Agrupados
Percentil 20	35	35

TABLA 50. VALORES CALCULADOS DEL PERCENTIL 20 EN DATOS SIMPLES Y AGRUPADOS

Necesitan tener un nivel superior a 35 para no ser tratados
vii.

	Simples	Agrupados
Percentil 15	34,35	33,13

TABLA 51. VALORES CALCULADOS DEL PERCENTIL 15 EN DATOS SIMPLES Y AGRUPADOS

Cualquier persona que haya obtenido un valor inferior a 34 puntos (o 33 en datos agrupados) deberá someterse urgentemente a terapia

Ejercicio 3

	Simples	Agrupados
Media	68,32	68,18
Mediana	70,5	71,67
Moda	68	78,5
Rango	74	83
Desv. Estándar	21,53	21,82
Coef. Asimetría	-0,50	-0,48
Coef. Variación	3,17	3,12

TABLA 52. VALORES DESCRIPTIVOS DEL TERCER EJERCICIO

En cuanto al tipo de distribución se puede observar que $\bar{X} < Md < Mo$ y que el coeficiente de asimetría es negativo por lo tanto se determina un sesgo negativo; para el caso esta distribución es buena dado que la tendencia de los resultados es hacia valores altos en liderazgo, el gráfico a continuación da una idea sobre esto.

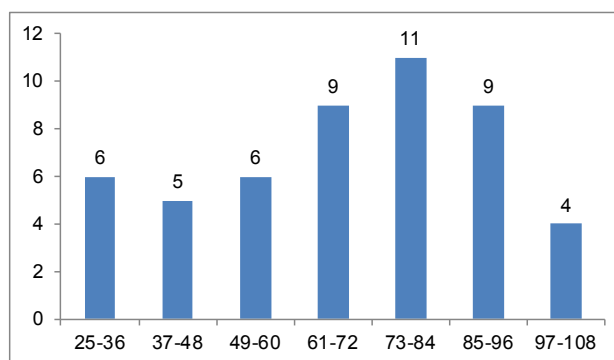


FIGURA 103: GRÁFICO DE LA DISTRIBUCIÓN EN DATOS AGRUPADOS

Dado que al calcular el percentil 75 el resultado es 87, todos aquellos líderes de proyecto que hayan superado este valor se mantendrán para proyectos futuros.

Los valores de los cuartiles uno y dos son respectivamente: 54.5 y 70.5 (nótese que es igual al valor de la Mediana en datos simples).

Por último, aquellos que no superen el valor de 66.8 (quintil 2) no serán tomados en cuenta para próximos proyectos y los que sí obtengan valores superiores podrán formar parte de nuevos proyectos.

Ejercicio 5

Dadas las condiciones de la variable, no hace falta hacer grupos dado que aquellos que hayan alcanzado un puntaje inferior al percentil 35 tendrán dificultades para socializar, el resto no manifestará problemas en este sentido.

Por lo tanto, los cálculos se harán solo en datos simples.

Puntaje correspondiente al percentil 35: 15; es decir todos aquellos que no superen este valor se les considerará con problemas.

Puntaje correspondiente al percentil 65: 18.85; por tanto, a todos aquellos que hayan obtenido este o mayor puntaje serán personas de condiciones normales en cuanto a socialización.

Para resolver lo indicado en el literal “d”, debe calcularse el percentil 80 (ver gráfico).

Puntaje correspondiente al percentil 80: 21; por ello todo aquel que haya obtenido un puntaje superior a 21 deberá ser considerado como una persona con alta capacidad de socialización.

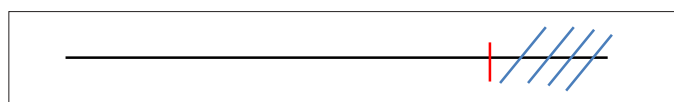


FIGURA 104: UBICACIÓN DEL VALOR QUE CUMPLE LA CONDICIÓN

El punto rojo separa hacia la derecha al quinto superior (20%) por tanto para los cálculos debe hacerse con el percentil 80 (por definición de medidas de posición)

Ejercicio 7

Media	14,3
Mediana	14
Moda	14
Rango	10
Desv. Estándar	2,49
Coef. Asimetría	0,20
Coef. Dispersión	0,12
Cuartil 3	16
Percentil 90	18

TABLA 53. VALORES DESCRIPTIVOS DEL EJERCICIO 7

En cuanto al ideal esperado esta muestra **no** cumple con la expectativa ya que sus valores centrales no llegan al mínimo requerido de 15, esto significa que en el grupo existe cierta deficiencia en referencia a la capacidad objetiva de observación o percepción.

En cuanto a las otras condiciones del ejercicio, se determina lo siguiente:

El 25% de ellos puede optar por cargos que requieran un mejor desarrollo de esta capacidad, es decir aquellos que hayan obtenido un puntaje superior a 16.

En lo referente a las personas a quienes se les puede ofrecer un cargo administrativo mayor, estas serán aquellas que superen un puntaje de 18 correspondiente al percentil 90.

CAPÍTULO 6:

CUATRO CONCEPTOS ADICIONALES

Considero importante tratar cuatro temas poco desarrollados en muchos libros, dos de ellos procesales y otros dos conceptuales complementarios y relacionados con la media. El primero referente a la limitada aplicación de las Tablas Dinámicas en Excel, otro referente al proceso abreviado de encontrar Medidas de Tendencia Central, Dispersión y Forma que también es una herramienta de Excel, el tercero relacionado con el cálculo de la Media Ponderada para análisis de datos en los cuales el “peso” de la variable es distinto según quién o qué cualifique su valor y el cuarto se refiere al cálculo, aplicación e importancia de la media acotada.

PRIMER TEMA. HISTOGRAMA

En cuanto a las Tablas Dinámicas, cabe indicar que, aunque nos han permitido desarrollar los ejercicios de este libro especialmente para encontrar resultados cuando se agrupan datos, tienen a mi juicio una falencia, y es que son algo rígidas en cuanto se refiere a este aspecto.

En cada uno de los ejercicios tanto de ejemplos desarrollados como de los propuestos, se ha sugerido una determinada amplitud de intervalo, dicha amplitud se mantiene fija para cada uno de los grupos establecidos según la instrucción; pero ¿será que al establecer una determinada amplitud conviene siempre mantenerla para todos los intervalos?

Propongo el siguiente caso: cuando se trate de hacer un estudio sobre el desarrollo según las edades en niños entre 12 y 60 meses, por ejemplo, y los valores de dicha variable fluctúen entre 5 y 30 es obvio que se deberá hacer grupos según las distintas edades, ya que no se puede esperar el mismo nivel de desarrollo entre un niño de 12 meses y otro de 48 por ejemplo; por lo tanto si se desea hacer grupos no es conveniente que haya una misma amplitud para todos ya que una diferencia de 6 meses entre un niño de 12 y otro de 18 no será la misma que entre un niño de 36 meses y otro de 42.

Otro ejemplo puede ser en cuanto a salarios, una diferencia de \$100 entre dos personas que ganan \$400 y \$500 es mucho más significativa que aquellas cuyos salarios son \$3000 y \$3100, ya que en el primer caso la diferencia representa un 25% respecto al sueldo más bajo en cambio en el segundo caso representa solo el 3.33%.

Para solucionar estos inconvenientes se puede recurrir a una función de Excel que es el “histograma”, a continuación, propongo un ejemplo sobre su uso y aplicación.

Ejemplo

Los puntajes obtenidos en una prueba aplicada a aspirantes para ingresar a una empresa, son los entregados en el cuadro siguiente:

75	89	67	29	76	64	29	67
62	23	90	95	48	18	69	64
45	77	69	20	56	74	58	34
88	58	27	74	59	45	21	57
20	42	60	35	25	87	77	60
48	57	18	54	87	67	23	92
28	19	47	81	67	47	27	57
83	55	67	33	95	75	94	89
74	43	87	48	22	61	69	46
92	59	37	26	92	19	66	89
42	18	75	51	37	57	49	18
54	22	20	64	67	27	86	25
86	71	65	83	47	62	30	56
38	49	89	20	20	43	59	34

Se conoce lo siguiente respecto a estos datos:

1. El puntaje máximo posible es de 100
2. El puntaje mínimo requerido para pasar a la segunda fase del proceso corresponde al percentil 65.
3. Las personas que hayan obtenido puntajes entre el percentil 65 y el percentil 60 pasarán a la base de datos de la organización para futuros procesos, los demás serán descartados definitivamente.
4. Aquellas personas que tengan un puntaje superior al percentil 80, tendrán una puntuación “bonificada” para la siguiente etapa.

Elabore un informe a la gerencia con opiniones y sugerencias.

Elementos a tomar en cuenta:

1. Según los datos del ejercicio existen algunos “valores críticos” a tomar en cuenta, estos son: 100 como puntaje máximo, percentiles 60, 65 y 80.
2. No se establece un valor para la amplitud por tanto esta puede ser escogida libremente según algún criterio particular.

Desarrollo:

Si se toma en cuenta los percentiles indicados, sus respectivos valores para datos simples son:

Percentil 60	62
Percentil 65	66,15
Percentil 80	75,8

TABLA 54. VALORES CALCULADOS EN DATOS SIMPLES DE LOS PERCENTILES INDICADOS

La interpretación sería que el 60% obtuvo valores inferiores a 62, el 65% valores inferiores a 66.15 y el 80% valores inferiores a 75.8; luego de esto debería aplicarse una regla de

tres para cada caso y se determinaría cuántos aspirantes hay en cada caso. Los resultados son los siguientes:

67 personas obtuvieron menos de 62 (percentil 60)

73 personas obtuvieron menos de 66.15 (percentil 65)

90 personas obtuvieron menos de 75.8 (percentil 80)

Como para los casos segundo y tercero están incluidas las personas del primer grupo, debemos restar para establecer cuántos aspirantes están en cada caso, así tenemos:

De las 73 personas que obtuvieron puntajes inferiores al percentil 65 (no superan la prueba), 6 de ellas (diferencia entre 73 y 67) estarán dentro del grupo que pasarán a la base de datos para futuros procesos y 67 serán descartadas definitivamente.

Si 73 personas no lograron el puntaje mínimo del percentil 65 entonces la diferencia pasa al segundo proceso, es decir 39 aspirantes y de esas 39 personas, 22 de ellas tendrán la puntuación “bonificada” para la segunda etapa del proceso.

Gráficamente se puede resumir así:

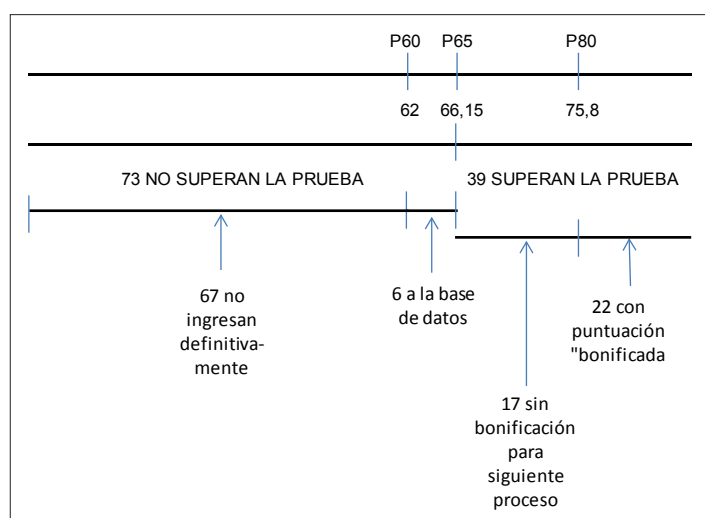


FIGURA 105: REPRESENTACIÓN GRÁFICA DE LAS CONDICIONES DADAS PARA EL EJEMPLO 1

Si quisiéramos crear una Tabla Dinámica con tres valores específicos correspondientes a los percentiles 60, 65 y 80 no podríamos; es por ello que debemos recurrir a la función “histograma” y crear grupos con los valores específicos, así:

En el menú principal de Excel en el campo **datos** verificar si se encuentra habilitada la herramienta “**análisis de datos**”, si no lo está se deberá habilitar siguiendo los siguientes pasos:

1. Haga “click” en “**Archivo**” y luego bajo el parámetro “Ayuda” encuentra “**Opciones**” si hace click allí encontrará la opción “**Complementos**”, haga “click” allí.

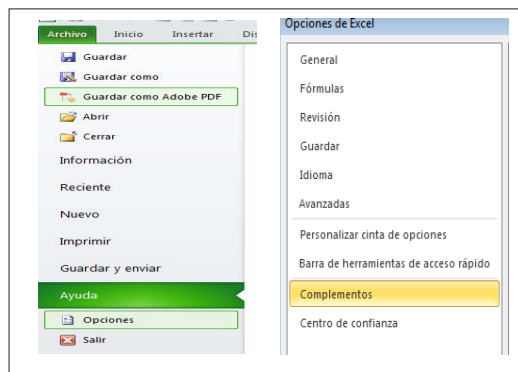


FIGURA 106: CAPTURA DE PANTALLAS PARA HABILITAR "COMPLEMENTOS" EN EXCEL, PASOS 1 Y 2

2. En la parte baja del último cuadro de diálogo (Complementos), encuentra un casillero que dice **"Complementos en Excel"** y a su derecha **"ir"**
3. Haga *click* en **"ir"**
4. En el cuadro de diálogo que aparece, ponga un *"check"* en el casillero **"herramientas para análisis"** y luego **"aceptar"**

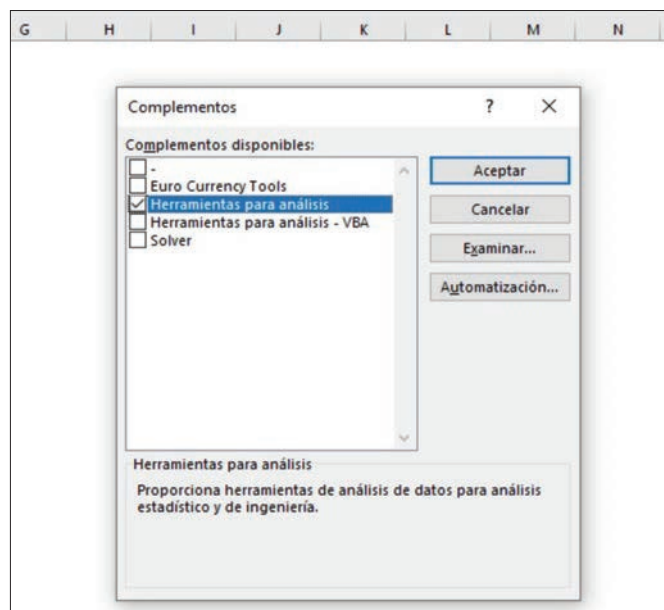


FIGURA 107: CAPTURA DE PANTALLA QUE INDICA LAS HERRAMIENTAS A HABILITAR

Se habrá habilitado ya entonces la herramienta "análisis de datos".

Antes de utilizar esta herramienta de análisis, se debe hacer lo siguiente:

1. Los datos deben tener un rótulo en la primera fila, y
2. Debe especificar en alguna celda un rótulo que establezca a qué se refieren los valores bajo ese título y sin dejar celdas vacías hay que digitar los valores críticos para el análisis, así:

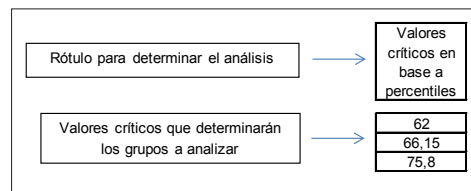


FIGURA 108: IDENTIFICACIÓN DE VALORES CRÍTICOS SEGÚN LA CONDICIÓN

Estamos listos entonces para utilizar la herramienta, en el menú principal en **“Datos”** al final del banner se encuentra ya habilitada la función **“Análisis de datos”**; haga *“click”* y se desplegará el siguiente cuadro de diálogo:

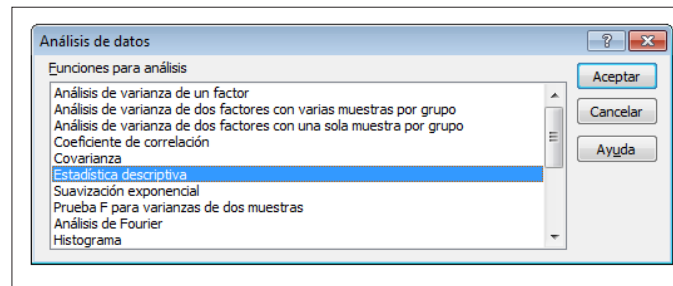


FIGURA 109: CUADRO DE DIÁLOGO QUE ABRE ALGUNAS FUNCIONES DE ANÁLISIS EN EXCEL

Dentro de él busque la función **“Histograma”** según se indica en la figura 110

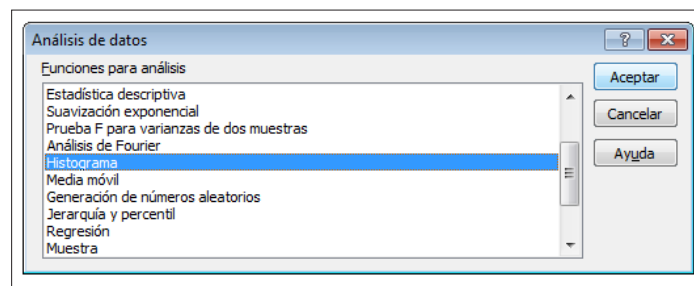


FIGURA 110: FUNCIÓN "HISTOGRAMA" DENTRO DE LA HERRAMIENTA ANÁLISIS DE DATOS

Escoja **“Aceptar”** y se despliega el siguiente cuadro de diálogo:

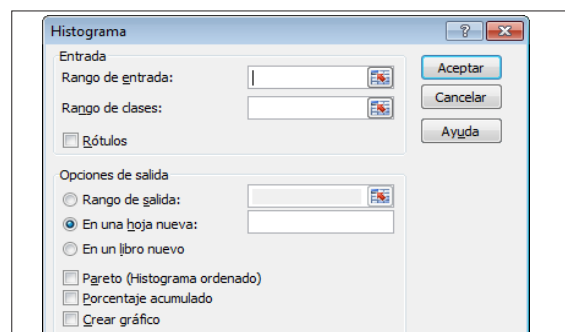


FIGURA 111: CUADRO DE DIÁLOGO QUE PERMITE APLICAR ESTA FUNCIÓN SEGÚN LAS CONDICIONES ESPECIFICADAS

Luego de ello, siga los siguientes pasos:

1. En “**Rango de entrada**” resalte desde el nombre del campo hasta el último dato.

Prueba para aspirantes a una empresa
75
62
45
88
20
48
28
83
74
92
42
.
.
.
.

FIGURA 112: VARIABLE Y VALORES Y A ANALIZAR

2. En “**Rango de clases**” resalte el rótulo y los valores críticos

Valores críticos en base a percentiles
62
66,15
75,8

FIGURA 113: RÓTULO DE LAS “CLASES” Y VALORES CRÍTICOS SEGÚN CONDICIÓN

3. Habilite (✓) el casillero “**Rótulos**”
4. En “Opciones de salida” escoja lo que desee
5. De los últimos tres casilleros escoja los dos últimos “**Porcentaje acumulado**” y “**Crear gráfico**”

El resultado para nuestro ejercicio será el siguiente:

Valores críticos en base a percentiles	Frecuencia	% acumulado	Frecuencia acumulada *
62	68	60,71%	68
66,15	5	65,18%	73
75,8	16	79,46%	89
y mayor...	23	100,00%	112

TABLA 55. RESULTADOS QUE ARROJA EL EXCEL LUEGO DEL PROCESO INDICADO

*En realidad la cuarta columna “Frecuencia acumulada” no es parte del proceso obtenido en Excel, pero se sugiere crearla.

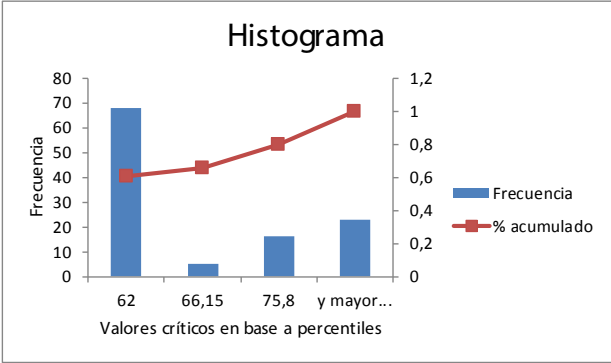


FIGURA 114: GRÁFICO QUE PRESENTA EL EXCEL LUEGO DEL PROCESO SEGUIDO

NO INGRESAN	68	No ingresan definitivamente
	5	Base de datos
PASAN A 2do. PROCESO	16	Sin bonificación para siguiente proceso
	23	Bonificados para siguiente proceso

FIGURA 115: ANÁLISIS SEGÚN LOS RESULTADOS OBTENIDOS

Como se puede notar, esta herramienta ha creado cuatro intervalos cuya lectura debe hacerse de la siguiente manera:

		Valores críticos en base a percentiles
Primer grupo	Valores "HASTA" 62	62
Segundo grupo	Valores "HASTA" 66,15	66,15
Tercer grupo	Valores "HASTA" 75,5	75,8
Cuarto grupo	Valores mayores a 75,8	y mayor...

FIGURA 116: INTERVALOS CREADOS POR EXCEL SEGÚN LAS CONDICIONES DADAS

Es decir, los grupos pueden identificarse así:

Intervalos	Frecuencia
18 - 62	68
63 - 66,15	5
66,16 - 75,8	16
75,9 - 95	23

TABLA 56. IDENTIFICACIÓN DE LOS LÍMITES DE CADA INTERVALO CREADO

Es claro que esto no se puede lograr utilizando Tablas Dinámicas, por ello la importancia de esta herramienta como complemento de análisis.

SEGUNDO TEMA. ESTADÍSTICA DESCRIPTIVA ABREVIADA

En cuanto al segundo tema **proceso abreviado de encontrar Medidas de Tendencia Central y Dispersión**, aquí explicaré, con los mismos datos del ejemplo anterior, cómo funciona este sistema que permite encontrar con rapidez estos valores en datos simples.

Al igual que el caso anterior, los datos deben estar en una sola columna con un rótulo en la primera celda y sin espacios entre el rótulo y el primer dato, así:

Prueba para aspirantes a una empresa
75
62
45
88
20
48
28
83
74
92
42
.
.
.
.

FIGURA 117: VARIABLE Y VALORES A ANALIZAR

Se procede igual que el caso anterior, así:

En el menú principal en **“Datos”** al final del banner se encuentra ya habilitada la función **“Análisis de datos”**; haga **“click”** y se desplegará el siguiente cuadro de diálogo:

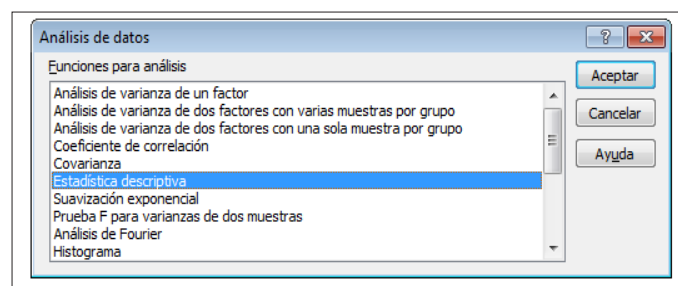


FIGURA 118: CUADRO DE DIÁLOGO PARA INICIAR EL PROCESO

Dentro de él busque la función “Estadística descriptiva” según se indica en la figura anterior

Escoja “Aceptar” y se despliega el siguiente cuadro de diálogo:

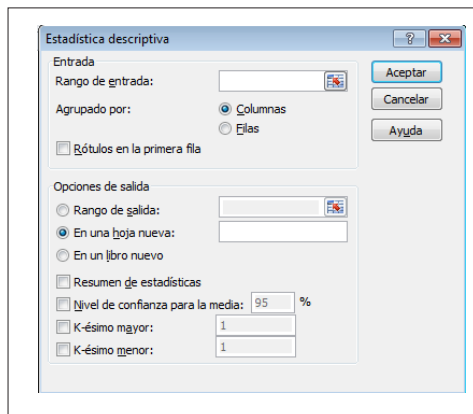


FIGURA 119: CUADRO DE DIÁLOGO PARA REALIZAR EL PROCESO

En “Rango de entrada” resalte desde el nombre del campo hasta el último dato.

Prueba para aspirantes a una empresa
75
62
45
88
20
48
28
83
74
92
42
.
.
.
.

FIGURA 120: VARIABLE Y VALORES RESALTADOS

En “Agregado por:” mantenga la opción “Columnas”

Habilite (✓) el casillero “Rótulos en la primera fila”

En “Opciones de salida” escoja lo que desee

De los cuatro últimos casilleros escoja solo los dos primeros: “Resumen de estadísticas” y “Nivel de confianza para la media:” (los demás casilleros no serán útiles para efectos de este libro)

El cuadro de diálogo entonces ya con estos pasos queda así:

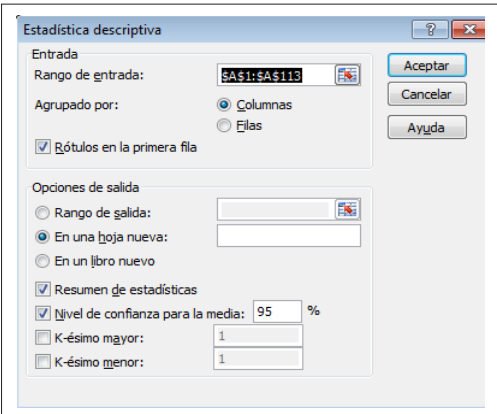


FIGURA 121: CUADRO DE DIÁLOGO LLENO CON LOS DATOS DEL EJERCICIO

Haga “click” en “Aceptar” y el resultado será el siguiente (luego de darle forma):

Prueba para aspirantes a una empresa	
Media	54,79464286
Error típico	2,204992375
Mediana	57
Moda	67
Desviación estándar	23,33544587
Varianza de la muestra	544,5430341
Curtosis	-1,126213447
Coefficiente de asimetría	-0,024121761
Rango	77
Mínimo	18
Máximo	95
Suma	6137
Cuenta	112
Nivel de confianza(95,0%)	4,369339591

TABLA 57. RESULTADOS ENCONTRADOS AL FINALIZAR EL PROCESO

Este cuadro resume todo lo concerniente a Medidas de Tendencia Central, Dispersión y Forma en datos simples; como se podrá notar este proceso abrevia muchísimo lo que hemos venido haciendo hasta este punto.

Aclaro un poco sobre el dato correspondiente a “Nivel de confianza (95%)”; este valor determina el intervalo de confianza para el valor de la media si la distribución de la población es normal (de esto hablaremos en un capítulo posterior).

TERCER TEMA. MEDIA PONDERADA

Respecto al tercer tema relacionado con el cálculo de la Media Ponderada, propongo un ejemplo bastante sencillo pero de mucha aplicación referente a evaluación (desempeño) escolar.

Supongamos que la nota del quimestre en una determinada materia está dada por los siguientes aportes: Tareas, Trabajos en clase, Lecciones escritas, Lecciones orales y Prueba de unidad, si cada una de ellas se calificaría sobre 10 puntos y al final se haría un promedio, esto no discriminaría el valor de una tarea (que el alumno puede bien haber copiado) respecto al de una prueba que por concepto es más fuerte.

Para evitar esto se recomienda utilizar un sistema que permita dar el verdadero valor a cada una de estas actividades y sopesando respecto a su nivel de esfuerzo y de dificultad.

Pongo a consideración el siguiente ejemplo en dos cuadros comparativos, en el primero (figura 122) se ha obtenido el promedio simple (media) y en el segundo (figura 123) se ha calculado la media ponderada.

	TAREAS	TRABAJOS EN CLASE	LECCIONES ESCRITAS	LECCIONES ORALES	PRUEBA DE UNIDAD	PROMEDIO
ALUMNO						
Acosta Grace	10	4	6	6	3	5,8
Álvarez Luis	3	4	6	6	10	5,8
Benítez Diego	10	0	3	5	6	4,8
Díaz Sandra	1	9	4	8	1	4,6
Flores Isabel	3	2	1	1	10	3,4
García Gustavo	1	7	2	4	0	2,8
Gómez Gabriel	2	3	1	10	7	4,6
Huertas María	2	3	1	7	10	4,6
López Consuelo	9	4	0	5	10	5,6
Martínez Jaime	1	6	9	9	9	6,8
Medina Aitor	2	10	6	4	2	4,8
Pérez Verónica	7	6	5	1	0	3,8
Ramírez Cristina	0	4	4	5	1	2,8
Rodríguez Juan	2	3	4	9	10	5,6
Romero Mauricio	2	3	0	10	8	4,6
Ruiz Antonio	3	2	3	3	10	4,2
Sánchez Alejandro	4	0	3	1	7	3,4
Sosa Fernanda	10	3	2	1	1	3,4
Torres Miguel	2	5	7	5	1	4

FIGURA 122: VALORES DE CADA ÍTEM Y PROMEDIO POR ESTUDIANTE

	TAREAS	TRABAJOS EN CLASE	LECCIONES ESCRITAS	LECCIONES ORALES	PRUEBA DE UNIDAD	MEDIA PONDERADA
PONDERACIÓN	5%	10%	20%	25%	40%	
ALUMNO						
Acosta Grace	10	4	6	6	3	4,8
Álvarez Luis	3	4	6	6	10	7,25
Benítez Diego	10	0	3	5	6	4,75
Díaz Sandra	1	9	4	8	1	4,15
Flores Isabel	3	2	1	1	10	4,8
García Gustavo	1	7	2	4	0	2,15
Gómez Gabriel	2	3	1	10	7	5,9
Huertas María	2	3	1	7	10	6,35
López Consuelo	9	4	0	5	10	6,1
Martínez Jaime	1	6	9	9	9	8,3
Medina Aitor	2	10	6	4	2	4,1
Pérez Verónica	7	6	5	1	0	2,2
Ramírez Cristina	0	4	4	5	1	2,85
Rodríguez Juan	2	3	4	9	10	7,45
Romero Mauricio	2	3	0	10	8	6,1
Ruiz Antonio	3	2	3	3	10	5,7
Sánchez Alejandro	4	0	3	1	7	3,85
Sosa Fernanda	10	3	2	1	1	1,85
Torres Miguel	2	5	7	5	1	3,65

FIGURA 123: VALORES DE CADA ÍTEM Y CÁLCULO DE LA MEDIA PONDERADA PARA CADA ESTUDIANTE

En el segundo cuadro he resaltado tres casos para análisis, en el primero (en verde) los dos estudiantes tienen las mismas notas pero en Tareas y Prueba de Unidad están intercambiados, si compara los resultados entre el resultado del cuadro 1 con el cuadro 2 de estos estudiantes, notará que existe una buena diferencia, esto debido a que el peso dado al ítem “Prueba de Unidad” es mayor (por su nivel de complejidad) que el dado al ítem “Tareas”; si no se hubiese ponderado, ambos tendrían el mismo promedio, ¿le parece esto justo?

El segundo ejemplo (en lila), los dos estudiantes se diferencian solo en cinco centésimas, aquí se ve que la señorita Flores no hizo muy bien las cosas durante el proceso pero le fue perfecto en la prueba, en cambio al señor Benítez que sin haber hecho un gran esfuerzo tampoco en el proceso especialmente en lo referente a “trabajos de clase”, no le fue también ya en la nota final, esto porque aunque haya obtenido diez puntos en deberes, en la prueba no le fue tan bien (a esto me refiero cuando digo por ejemplo que, si las tareas son copiadas, aunque tenga buena nota en ese ítem la recompensa o “peso” no es la misma que el éxito en la prueba).

Para el tercer caso, la diferencia de nota no es muy significativa (45 centésimos), la única diferencia está en el peso dado para cada uno de los dos últimos ítems, pero compare esas notas con los valores del primer cuadro.

Con esto he querido hacer notar que el valor del promedio (media) de una variable hay que saber manejarlo especialmente cuando dicho promedio se obtiene de distintos ámbitos para la misma variable.

Pongo también a consideración el siguiente caso que es muy común para los reportes de calificaciones de los estudiantes especialmente del nivel medio. En muchas instituciones se suele añadir al reporte de notas el promedio de las calificaciones como un indicador del rendimiento académico; nada más simple y al mismo tiempo nada más “mentiroso” en cuanto al verdadero análisis que debe hacerse en cuanto a este aspecto. La razón de este parecer es que a mi juicio el rendimiento académico debe valorarse por áreas de estudio y no mezclar a todas las materias como si todas evaluarían las mismas capacidades de los estudiantes.

Propongo el siguiente ejemplo de dos hermanas mellizas (las voy a llamar Camila y Emilia) que obviamente están bajo las mismas condiciones en cuanto a ambiente, hogar, amistades, curso, edad, entre otras.

Los reportes de calificaciones de Camila y Emilia son los siguientes:

	CAMILA	EMILIA
Matemáticas	6.5	9.8
Física	6	10
Química	5.5	10
Literatura	8.5	5
Sociales	8.5	4
Idioma (inglés)	9	6
Cultura física	7	6
Biología	5	10
Computación	7.25	10
Historia	9	6.5
Lógica y Ética	9.5	6.75
Teoría del conocimiento	9	6.5

El promedio de Camila es 7.56 y el de Emilia es 7.55, según esto y en el criterio de más instituciones, departamentos de orientación y profesores de los que uno se puede imaginar, es que ambas hermanas tienen el mismo rendimiento académico; pero en verdad lo único que tienen en común es el promedio y su genética, ya que en cuanto a las capacidades en las distintas áreas de estudio de una y otra en realidad son muy diferentes.

Tan solo observando las notas en las materias señaladas se puede concluir que Camila tiene falencias en lo referente a las materias que tradicionalmente se conocen como más complejas (Matemáticas, Física, Química, Biología) cosa que no ocurre con Emilia, pero a su vez Emilia tiene serias dificultades en materias referentes al ámbito social (Literatura, Sociales, Lógica y Ética, Teoría del Conocimiento).

Entonces los estudiantes no deberían ser “juzgados” ni comparados en base al promedio de calificaciones y los departamentos de orientación, profesores y padres de familia deben hacer un análisis de calificaciones en cuanto al área de estudio que representan las distintas asignaturas.

Este error de presentar el promedio de calificaciones como elemento determinante del aprovechamiento del estudiante, se basa en que a todas las materias se las “mete en el mismo saco” y no debería ser así, tiene que hacerse un análisis por temáticas de estudio, esto significa que puede hacerse promedios según el área de estudio de las materias: ciencias sociales, ciencias exactas, ciencias naturales, área tecnológica, idiomas, deportes, según cómo lo decida cada institución. Cuando se hace de esta manera, estudiantes, profesores, padres de familia, institución tendrán parte de la información necesaria para determinar en qué áreas debe reforzarse al estudiante según sus calificaciones.

CUARTO TEMA. MEDIA ACOTADA

El último tema se refiere a la aplicación de la media acotada, esta medida es muy importante ya que trata de quitar los “picos” que pueden darse en determinada variable y que, como dijera en páginas anteriores respecto a la media, distorsionarán su valor.

Por ejemplo, revisemos los siguientes datos:

30	37
2	45
67	70
74	73
53	70
17	84
42	30
48	84
55	70
82	45

En *fx* (barra de fórmulas de Excel) busque la función MEDIA.ACOTADA, obtendrá el siguiente cuadro de diálogo en el que deberá llenar en “Matriz” las celdas donde se encuentran los datos a analizar y en “Porcentaje” establecer el valor con el que quiere acotar los datos según se muestra en la Figura 124.

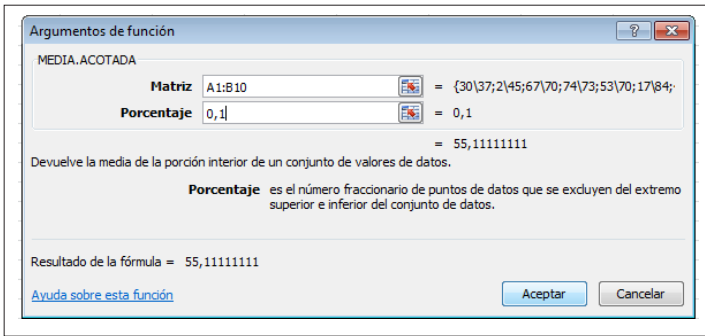


FIGURA 124: CUADRO DE DIÁLOGO DE LA FUNCIÓN CON LOS VALORES ESTABLECIDOS PARA EL CÁLCULO

Media	53,9
Media acotada al 10%	55,11

TABLA 58. VALOR DE LA MEDIA LUEGO DE ACOTAR AL 10% LOS DATOS

El valor de 53.9 es el promedio general y el valor de 55.11 (diferencia de 1.79 puntos) es el nuevo promedio luego de haber pedido al Excel a través de la función “*Media.Acotada*” que “corte” los valores extremos correspondientes al 10% de los datos; con esa instrucción Excel eliminó de la muestra el valor más bajo (2) y el más alto (84), fíjese que de todas maneras queda un valor de 84 en la muestra.

Veamos un ejemplo en tema de salarios en una empresa estructurada por 8 áreas cuyas cabezas tienen el salario máximo (\$5000) y el salario menor es de \$500.

638	578	657	600	600
762	818	1029	500	741
1048	888	1070	936	875
1165	1399	1072	1346	1046
1225	1463	1210	1479	1106
1623	1586	1250	1723	1177
1661	1791	1620	1756	1633
1662	1799	1752	1849	2023
1812	1851	1769	1929	2035
1827	2246	1843	2072	2062
2038	2248	2000	2377	2212
2072	2265	2357	2400	2524
2114	2352	2449	2441	2548
2122	2743	2514	2557	2574
2398	2878	2539	2798	2809
2403	2893	2872	2891	2898
2717	2921	3133	3130	3023
2752	3014	3375	3291	3046
2776	3163	3458	3357	3140
2807	3215	3460	3545	3525
2930	3316	3466	3598	3541
3128	3407	3733	4095	3715
3523	3445	3823	4240	3750
4119	3473	3944	4461	3954
4352	3515	4098	4614	4244
4418	3784	4251	4669	4370
4569	3929	4623	4700	4413
4637	4324	4786	4866	4447
5000	5000	5000	5000	4478
4000	5000	5000	5000	5000

Media	2769,89
Media acotada al 10%	2765,96
Media acotada al 20%	2757,28

TABLA 59. VALORES DE LA MEDIA Y MEDIA ACOTADA AL 10% Y 20%

Como se puede observar el promedio obtenido sin “cortar picos” es más elevado y por tanto los otros darán una idea más real de la situación en cuanto a este dato. Dependerá de algunos factores como el número de funcionarios con valores muy elevados respecto al resto o también si el promedio deberá o no hacerse tomando en cuenta a los funcionarios de alto rango.

El siguiente caso representa las notas de una materia calificada sobre 10 puntos, fíjese que solo hay una nota muy baja y la que le “sigue” tiene 6 puntos más,

8	9
7	10
8	8
7	9
9	7
9	1
7	10
8	9
8	10
10	8

El cuadro siguiente expresa los resultados de media y media acotada a varios porcentajes.

Media	8,1		
Media Acotada al 10%	8,39	Diferencia	0,29
Media Acotada al 15%	8,39	Diferencia	0,29
Media Acotada al 20%	8,38	Diferencia	0,28

TABLA 60. VALORES DE LA MEDIA Y MEDIAS ACOTADAS CON VARIOS CRITERIOS

Fíjese que los valores de acotación no tienen diferencia mayor pero sí se diferencian con el promedio clásico en 29 centésimos.

Si eliminamos (a mano) la nota muy baja (uno) cambia a lo siguiente:

Media	8,41		
Media Acotada al 10%	8,41	Diferencia	0,00
Media Acotada al 15%	8,40	Diferencia	-0,01
Media Acotada al 20%	8,40	Diferencia	-0,01

TABLA 61. VALORES ACOTADOS LUEGO DE LA ELIMINACIÓN DEL DATO MENOR

Aunque no se ha aplicado un sistema de acotación, se recomienda en algunas ocasiones hacer el ejercicio sin tomar en cuenta uno o dos datos de los extremos (eliminando “a mano”) ciertos datos que pueden estar distorsionando; en este caso se puede ver que no hay diferencia entre el promedio común y los promedios obtenidos en cada acotación, pero sí difieren los resultados respecto al promedio original.

CAPÍTULO 7:

DISTRIBUCIÓN NORMAL

El concepto de curva normal fue desarrollado por Gauss y Laplace, aunque tiempo atrás, en 1733, DeMoivre fue el primero en obtener la ecuación de dicha curva, y por tal motivo hoy se conoce como curva de Gauss o curva de DeMoivre, o bien, curva de campana (González Betanzos, Escoto Ponce de León, & Chávez López, 2017, p. 60)

En capítulos anteriores se han realizado análisis del comportamiento de la variable en cuanto a las Medidas de Tendencia Central y Dispersión, al tipo de distribución que se haya presentado y estableciendo ciertas condiciones que permiten determinar qué grupo dentro de la muestra las cumple. También se revisaron varios tipos de distribución de la variable que a manera de resumen se presenta a continuación:

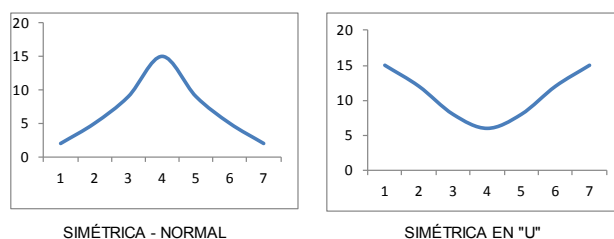


FIGURA 125: CURVAS SIMÉTRICAS

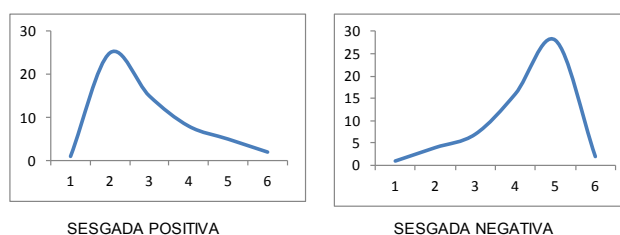


FIGURA 126: CURVAS ASIMÉTRICAS Y SESGADAS

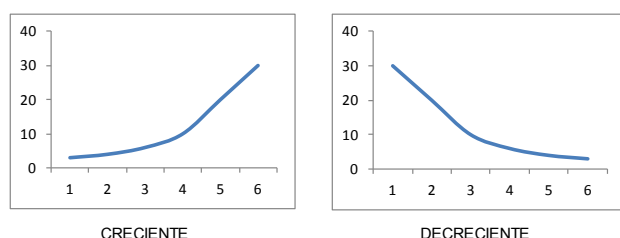


FIGURA 127: CURVAS CRECIENTE Y DECRECIENTE, CURVAS INVERTIDAS

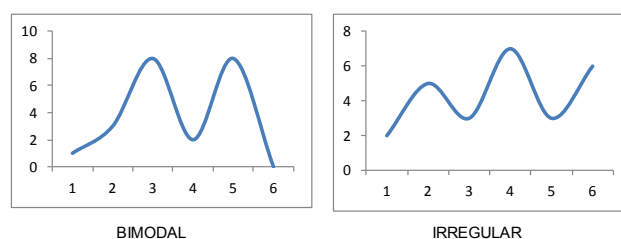


FIGURA 128: CURVAS DE DISTRIBUCIÓN IRREGULAR

Cada una de ellas y cualquier otra que se presente, obligarán a una distinta interpretación pero que, mientras no se conozca de qué variable se trata, no se podrá “calificar” dicho comportamiento como bueno o malo según los intereses de la investigación.

En lo referente a la distribución Normal se conoce que, en el estudio de una variable y especialmente en variables referentes al Comportamiento Humano, hay un tipo de distribución muy particular que vale la pena detenerse en su estudio, ya que se espera que esta forma de distribución de la variable se presente en condiciones generales a cualquier grupo de estudio.

Cuando en el tercer capítulo se revisaron los distintos tipos de distribución de una variable, ya se habló sobre la Distribución Normal y las condiciones que se requiere cumpla para considerarla como tal; paso entonces a recordar ciertos elementos de juicio.

- i. Para el caso de la distribución simétrica (normal) las medidas de tendencia central son iguales (realmente muy cercanas entre sí). En este caso se debe establecer lo siguiente: será muy difícil que Media, Mediana y Moda sean exactamente iguales, por ello, y a criterio del investigador, se analizará cuán cercanos considera son los valores, de qué variable se trata y sobre cuánto se mide ésta para decidir si la distribución califica como normal.
 - ii. El coeficiente de asimetría es un valor que determina cómo están distribuidos los datos de una muestra y, según lo establecido en el capítulo 3, este coeficiente puede presentar los siguientes casos:
 1. c.a. > 0 Asimetría positiva
 2. c.a. < 0 Asimetría negativa
 3. c.a. $= 0$ Simetría (distribución Normal)
 - iii. Dado que en un determinado estudio será muy difícil encontrar que la variable presente una distribución cuyo coeficiente de asimetría sea cero, se considera entonces que si los valores de dicho coeficiente están dentro del intervalo de ± 0.5 , a esa distribución se la reconocerá como simétrica, es decir si c.a. $[-0.5; 0.5]$ la distribución se puede decir que cumple la característica de ser normal.
- De todas maneras, esto también deberá ser analizado mediante otros valores como las medidas de tendencia central, la desviación estándar, la amplitud de los intervalos (datos agrupados) y como ayuda visual con un gráfico.

El estudio de este tipo de distribución se aplica mucho en todo tipo de investigación, pero toma un matiz muy especial en las Ciencias Sociales, Ciencias del Comportamiento Humano y Ciencias de la Educación; la razón es que en este tipo de disciplinas las distintas variables que forman parte de estos estudios sugieren que se dé este tipo de distribución como “lo esperado”, dado que la concentración de sujetos de estudio debería ubicarse en valores cercanos a la Media (valor referencial), es así que los extremos de la curva sugieren que pocos individuos presentan valores alejados del valor medio.

Como ejemplo se puede citar variables como la estatura, la inteligencia general (C.I.), el desempeño escolar, de ellas se espera que la distribución de una población tome la forma de Campana de Gauss, es decir que los resultados en esa población se concentren en valores conocidos o aceptados históricamente o por estudios previos; por ejemplo la estatura media de los hombres en el Ecuador se conoce es de 167 cm, esto quiere decir que si se hace un estudio sobre este tema, la gran cantidad de los varones a estudiar tendrán estaturas muy cercanas a este valor y habrá pocos hombres con estaturas muy inferiores o superiores; algo similar ocurrirá con el valor del C.I. y cualquier otra variable.

De otro lado, la Distribución Normal no solo es una forma de presentarse los datos en cuanto a su frecuencia, es una distribución que habla sobre la probabilidad de que un evento ocurra en base a una condición; desde cierto punto de vista su tratamiento es muy parecido al de las medidas de posición.

No es necesario realizar un estudio de Probabilidades para entender y manejar esta distribución en la práctica (es una de las razones por las que este libro no trata ese tema), lo que sí es importante es determinar las características específicas de la Distribución.

CARACTERÍSTICAS DE LA DISTRIBUCIÓN NORMAL

Dado que esta distribución es teórica respecto a los puntajes que se pueden dar en una población, la ecuación de la curva se define de la siguiente manera:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Las características específicas de esta curva son las siguientes:

- i. Es simétrica respecto al valor de la Media (en realidad se podría decir simétrica respecto a las Medidas de Tendencia Central)
- ii. Es infinita en sus extremos (se conoce como asintótica al eje de abscisas)
- iii. La región bajo la curva establece una probabilidad total – en la práctica – del 100%, repartido por su carácter de simétrica en 50% a cada lado de la Media.

Pero esta ecuación es para encontrar el lugar geométrico que cumple esta característica y ese lugar geométrico no es otra cosa que la Campana de Gauss; en cambio para el cálculo de la probabilidad según una condición determinada la fórmula es la siguiente:

$$z = \frac{x - \mu}{\sigma}$$

Es importante señalar que la curva normal representa una estandarización, es decir que sirve para cualquier variable cuya distribución de frecuencias indique tener esta característica. Para ello es indispensable aclarar que esta estandarización supone lo siguiente en cuanto a la fórmula se refiere: la media (μ) se considera el punto inicial y por tanto toma un valor de 0 ($\mu = 0$) y la desviación estándar (σ) toma el valor de 1 ($\sigma = 1$); es por esta razón que las tablas de distribución normal parten de cero y según ciertas publicaciones se encuentra hasta un valor “z” de 4 o más inclusive; pero en la práctica se dice que si “z” = 3 (o “z” = - 3) se ha abarcado ya el 50% de la curva.

¿Pero qué es “Z” ?, este valor se define como el número de desviaciones estándar (se podría decir la distancia) a la que se encuentra un determinado valor de la variable respecto de la Media y el signo determinará si se encuentra por sobre la media o es menor a ella; su explicación se hace más fácil de manera gráfica.

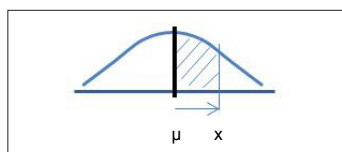


FIGURA 129: DISTANCIA DE “X” RESPECTO A LA MEDIA

El valor “x” (condición) se ha alejado linealmente de la Media una determinada cantidad y ha generado a su vez un área bajo la curva, el valor numérico de ese “alejamiento” es “Z” y el área bajo la curva será la probabilidad de que ocurra el evento según una condición previamente determinada. Los valores de “μ” y “σ” representan la Media y Desviación Estándar respectivamente.

Al aplicar la fórmula se encontrará un valor numérico que nos indicará a cuánta distancia (linealmente hablando) se encuentra ese valor respecto a la media y ese dato nos permitirá buscar en tablas previamente elaboradas el valor de la probabilidad (área señalada) según la condición establecida.

Por ejemplo: si la media de calificación de una determinada materia es 7.8 (se califica sobre 10 puntos) y la desviación estándar es 1.12; ¿cuán probable es que una persona obtenga un valor superior al promedio pero menor a 9?

$$z = \frac{x - \mu}{\sigma}$$

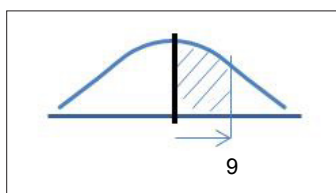


FIGURA 130: REPRESENTACIÓN DE LA CONDICIÓN DADA

Media (μ): 7.8

Desviación estándar (σ): 1.12

Valor crítico: 9

Condición: valores que se encuentren entre 7.8 y 9

Resultado de la fórmula: **1.07**

Esa será la “distancia” (en desviaciones estándar) entre la media y el valor crítico.

Este valor lineal habrá generado un área bajo la curva que representará la probabilidad de que ese evento ocurra.

Para encontrar la probabilidad entonces se debe buscar en la tabla (Anexo 1) el valor correspondiente según se observa en la Figura 126.

El signo de “z” determinará si el valor referencial dado en la condición es mayor (signo positivo) o menor (signo negativo) del valor de la media. En el caso del ejemplo desarrollado, “z” fue positivo dado que 9 (valor de la condición) es mayor que 7.8 (valor de la media).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0,1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0,2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0,3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0,4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0,5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0,6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0,7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0,8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0,9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1,0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1,1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1,2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1,3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1,4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319

FIGURA 126: VALOR DE LA PROBABILIDAD DE OBTENER UNA NOTA ENTRE 7,8 Y 9

Para ubicar el valor de 0.3577, debemos hacer lo siguiente:

En este caso “z” resultó ser **1.07**, los dos primeros valores (resaltados en negrita) se buscan en la columna izquierda y el segundo decimal (subrayado) en la fila superior, la “doble entrada” nos dará el valor a buscar.

Por tanto se ha determinado que la probabilidad de obtener una nota superior a la media pero no mayor a 9 será de 0.3577 o 35.77%.

Cabe recalcar que los porcentajes de probabilidad siempre se dan entre la Media y el valor de la variable del cual se desea conocer su probabilidad.

Veamos algunos casos:

Ejemplo 1

Una variable tiene comportamiento normal con $\mu = 10$ y $\sigma = 2$. Encuentre las probabilidades siguientes:

a) $p(x < 13.5)$ b) $p(x < 8.2)$ c) $p(9 < x < 10.6)$ d) $p(x > 15)$ e) $p(x > 5)$

Para resolver problemas referentes a la Distribución Normal, es muy importante hacer un dibujo en el cual se interprete, a través de un gráfico, la condición dada.

Según la condición del primer literal debe encontrarse la probabilidad de que un valor de la variable sea **menor a 13.5**, este valor es mayor a la Media por ello se coloca a la derecha y la región de probabilidad que nos dará la tabla según el valor “z” será el que esté entre 13.5 y 10; sin embargo esta región no expresa la totalidad de la condición, ya que otros valores menores a 13.5 están en el lado izquierdo de la curva; por tanto, y tomando en cuenta la característica simétrica de la curva, todos esos valores representan per se un 50% de probabilidad, es decir la probabilidad total de cumplir la condición estará dada por dos regiones: aquella que está entre la Media y 13.5 y la mitad de la curva bajo la Media.

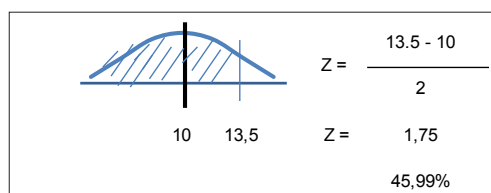


FIGURA 131: SOLUCIÓN LITERAL a

“z” da un valor de 1.75, al buscar en la tabla se determina una probabilidad de 0.4599, este valor en porcentaje será 45.99% y corresponderá a la región entre 10 y 13.5, por tanto, deberá sumarse el 50% de probabilidad de aquellos valores que siendo menores a 13.5 están en el otro lado de la curva.

Entonces para cumplir la condición de determinar la probabilidad de obtener un valor menor a 13.5 se debe sumar los dos porcentajes de probabilidad que cumplen dicha condición, es decir 45.99% más 50%, en símbolos sería así:

$$P(x < 13.5) = 45.99\% + 50\% = 95.99\%$$

El segundo caso pide encontrar la probabilidad siguiente: $p(x < 8.2)$, el gráfico sería así:

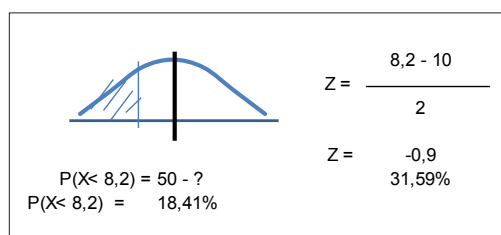


FIGURA 132: SOLUCIÓN LITERAL b

La región rayada es la que cumple la condición, pero hay que recordar que los porcentajes de probabilidad en tabla siempre dan la región comprendida entre la Media y el valor de la condición.

El porcentaje de probabilidad (31.59%) que se determina es el que está entre 10 y 8.2, pero no corresponde a la interpretación ni al gráfico, por tanto y dado que la mitad de la curva representa el 50%, la región que cumple la condición y que está señalada corresponderá a la diferencia entre este valor y el encontrado, en símbolos sería así:

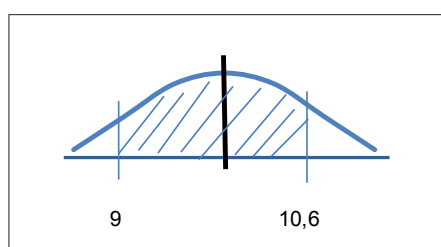


FIGURA 133: GRÁFICO DE LA TERCERA CONDICIÓN

$$P(x < 8.2) = 50\% - 31.59\% = 18.41\%$$

El tercer caso pide encontrar la probabilidad siguiente: $p(9 < x < 10.6)$ el gráfico sería así:

La región rayada cumple la condición, $(9 < x < 10.6)$ debe leerse “entre” los valores 9 y 10.6.

Para este caso se deben realizar los dos cálculos por separado, uno con el valor 9 y otro con 10.6, así:

$$\begin{array}{ll}
 Z = \frac{9 - 10}{2} & Z = \frac{10,6 - 10}{2} \\
 Z = -0,5 & Z = 0,3 \\
 \downarrow & \downarrow \\
 19,25\% & 11,79\%
 \end{array}$$

Los porcentajes encontrados representan la región de probabilidad entre cada valor y la Media, como esto coincide con la región señalada, entonces el resultado será la suma de ambos, entonces:

$$P(9 < x < 10.6) = 19.25\% + 11.79\% = 31.04\%$$

Lo solicitado en el cuarto caso: $P(x > 15)$, se traduce en el siguiente gráfico y cálculo:

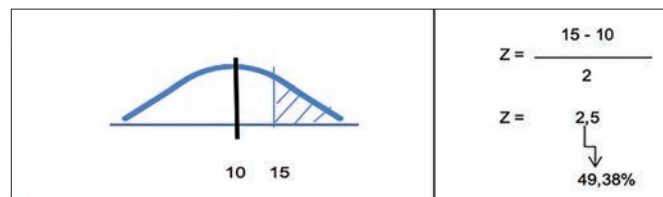


FIGURA 134: CUARTA CONDICIÓN

El porcentaje encontrado en tablas será la región entre la Media y 15 pero ese valor no corresponde a lo solicitado, por tanto, deberá restarse ese valor del 50% que corresponde a la mitad de la curva (por definición) y entonces:

$$P(x > 15) = 50\% - 49.38\% = 0.62\%$$

La última pregunta: $P(x > 5)$ expresada en la figura 131 y su cálculo será así:

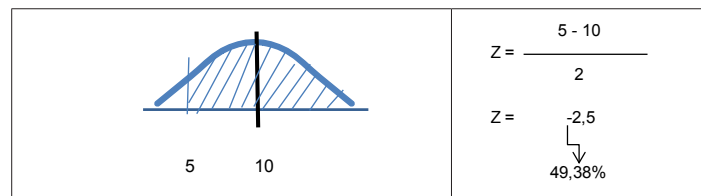


FIGURA 135: QUINTA CONDICIÓN

El valor de 49.38% indica la región de la curva entre 5 y 10 a esto hay que sumar el 50% correspondiente al resto de valores que están más allá de la Media y que cumplen la condición de ser mayores a 5, por tanto, el resultado de esta pregunta sería:

$$P(x > 5) = 49.38\% + 50\% = 99.38\%$$

EJERCICIOS DE APLICACIÓN DEL CAPÍTULO

RECOMENDACIÓN: En la práctica resulta más sencillo¹ realizar los ejercicios “a mano”, ya que la fórmula es muy simple, lo que sí es fundamental es realizar un dibujo que exprese la condición para cada ítem del ejercicio.

¹ En internet hay algunas aplicaciones que ayudan a realizar los cálculos

Ejemplo 2

Ud. es jefe de personal y le han pedido una opinión sobre la seguridad para utilizar el ascensor de la empresa. Luego de un estudio del uso de los elevadores, se conoce que si la capacidad es de 8 personas, la media del peso total es de 550 kg. con una $\sigma = 66.71$ kg. Cuál es la probabilidad de que el peso total de 8 personas: a) exceda de 590 kg.; b) sea menor de 400 kg.; c) esté entre 380 kg. y 430 kg.; d) esté entre 450 kg. y 600 kg.

Literal “a”

El gráfico indica la posición del valor “crítico” que en este caso es 590 y la región señalada nos hace notar que esa región cumple con la condición: el peso debe exceder de 590.

Al aplicar la fórmula se encuentra un valor de “Z” equivalente a 0.59, es decir el peso de 590 está a 59 centésimos (de desviación estándar) de distancia de la media y hacia la derecha (dado que el signo es positivo).

Se busca en la tabla de distribución normal (Anexo 1) y se encuentra un valor de 0.2224, mismo que representa la probabilidad de obtener un valor entre la media y el valor crítico; por lo tanto, ese valor deberá restarse de 0.5 (recuerde que la mitad de la curva representa el 50% del todo) para obtener la probabilidad complementaria que es la representada en el gráfico.

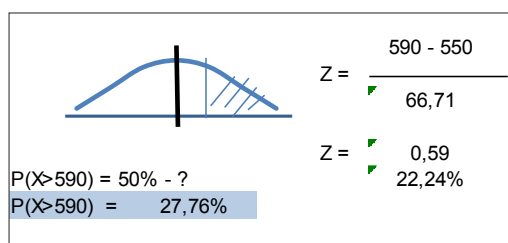


FIGURA 136: SOLUCIÓN LITERAL “a” EJEMPLO 2

De acuerdo al proceso entonces, la región que cumple la condición de que el peso exceda de 590 Kg. dará una probabilidad del 27.76% (diferencia entre 50% y 22.24%)

Literal “b”

El gráfico indica la posición del valor “crítico” que en este caso es 400 y la región señalada nos hace notar que esa región cumple con la condición: el peso debe ser menor a 400.

Al aplicar la fórmula se encuentra un valor de “Z” equivalente a -2,24 esa es la distancia desde la media hasta el peso de 400 hacia la izquierda (el signo negativo indica eso).

Se busca en la tabla de distribución normal (Anexo 1) y se encuentra un valor de 0.4875, pero ese porcentaje es la probabilidad de obtener un valor entre la media y el valor crítico; por lo tanto ese valor deberá restarse de 0.5 (la mitad de la curva representa el 50% del todo) para obtener la probabilidad complementaria que es la representada en el gráfico.

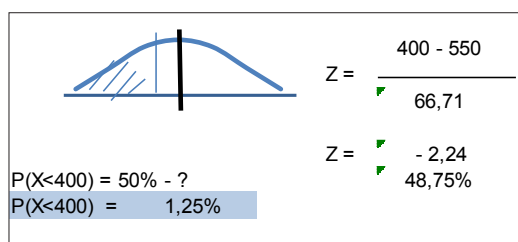


FIGURA 137: SOLUCIÓN LITERAL “b” EJEMPLO 2

De acuerdo al proceso entonces, la región que cumple la condición de que el peso sea menor de 400 Kg. dará una probabilidad del 1.25% (diferencia entre 50% y 48.75%).

Literal “c”

En este caso hay dos valores críticos, el gráfico indica eso, los dos valores son menores a la media por lo tanto se puede predecir ya que los valores de “Z” a encontrar serán negativos. La región señalada nos hace notar que esa parte de la curva indica la probabilidad a encontrar con la condición de que el peso total sea mayor a 380 pero menor a 430.

Al aplicar la fórmula para cada caso se encuentra un valor de “Z₁” equivalente a -2,54 esa es la distancia desde la media hasta el peso de 380 (el signo negativo era lo esperado); para el segundo punto crítico el valor de “Z” es de -1.79.

Se busca en la tabla de distribución normal los valores para cada caso y encontrándose un valor de 0.4945 para el primero y de 0.4633 para el segundo.

Cada uno de estos porcentajes establece una región de probabilidad entre la media y el respectivo valor, pero nos interesa exclusivamente la región señalada; por lo tanto, deberemos restar el porcentaje mayor del menor para obtener la probabilidad pedida, esto se establece en la figura siguiente.

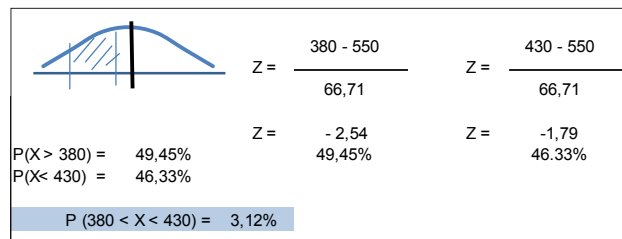


FIGURA 138: SOLUCIÓN LITERAL “c” PRIMER EJERCICIO DE APLICACIÓN

Literal “d”

En este caso también hay dos valores críticos pero estos se encuentran a distintos lados respecto a la media. La región señalada nos hace notar que al encontrar los valores “Z” y a su vez las regiones de probabilidad que generen, ambas serán parte de la solución (recuerde que los valores de tabla indican regiones de la curva entre la media y un valor crítico).

En este caso para encontrar el resultado según la condición dada, deberá sumarse la probabilidad que arroje cada una de las dos regiones, según se observa en la figura 139.

Al aplicar la fórmula para cada caso se encuentra un valor de “Z₁” equivalente a -1,49 esa es la distancia desde la media hasta el peso de 450 (el signo negativo es lo esperado); para el segundo punto crítico el valor de “Z” es de 0.74 (el signo obviamente debe ser positivo).

Se busca en la tabla de distribución normal los valores para cada caso y se encuentra un valor de 0.4319 para el primero y de 0.2704 para el segundo.

Cada uno de estos porcentajes establece una región de probabilidad entre la media y el respectivo valor, y como ambas regiones son las que cumplen al mismo tiempo la condición deberemos sumar dichos porcentajes para obtener la probabilidad pedida, esto se establece en la figura.

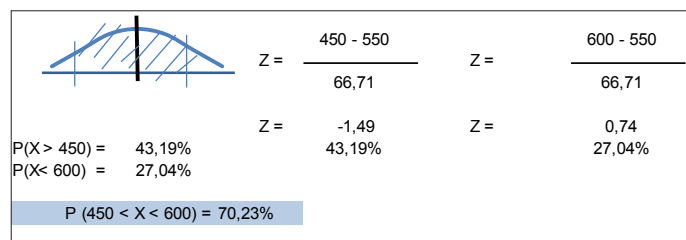


FIGURA 139: SOLUCIÓN LITERAL "d" PRIMER EJERCICIO DE APLICACIÓN

Ejemplo 3

La vida útil de una prueba psicológica tiene distribución aproximadamente normal con $\mu = 3.1$ años y $\sigma = 1.2$ años. a) Cuál es la probabilidad de que esta prueba haya sido aplicada dentro del primer año de salida al mercado b) cuál es la probabilidad de que esta prueba sea aplicada durante un tiempo no mayor a 5 años.

Literal "a"

El gráfico indica la posición del valor "crítico" que en este caso es un año y la región señalada nos hace notar que esa región cumple con la condición: se busca determinar la probabilidad de que la prueba se aplique antes de un año de haberse publicado.

Al aplicar la fórmula se encuentra un valor de "Z" equivalente a -1.75, al buscar en la tabla este valor corresponde al 0.4599 que sería la probabilidad de estar entre la media y el primer año, pero nos interesa lo que se conoce como la "cola" de la curva y ese valor se determina restando del 50% que corresponde a la parte izquierda de la campana según se muestra en la figura 140.

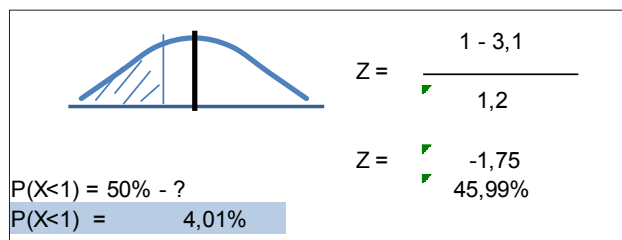


FIGURA 140: SOLUCIÓN LITERAL "a" TERCER EJEMPLO

Literal "b"

El gráfico indica la posición del valor "crítico" que en este caso es el quinto año, como la condición es buscar la probabilidad de que esa prueba sea aplicada hasta en cinco años, la

región señalada indica dos partes de la curva: la que equivale a la parte izquierda de la campana y la que corresponde a la región entre los cinco años y el valor de la media.

Como es lógico pensar, parte de la respuesta ya se conoce dado que la región de la izquierda representa el 50% por lo tanto solo hace falta calcular la región de probabilidad de la segunda región.

Al aplicar la fórmula se encuentra un valor de “Z” equivalente a 1.58, al buscar en la tabla este valor corresponde a 0.4429 que sería la probabilidad de estar entre la media y el quinto año, por lo tanto a este valor habrá que sumar el dato ya conocido y entonces la probabilidad de que dicha prueba se aplique dentro de los cinco primeros años será de 94.29% que es lo que se muestra en la figura 137.

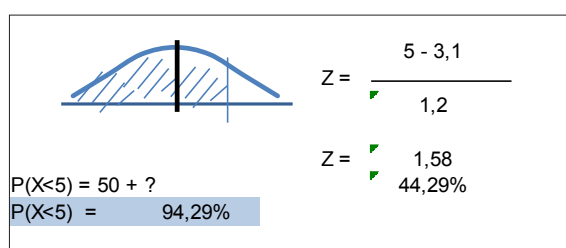


FIGURA 141: SOLUCIÓN LITERAL "b" TERCER EJEMPLO

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

- Para determinar ciertas aptitudes de los estudiantes de bachillerato de una institución educativa, se aplicó un test que permite determinar la capacidad de clasificar objetos en base a varias características, la distribución de los resultados tuvo un comportamiento normal con media de 100 y desviación estándar de 20
 - Qué porcentaje de los puntajes se halla entre 75 y 90
 - Si para ser parte del grupo con aptitudes especiales se requiere una puntuación mínima de 115, qué probabilidad hay de ser parte de los escogidos.
- En un examen de Estadística calificado sobre 100, la media fue de 78 y la varianza 100. Determinar las probabilidades de obtener:
 - Máximo 93
 - No más de 62
 - Más de 80 pero máximo 90
- A 300 estudiantes se les aplicó una prueba de motricidad cuya distribución fue normal con media 68 y desviación estándar 9
 - Cuántos estudiantes obtuvieron un puntaje mayor a 72
 - Cuántos estudiantes obtuvieron un puntaje menor o igual a 60
 - Si para ser aceptado en un curso especial se necesita un puntaje de al menos 75, cuál es la probabilidad de ser rechazado

4. En una prueba de atención (concentración) a 80 niños se determinó una media de 2.3 minutos y una desviación estándar de 0.41 minutos y la distribución resultó ser simétrica. Determinar las probabilidades siguientes: a) que se concentre tan solo durante 1.5 minutos b) si se considera que una concentración más allá de 3 minutos es extraordinaria, cuál es la probabilidad de que esto ocurra, c) que un niño se concentre entre 1 y 2 minutos, d) que un niño se concentre no más de 4 minutos, e) si a los niños que se distraen al primer minuto se les envía a un programa especial, cuántos estarían en ese caso.
5. Se conoce que un test psicológico es calificado sobre 100 puntos y que está dirigido para personas adultas. Los siguientes son los resultados de dos grupos separados por género. Determine las MTC y si considera que la distribución es normal encuentre las probabilidades siguientes: a) de que un hombre o una mujer obtengan menos de 50, b) que una mujer obtenga más de 80, c) que un hombre obtenga más de 70 d) la probabilidad de que las mujeres superen el promedio de los varones

MUJERES

50	80	82	53	82	81	81	70
53	63	72	81	82	72	80	70
82	64	80	63	81	73	84	71
60	63	53	60	80	70	75	60
70	64	81	67	76	72	82	75
75	61	79	65	75	70	78	71
44	70	60	63	83	72	72	60
71	72	81	65	72	53	80	73
80	64	23	75	50	60	75	70
82	53	79	65	45	60	69	73

HOMBRES

57	86	85	85	74	54	75	82
85	86	76	84	56	57	67	76
67	85	77	88	75	82	68	84
80	84	74	84	74	40	67	57
71	80	72	86	83	74	68	70
69	79	74	82	75	80	65	81
67	87	76	76	64	48	74	64
69	76	80	84	77	75	76	85
79	54	64	79	74	84	68	82
69	85	64	73	77	86	57	83

6. En una muestra de 1000 casos, la media de una cierta prueba psicológica era de 14.4 y varianza de 6.25
- Suponiendo que los puntajes se distribuyen normalmente, calcular:
 - Cuántos individuos logran puntajes entre 12 y 16
 - Cuántos tendrán al menos 18
 - Cuántos tendrán un puntaje no mayor a 6
 - Cuál es la probabilidad de que un individuo elegido al azar logre un puntaje mayor a la moda
7. Supongamos que un conocido nos dice que ha obtenido en un test de inteligencia una puntuación CI igual a 95. Asumiendo que las puntuaciones en dicho test de inteligencia

se distribuyeron simétricamente y sabiendo que estas tienen media 100 y desviación estándar 6 a) ¿qué le diría a usted respecto a esta situación en referencia a la población? b) ¿qué porcentaje de sujetos es de esperar que obtengan un valor inferior o igual a la moda? c) ¿qué porcentaje de sujetos es de esperar que obtengan un valor superior a 115? d) ¿entre qué valores de CI se encuentra el 50% central de los sujetos?

8. Los datos a continuación se refieren a una investigación sobre los niveles de destreza en jóvenes entre 15 y 20 años, esta variable se mide sobre 120 puntos y se conoce que los valores medios están entre 95 y 105. Si considera que la distribución es normal y si se espera también que una quinta parte de la muestra supere un valor de 110, ¿se cumple esta expectativa? Se pide también determinar las siguientes probabilidades: que algún joven obtenga una puntuación no menor a 100, que obtenga una puntuación de al menos la media, que obtenga una puntuación no menor a 98 ni mayor a 110, que obtenga una puntuación menor a una desviación estándar de la media.

96	94	84	116	110	99	106	119	90	71	116	98
96	94	84	105	110	99	75	110	120	108	116	107
88	94	84	105	110	99	101	91	90	96	95	80
89	94	100	113	111	108	78	97	103	108	85	101
115	94	85	105	111	90	104	119	90	119	116	85
100	94	116	101	90	99	117	97	109	79	117	101
100	94	85	106	111	86	82	104	91	108	95	101
107	94	85	106	112	97	103	97	79	107	117	79
108	112	115	106	108	98	98	78	116	76	92	102
100	112	86	106	91	100	80	79	91	109	117	102
101	95	86	71	112	95	102	97	116	81	107	119
101	105	86	106	112	74	118	97	91	109	118	102
86	95	116	106	101	100	82	98	98	119	96	73
107	106	87	75	113	74	97	82	91	110	118	103
107	95	87	106	90	74	119	98	78	118	97	106
100	75	88	107	114	92	104	118	92	81	75	103
100	95	88	72	114	99	79	98	83	93	119	98
95	95	115	78	91	99	119	77	100	119	80	103
95	101	77	99	118	100	108	98	92	96	119	83
100	95	89	107	105	90	117	80	92	75	78	99
93	108	89	76	116	98	83	119	79	109	120	104
100	96	89	107	91	98	77	82	93	119	95	71
89	71	117	70	116	97	102	98	103	70	120	100
89	96	92	107	120	101	82	70	71	105	84	71
113	96	120	111	92	101	104	98	93	119	119	105

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1

Como primera consideración, debe asegurarse que la distribución de los datos es simétrica (normal), caso contrario no se podría desarrollar.

En este caso, la redacción del ejercicio confirma que efectivamente los datos se distribuyen simétricamente por tanto se asegura el comportamiento normal, lo cual supone que las medidas de tendencia central son iguales o muy similares entre sí.

Como se había sugerido, es importante establecer un gráfico según la condición dada, esto se presenta a continuación en las figuras 142 y 143, en las cuales también aparece el desarrollo de cada uno de los literales del ejercicio.

Literal “a”

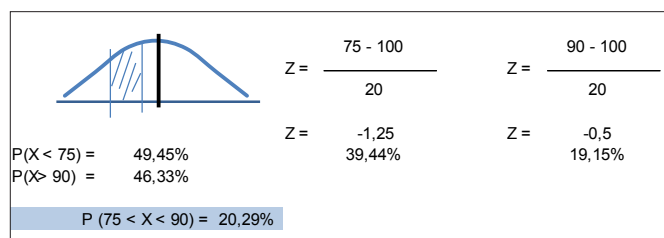


FIGURA 142: SOLUCIÓN LITERAL “a”

Literal “b”

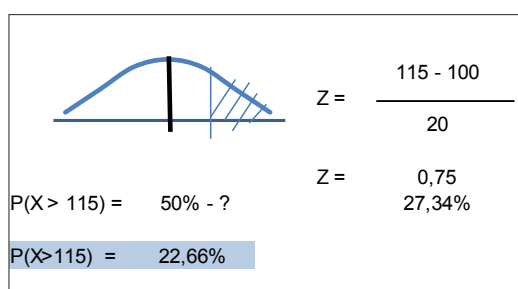


FIGURA 143: SOLUCIÓN LITERAL “b”

Ejercicio 3

Literal “a”

El gráfico indica que el valor crítico está a la derecha de la media y que la región de probabilidad a encontrar es la “cola” de la curva, por tanto, al encontrar el porcentaje que corresponda en la tabla al valor de “z”, se deberá restar del 50%. En la figura 144 se ha desarrollado el ejercicio en su totalidad.

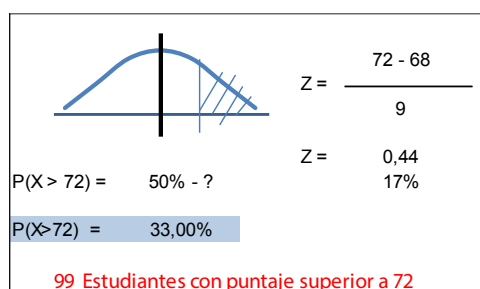


FIGURA 144: SOLUCIÓN LITERAL “a”

Pero este ejercicio solicita determinar el número de estudiantes que cumplen la condición y no exclusivamente la probabilidad y, dado que hay un total de 300, el 33% que tienen puntajes superiores a 72 equivaldrá a 99 estudiantes.

Literal “b”

El gráfico indica que el valor crítico está a la izquierda de la media y que la región de probabilidad a encontrar es la “cola” de la curva, por tanto, al encontrar el porcentaje que corresponda en la tabla al valor de “z”, se deberá restar del 50%. La figura 145 muestra el desarrollo del ejercicio según la condición dada.

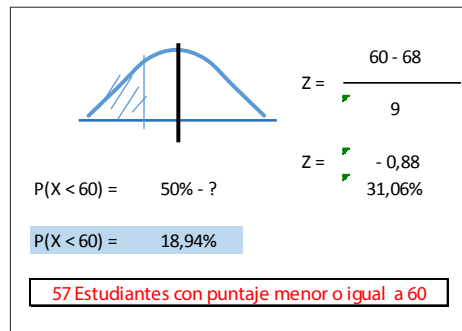


FIGURA 145: SOLUCIÓN LITERAL “b”

Al igual que el literal anterior este ejercicio solicita determinar el número de estudiantes que cumplen con tener puntajes iguales o inferiores a 60 y no exclusivamente la probabilidad, en este caso se determina que el 18.94% arroja un valor de 56.82; pero como es lógico el resultado final no puede tener decimales (no hay 82 centésimas de estudiante); se recomienda siempre que cuando esto ocurra se redondee al entero superior, sin importar el valor de la parte decimal, es decir que si en un caso el cálculo diera por ejemplo 35.02, igual se redondee a 36 el resultado.

Literal “c”

Para encontrar la solución hay que tomar en cuenta la condición de este ejercicio y la pregunta en sí; el valor crítico es para ser aceptado, pero preguntan la probabilidad de ser rechazado.

En la figura siguiente se muestra el desarrollo del literal y la parte rayada indica el área de la curva que cumple con lo solicitado en la pregunta.

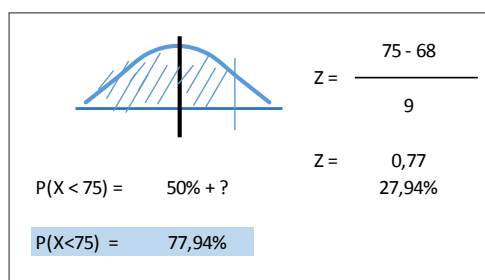


FIGURA 146: SOLUCIÓN LITERAL "c"

Al porcentaje de 27.94% hay que sumar el 50% adicional, dado que a partir de un puntaje de 75 se acepta, pero preguntan la probabilidad de ser rechazado y por tanto de allí hacia valores inferiores cumplen con lo pedido en el literal.

Ejercicio 5

Literal "a"

Este ejercicio presenta valores numéricos de los resultados de un test aplicado a dos grupos de personas y en la redacción se pide encontrar los valores de las MTC y comprobar si la distribución cumple con la condición para ser considerada simétrica.

Al desarrollar el ejercicio y encontrar los valores de Media, Mediana y Moda se encuentra para cada grupo lo siguiente:

Mujeres: Media: 69.32 Mediana: 71.5 Moda: 70

Hombres: Media: 74.2 Mediana: 76 Moda: 74

Con estos valores se debe decidir si la distribución para cada caso es normal.

Como la máxima diferencia en el caso de las mujeres es de 2.175 puntos (entre mediana y media) y en el caso de los hombres es de 2 puntos; estos valores hacen ver que las medidas de tendencia central son muy cercanas entre sí y por tanto se puede decir que la distribución sí es normal.

En ese caso procedemos a resolver el ejercicio según cada una de las condiciones dadas.

Literal "a"

En la figura 147 se expone el proceso para cada caso

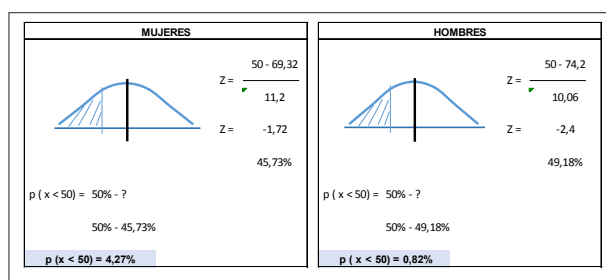


FIGURA 147: SOLUCIÓN DEL LITERAL "a" PARA CADA CASO

Según esto, es más probable que una mujer obtenga menos de 50 respecto a la probabilidad de que algún hombre lo haga.

Literal “b”

El valor crítico es 80 y se pide la probabilidad de obtener un puntaje superior a 80, por lo tanto interesa la parte rayada de la curva según el gráfico presentado en la figura 148

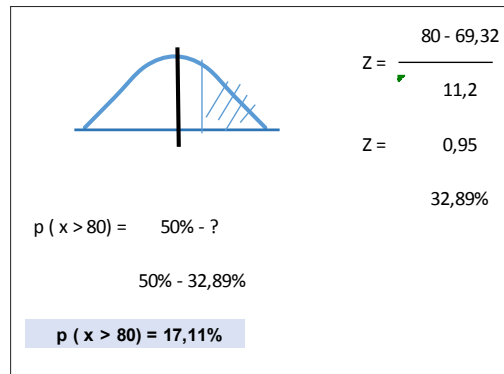


FIGURA 148: SOLUCIÓN LITERAL “b”

Entonces la probabilidad de que una mujer obtenga un puntaje superior a 80 en el test es de 17.11%

Literal “c”

Siendo el valor crítico menor a la media del grupo de los hombres, el valor de “z” será negativo y la región indicada en la figura 149 será la que cumpla la condición.

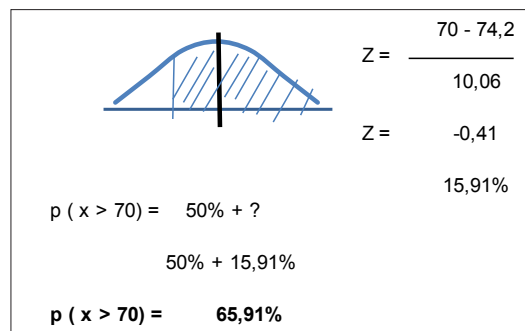


FIGURA 149: SOLUCIÓN LITERAL “c”

Literal “d”

Esta condición es interesante dado que no nos dan explícitamente un valor crítico, pero dado que ya se conocen los resultados de las medidas de tendencia central de cada grupo, la pregunta se transforma en lo siguiente: encontrar la probabilidad de que las mujeres superen un puntaje de 74.2.

En la figura 150 se desarrolla este literal.

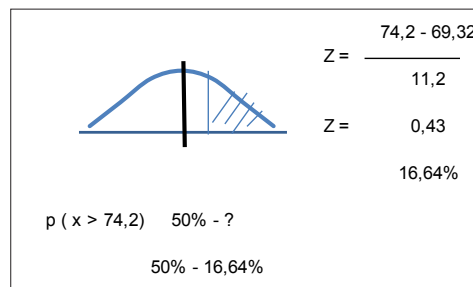


FIGURA 150: SOLUCIÓN LITERAL "d"

Ejercicio 7

Literal "a"

Supongamos que un conocido nos dice que ha obtenido en un test de inteligencia una puntuación CI igual a 95. La decisión dependerá de cada profesional, pero como sugerencia se debe tomar en cuenta este valor está a 0.83 valores "z" de la media, es decir a menos de una desviación estándar, lo cual le ubica dentro del intervalo medio.

Literal "b"

Esta pregunta es suspicaz ya que no tenemos datos para conocer el valor de la moda, pero aplicando la teoría de distribución normal hay que suponer que tanto media como mediana y moda tienen el mismo valor y que cualquiera de ellas divide a su vez en dos partes iguales a la curva.

Por ello da lo mismo utilizar cualquiera de ellas, es más, en realidad no interesa ni siquiera el valor como tal dado que la probabilidad de tener un valor inferior a la media (moda o mediana) es la misma que la de ser superior a cualquiera de ellas.

Para este caso entonces no hace falta dibujar la condición y el resultado será que el 50% de los sujetos tendrá la probabilidad de obtener un valor de C.I. inferior a la moda.

Literal "c"

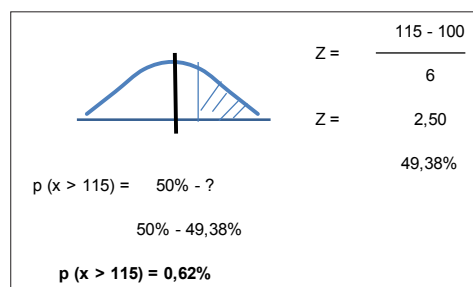


FIGURA 151: SOLUCIÓN LITERAL "b"

Literal “d”

50% central significa que hay 25% de probabilidad simétricamente repartido alrededor de la media como se muestra en la figura 152

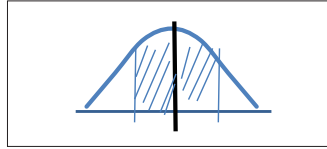


FIGURA 152: POSICIÓN DE VALORES “X” CADA UNO A CIERTA “DISTANCIA” DE LA MEDIA

En este caso nos están pidiendo los valores específicos del resultado del test que cumplan con la condición dada.

Para ello es necesario despejar, de la fórmula de la distribución normal, el valor de “x”

$$x = z \delta + \mu$$

De ella no conocemos el valor de “z” pero sí el valor de la probabilidad, por tanto, hay que buscar en la tabla el valor más cercano a una probabilidad del 25%; dicho valor es 0.2486 y corresponde a un valor de “z” de 0.67.

Pero tenemos dos valores que cumplen la condición a cada lado de la media, entonces para aplicar la fórmula hay que tomar en cuenta que para encontrar el valor x_1 se debe usar un “z” negativo.

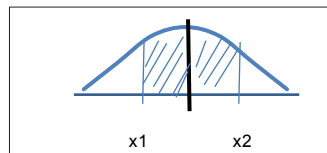


FIGURA 153: POSICIÓN DE LOS VALORES “X” ALREDEDOR DE LA MEDIA

Utilizando la fórmula para cada caso tenemos:

Para una probabilidad de 25%, el valor de Z es 0.67

Fórmula a utilizar: $X = Z \cdot \sigma + \mu$

$$X_1 = -0.67 \cdot 6 + 100 \quad X_2 = 0.67 \cdot 6 + 100$$

$$X_1 = 95.98 \quad X_2 = 104.02$$

Estos dos valores garantizan que el 50% de los sujetos que se encuentran a la misma distancia de la media a cada lado tendrán cocientes intelectuales entre 95.98 y 104.02.

CAPÍTULO 8:

ESTADÍSTICA INFERENCIAL

La falacia del cuadro estadístico estriba en que es unilateral, en la medida en que representa sólo el aspecto promedio de la realidad y excluye el cuadro total. La concepción estadística del mundo es una mera abstracción, y es incluso falaz, en particular cuando atañe a la psicología del hombre.

Carl Jung

La diferencia entre la Estadística Descriptiva y la Inferencial, radica — a mi juicio — en que la primera hace énfasis en el estudio de una sola variable a analizar a través de la muestra para tratar de explicar su comportamiento y con ello poder expresar comentarios y pocas conclusiones que podrían extenderse a la población siempre y cuando la técnica de muestreo utilizada sea correcta y que la Estadística Inferencial procura encontrar elementos de juicio que le permitan determinar si lo que ocurre en la muestra es aplicable — con cierto grado de confiabilidad — a la población que ha sido objeto de estudio relacionando al menos dos variables que luego de un análisis más completo poder inferir, pronosticar, comentar y concluir con mayor certeza. En pocas palabras la primera se suscribe a la muestra y la segunda hace hincapié en la población.

En este punto debo hacer un comentario respecto a los distintos libros de Estadística que encontramos en el mercado: estos libros suelen tratar los temas para hacer énfasis en nuevos conceptos, y por tanto muy pocos hacen un alcance cuando se refiere a “unir” temas que son complementarios en un análisis estadístico, me refiero a la necesidad de que en un análisis de relación no debe excluir la parte descriptiva.

En referencia a lo “descriptivo” e “inferencial”, por ejemplo, debemos tener claro que si bien es cierto ambos tratan de “descubrir” lo que ocurre en una población en función de muestras, el primero tan sólo se refiere al comportamiento de una variable y trata de explicarla en función de medidas muy específicas (tendencia central y dispersión fundamentalmente) y el segundo establece relaciones de una o más variables vs. otra que es objeto de estudio para poder hacer inferencias de la población con mayor o menor grado de confiabilidad, pero nada más.

En la experiencia de estos años a cargo de la materia de Estadística, he procurado unir estas dos grandes secciones de la Estadística diciendo a mis estudiantes que así como un

análisis descriptivo basado tan solo de medidas de tendencia central es muy pobre, analizar solamente la relación entre variables también lo es y por ello considero necesario que en todo análisis que se pretenda realizar entre dos o más variables debe hacerse también una descripción amplia de cada una de ellas para así tener más y mejores elementos de juicio.

EVALUACIÓN PREVIA DE PROYECTOS (EPP)

Para iniciar el estudio de la Estadística Inferencial con aplicación a la Psicología y Educación voy a intentar explicar lo que he llamado “Evaluación Previa de Proyectos” (EPP), que no es otra cosa que el análisis lógico que se da entre variables para conseguir un objetivo a partir de la Lógica Matemática.

Esta técnica vincula los conceptos de la Lógica Proposicional, transformándolos en Lógica Matemática, en la siguiente figura se resume esto, y luego se hace una aplicación a la Psicología con variables vinculantes dentro de ella; tomando en cuenta lo que se resume en la figura 154:

Lógica Formal	Lógica Matemática
Proposiciones (premisas)	Variables
Conectores lógicos (y , o)	Operadores numéricos [producto (*), suma (+)]
Valores de verdad (V , F)	Valores numéricos (1 , 0)
Combinaciones posibles : 2^n donde "n" es el número de variables analizadas	

FIGURA 154: TRANSFORMACIÓN DE LA LÓGICA PROPOSICIONAL A LÓGICA MATEMÁTICA

Hay que establecer y recordar lo siguiente:

1. Cada premisa debe poder ser evaluada como verdadera (V) o falsa (F), en otras palabras, que puede darse o no.
2. El conector lógico “y” significa la obligatoriedad de que las premisas relacionadas con él se cumplan al mismo tiempo y el conector lógico “o” indica que con cualquiera de las dos premisas que se den, el resultado es favorable, por ejemplo:

p: hoy es viernes (puede ser V o F)

q: hoy voy al cine (puede ser V o F)

Por tanto uniendo dichas premisas con cada conector se tiene los siguientes resultados:

PREMISAS		Proposición compuesta
p	q	$p \wedge q$
V	V	V
V	F	F
F	V	F
F	F	F

PREMISAS		Proposición compuesta
p	q	$p \vee q$
V	V	V
V	F	V
F	V	V
F	F	F

TABLA 62. PREMISAS CONECTADAS CON “y” Y CON “o”

CÓMO TRABAJA ESTE SISTEMA

El auténtico problema no es si las máquinas piensan,
sino si lo hacen los hombres
(B. F. Skinner)

En la Lógica proposicional solo pueden existir resultados de V y F sin importar la relación que se proponga entre las variables, pero con este sistema transformando a la Lógica Matemática podemos encontrar resultados superiores a 1 (V) ya que el conector lógico “o” permite sumar, esta es una de las grandes diferencias con la lógica proposicional ya que encontrar un valor “V” (verdadero) en realidad no dice mucho como lo explicaré más adelante.

¿Qué significa un resultado verdadero en la lógica proposicional? Pues no otra cosa que al combinar las premisas con los conectores usados en las proposiciones compuestas el objetivo se cumple, pero no nos dice si el resultado es lógico (aunque la teoría lo suponga así).

La evaluación de la combinación realizada en la proposición compuesta con los distintos conectores lógicos solo se hace en función de los resultados (verdaderos y falsos) alcanzados y no de manera detallada para saber si cada resultado responde o no al objetivo propuesto.

Tampoco se hace un análisis lógico y numérico en cuanto a la diferencia entre un resultado verdadero y otro ya que en realidad no es lo mismo cumplir con el objetivo cuando todas las variables se hicieron presentes o hacerlo cuando una o más fallaron.

Pueden hacerse muchísimas combinaciones entre las variables intervinientes y se debe establecer previamente cuáles son variables fundamentales y cuáles pueden ser complementarias.

Los conectores lógicos al transformarse en operadores numéricos permitirán encontrar resultados mayores a 1 que equivaldría al de verdadero en la lógica proposicional; esto puede llamar la atención, pero en realidad eso daría una pauta de la posibilidad lógica de que dicho resultado se dé con mayor razón.

La diferencia entonces radica en que el conector “y” obliga a que se produzcan los dos hechos y en cambio con el conector “o” basta que uno de los dos ocurra para que el resultado sea favorable.

Para explicar esta técnica de EPP y visualizar la diferencia y alcance de esta propuesta pongo a consideración un ejemplo sencillo. En muchos libros de Lógica Proposicional se proponen ejercicios como el siguiente: “encontrar el resultado de la siguiente proposición compuesta: $(p \wedge q \wedge r) \vee s$ ”; es decir que a partir de cuatro premisas en donde se establecen combinaciones posibles entre ellas, se propone una fórmula lógica en la cual las tres premisas iniciales se operarán con el conector \wedge (“y”) y la cuarta se añade con el conector \vee (“o”); el resultado se muestra en la tabla 63:

PREMISAS				COMBINACIÓN (proposición compuesta)
p	q	r	s	p y q y r o s
				En símbolos sería: (p ∧ q ∧ r) ∨ s
V	V	V	V	V
V	V	V	F	V
V	V	F	V	V
V	V	F	F	F
V	F	V	V	V
V	F	V	F	F
V	F	F	V	V
V	F	F	F	F
F	V	V	V	V
F	V	V	F	F
F	V	F	V	V
F	V	F	F	F
F	F	V	V	V
F	F	V	F	F
F	F	F	V	V
F	F	F	F	F

TABLA 63. COMBINACIÓN DE PROPOSICIONES CON LOS CONECTORES “y” y “o”

El ejercicio termina ahí, pero normalmente no se hace un análisis más profundo de la lógica de los resultados finales, además el objetivo es tan solo encontrar el resultado de dicha combinación de manera muy abstracta. El sistema de análisis propuesto (EPP) pretende ir más allá utilizando las mismas reglas del juego de la Lógica Proposicional, pero en procesos tangibles.

Para ello se propone un objetivo específico (proyecto) y se determinan variables (que equivaldrían a las premisas) mismas que tendrían relación con dicho objetivo y estas deberán combinarse con los conectores lógicos (operadores numéricos) estableciéndose su carácter de verdad con valores numéricos (1 y 0).

A continuación, expongo un ejemplo simple en donde una novia desea saber cuál será la combinación perfecta entre cuatro variables para que su ceremonia se realice con éxito:

Objetivo (proposición compuesta): realizar una boda.

Variables (premisas): Novio, Dinero, Documentos e Invitados.

Con claridad se puede observar que cada variable solo puede tener dos caracteres de verdad, es decir se dan o no (existen o no).

¿Le parece a usted que estas son las únicas variables a tomar en cuenta?, ¿considera que son las variables precisas?, si no es así puede hacer el ejercicio proponiendo otras o más variables; para efectos de este ejemplo serán suficientes.

Propongo la misma combinación simple propuesta antes para entender la diferencia (recuerde también que el conector “∧” se transforma en producto (*) y el conector “∨” se transforma en suma (+):

VARIABLES				COMBINACIÓN (proposición compuesta)
Novio (N)	Dinero (Di)	Documentos (D)	Invitados (I)	N y Di y D o I
				En fórmula sería: $N \wedge Di \wedge D \vee I$
1	1	1	1	2
1	1	1	0	1
1	1	0	1	1
1	1	0	0	0
1	0	1	1	1
1	0	1	0	0
1	0	0	1	1
1	0	0	0	0
0	1	1	1	1
0	1	1	0	0
0	1	0	1	1
0	1	0	0	0
0	0	1	1	1
0	0	1	0	0
0	0	0	1	1
0	0	0	0	0

TABLA 64. CONTINUACIÓN DE VARIABLES PARA ANALIZAR LA REALIZACIÓN O NO DE UNA BODA

Según la fórmula presentada, la novia considera que las tres primeras variables deben darse al mismo tiempo (por ello usa el conector “y”), y que la variable de los invitados no importa si se presenta o no; dentro de los esquemas prácticos tiene razón, aunque no parece tenerla tomando en cuenta las costumbres sociales, ¿no le parece?

Podrá notar que, en los resultados, en la primera fila, hay un valor no común y se preguntará qué relación tiene con V o con F, los otros resultados que son 1 o 0 obviamente se interpretarán igual $1 = V$ y $0 = F$.

Pasemos a interpretar ese primer valor, numéricamente es lógico y correcto porque operativamente da ese valor, pero ¿cómo debemos interpretar?, pues bien, dentro de todas las combinaciones posibles es la que más posibilidad tiene del éxito de su objetivo ya que todas las variables implicadas se dan, por tanto, es obvio que ese valor sea el más alto y esto significaría que la novia tiene “garantía” de que su proyecto se cumple satisfactoriamente.

Respecto a los otros resultados ocurre que son iguales a los de la combinación abstracta realizada con las premisas en la tabla 63 (la fórmula es la misma), pero ahora ya podemos hacer un análisis más minucioso sobre cada resultado y tendríamos lo siguiente:

VARIABLES				COMBINACIÓN (proposición compuesta)	Análisis de los resultados
Novio (N)	Dinero (Di)	Documentos (D)	Invitados (I)	N y D y Di o I	
				En fórmula sería: $N \wedge Di \wedge D \vee I$	
1	1	1	1	2	
1	1	1	0	1	
1	1	0	1	1	No lógico
1	1	0	0	0	
1	0	1	1	1	No lógico
1	0	1	0	0	
1	0	0	1	1	No lógico
1	0	0	0	0	
0	1	1	1	1	No lógico
0	1	1	0	0	
0	1	0	1	1	No lógico
0	1	0	0	0	
0	0	1	1	1	No lógico
0	0	1	0	0	
0	0	0	1	1	No lógico
0	0	0	0	0	

TABLA 65. ANÁLISIS DE CADA RESULTADO

¿Por qué no son lógicos esos siete valores?, en rigor los resultados son los mismos que los encontrados en la tabla 63 y eso es lo que la Lógica Proposicional acepta.

La diferencia está en las condiciones básicas, según la novia las variables Novio, Dinero y Documentos TIENEN que estar presentes y en los casos señalados algunas de ellas no están, es más, con las últimas ocho combinaciones ya se podía pronosticar que no puede realizarse la boda (es fácil darse cuenta el porqué), ya que en este caso hay una variable (el novio) que es totalmente imprescindible y sin embargo, según la lógica proposicional, en cuatro de ellos, ¡el resultado es favorable!, ¿esto no es lógico verdad?

La pregunta inmediata que surge es que eso va a ocurrir siempre dado que, si hay una variable básica, cuando esta no se presente cualquier combinación con ella resultará en un resultado fallido.

Pero esto ocurre en temas como el del ejemplo propuesto pero pocas veces ocurrirá en temas de diagnóstico de patologías, trastornos, síndromes o en casos de deserción laboral o en temas referentes a desarrollo cognitivo, o en otros que impliquen rendimiento académico ya que en unos u otros las variables serán distintas y de cierta importancia según cada caso y situación específica y no necesariamente se presentarán o serán indispensables para todos los casos y por tanto se podrán realizar combinaciones muy complejas que permitan establecer la fórmula “perfecta”, esto no significa tampoco que para intentar resolver alguna situación se deban crear cuadros como los presentados ni mucho menos.

Esas combinaciones complejas son las que en realidad se producen cerebralmente cuando al profesional de cualquier ciencia le presentan un caso a resolver, esa persona escoge las variables y las combina en función de la importancia y se hace una primera idea del camino a seguir para resolver el problema; todo esto depende de dos cosas en su orden de importancia: la experiencia y el conocimiento teórico.

Por ejemplo, si a un docente se le presenta un problema de un alumno con dificultades de aprendizaje, tendrá que establecer las variables que condujeron a eso y mirar el problema de manera holística; pero no necesitará hacer un cuadro para (según su experiencia) encontrar una “fórmula” de solución en la cual se hayan tomado en cuenta las variables implícitas; claro está que deberá pesar las más fuertes.

Lo que no tomamos en cuenta a diario es que en realidad esto lo hacemos siempre en nuestro cerebro y que dependiendo de la experiencia y conocimiento del tema encontraremos la solución (fórmula) con mayor o menor dificultad y rapidez.

Lo importante entonces de este análisis está en tratar de encontrar una combinación de variables que nos permita “garantizar” el logro de nuestro objetivo de la mejor manera posible. Por ejemplo, hay que decidir sobre cuáles son las variables precisas que se relacionan con el objetivo, de ellas cuáles son las más importantes, cómo deben estar “combinadas” dichas variables, etc.

Esta breve introducción nos permite ver lo siguiente:

1. Es necesario determinar con claridad cuál es el objetivo de un estudio estadístico.
2. Es imprescindible establecer las variables que realmente se relacionan con dicho objetivo.
3. Es importante la forma de combinar las variables para lograr una “fórmula perfecta”, es decir una fórmula que permita conseguir el objetivo de la mejor manera.
4. La expresión lógica propuesta, debe llevarnos a formular una propuesta que permita pronosticar el éxito o fracaso del objetivo planteado.

Esta expresión será más adelante una función – lineal o no – que permitirá pronosticar valores de la variable objetivo.

¿Cuál es entonces la implicación de esto a la Estadística Inferencial y cómo se relaciona con la Psicología o la Educación?

Lo que hemos llamado objetivo o proyecto, para la Estadística Inferencial será la **Variable Dependiente**, es decir algún síndrome, trastorno e incluso patología o también temas relacionados con la dificultad de aprendizaje de los estudiantes o de desarrollo; de cada uno de ellos será necesario conocer cuáles son sus causas; las premisas se transformaron ya en variables y se llamarán **Variables Independientes**, la proposición compuesta o fórmula lógica será una ecuación que dentro del proceso de investigación tanto dentro de lo Psicológico como pedagógico pasará a ser el resultado del diagnóstico.

En términos específicamente estadísticos presento a continuación un cuadro que resume la relación entre este sistema de análisis y el análisis de Regresión Lineal que lo estudiaremos más adelante.

	EVAL. PREVIA PROYECTOS	RELACIÓN DE VARIABLES	PROCESO PSICOLÓGICO	PROCESO EDUCATIVO
OBJETIVO	NECESIDAD DE ESTUDIO	VARIABLE DEPENDIENTE	PATOLOGÍA, TRASTORNO, SÍNDROME	PROBLEMAS DE APRENDIZAJE
VARIABLES	ELEMENTOS QUE INFLUENCIAN	VARIABLES INDEPENDIENTES	SÍNTOMAS	COMPAÑEROS, PADRES, PROFESORES
TIPO DE RELACIÓN	MÁS / MENOS IMPORTANTE	DIRECTA / INVERSA	CONFIRMACIÓN CON LA TEORÍA	DESEMPEÑO EN AULA
FUERZA DE RELACIÓN	Y / O	COEF. CORRELACIÓN Y DETERMINACIÓN	SÍNTOMAS MÁS EVIDENTES QUE OTROS	PROCESOS DENTRO DE AULA Y EN CASA
ERRORES	MAL USO O COMBINACIÓN DE ELEMENTOS	POSICIÓN (Se) INCLINACIÓN (Sb)	MALA INTERPRETACIÓN DE LOS SÍNTOMAS	FALTA DE OBSERVACIÓN A TIEMPO
RELACIÓN	FÓRMULA "PERFECTA"	ECUACIÓN	DIAGNÓSTICO	INTERVENCIÓN

FIGURA 155: APLICACIÓN Y RELACIÓN DE LA LÓGICA MATEMÁTICA A PROCESOS ESPECÍFICOS

CAPÍTULO 9:

RELACIÓN ENTRE VARIABLES

REGRESIÓN SIMPLE

Existe correlación negativa entre recuperación y terapia psicológica:
a más terapia psicológica, menor recuperación del paciente
(Hans Eysenck)

La creatividad está relacionada con nuestra capacidad
para encontrar nuevas respuestas ante viejos problemas
Martin Seligman

En los procesos de investigación en educación, en muchas ocasiones nos interesa conocer la posible relación que se puede manifestar entre dos o más variables. Para avanzar en el estudio de los hechos y fenómenos educativos buscamos la posible influencia que se puede dar entre aquellas variables que pueden intervenir en los resultados del aprendizaje y en otros ámbitos de la actividad educativa (Pérez Juste, García Llamas, Gil Pascual, & Galán González, 2009, p. 129).

Como hemos visto en un apartado anterior, uno de los temas básicos de la Estadística Inferencial es el de relacionar variables para poder predecir o pronosticar ciertos comportamientos de lo que será el objetivo de estudio.

Pero al igual de lo que se recomienda para la EPP, se debe tener mucho cuidado en seleccionar las variables que en verdad tienen relación con la problemática a estudiar; por ejemplo, si quiere hacer un estudio sobre el cociente intelectual podrá escoger variables como: la calidad de la alimentación, si hubo o no estimulación temprana, factores genéticos, ambiente entre otras; pero también podría escoger el color de ojos, el barrio en que habita, etc., pero como supondrá hay una gran diferencia entre las primeras y las dos últimas variables propuestas; de todas maneras estadísticamente siempre encontrará una fórmula que relacione cualquier variable con el cociente intelectual; es por ello que la determinación de las variables cobra una gran importancia para no cometer errores de diagnóstico por ejemplo.

Entonces el análisis de regresión no es otra cosa que establecer un modelo matemático (modelo lineal para efectos y alcance de este libro) en el cual se determina cómo se relacionan una o más variables (factores causales o variables independientes) con otra que será objeto de estudio (variable dependiente) y que a partir de dicho modelo se pueda realizar inferencias que se apliquen a una población de estudio.

REGRESIÓN LINEAL

Uno de los objetivos principales de la ciencia consiste en descubrir las relaciones entre variables, y la estadística ha desarrollado instrumentos apropiados para esta tarea. Así por ejemplo, en el campo de la psicología podemos preguntarnos si el rendimiento laboral en un tipo de puesto de trabajo guarda relación con la personalidad del trabajador, si el fracaso escolar es más probable en niños con determinadas circunstancias personales y familiares, si un determinado estilo de vida fomenta los estados depresivos... si los niveles de estrés antes de la intervención están relacionados con la inteligencia (Botella et al., 1997, p. 181).

REGRESIÓN LINEAL SIMPLE (DOS VARIABLES)

Para realizar predicciones podemos seguir dos caminos:

1º La intuición, basado en la experiencia y sentido común, pero parcial y ciega a algunos factores; además de no ser un método científico y

2º El conocimiento de situaciones específicas, basado en la realidad, observación, datos históricos o cualquier otro método objetivo.

La regresión y el análisis de correlación nos permiten determinar tanto la naturaleza como la fuerza de una relación entre dos variables; y si bien es cierto que la intuición es muy barata, el riesgo que existe al utilizarla es mucho más alto.

En el análisis de regresión hay que establecer algunos conceptos básicos:

Variables intervinientes

Las variables deben dividirse en dos, una que será el objeto de estudio (variable dependiente) y otra será la que influye en ella (variable independiente).

En Regresión Lineal Simple sólo se puede hablar de UNA variable independiente, no así en lo que respecta a la Regresión Multivariable que se estudiará más adelante, en ese caso usted podrá escoger tantas variables como considere necesario y es ahí cuando la comparación que he realizado con el sistema de evaluación previa de proyectos tomará más sentido.

La representación de cada una de las variables en el plano cartesiano se establece de la siguiente manera: Variable Independiente en el eje de abscisas ("x") y Variable Dependiente en el eje de ordenadas ("y").

La decisión de saber cuál será la variable dependiente (objeto de estudio) es la parte más importante de este proceso, para ello hay varias recomendaciones que nos pueden ayudar a decidir. En cuanto a ejercicios planteados, por ejemplo, hay que revisar bien y leer con atención la redacción de los mismos para saber si nos dan una “pista” que permita conocer qué se pretende estudiar; en algunos casos esta redacción no es muy clara y deberá primar la lógica. En referencia a temas de investigación el asunto obviamente es muy claro dado que la pregunta será: ¿qué se quiere investigar?, siendo la respuesta a este interrogante la variable dependiente.

TIPOS DE RELACIÓN ENTRE LAS VARIABLES

Las variables relacionadas presentarán lo que se conoce como tipos de relación, a saber: Directa e Inversa, estos se basan en la lógica, es decir debe haber una relación esperada (lógica) entre variables; Existen dos casos para establecer el tipo de relación:

1. Relación Directa. Las variables presentarán este tipo de relación cuando el incremento (disminución) de la una, determinará un incremento (disminución) de la otra; gráficamente según se muestra en la figura 156.

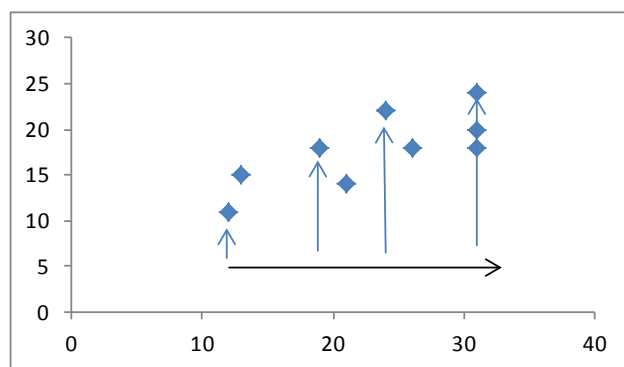


FIGURA 156: GRÁFICO DE UNA RELACIÓN DIRECTA ENTRE VARIABLES, A MEDIDA QUE “x” CRECE LA TENDENCIA DE “y” ES CRECER TAMBIÉN

Si nos fijamos en los puntos, según crecen los valores en el eje de abscisas (x), en el eje de ordenadas (y) en general también se observa un crecimiento; es importante resaltar esto de que no todos los valores de la variable dependiente tienen el mismo “nivel” o “rapidez” de crecimiento, esto dependerá de cada individuo estudiado.

2. Relación Inversa. Las variables tienen una relación inversa cuando el incremento (disminución) de la una, determina disminución (incremento) en la otra, según nos muestra la figura 157.

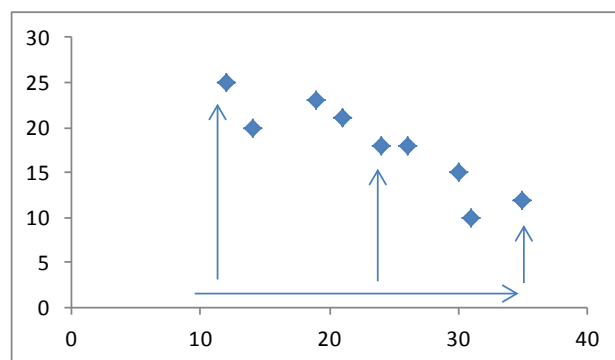


FIGURA 157: GRÁFICO DE UNA RELACIÓN INVERSA ENTRE VARIABLES, A MEDIDA QUE "x" CRECE LA TENDENCIA DE "y" ES A DECRECER

En la figura anterior se puede apreciar que, al aumentar los valores en la variable independiente, los de la variable dependiente tienden a disminuir.

En términos prácticos se sugiere que para determinar el tipo de relación entre las variables, siempre se haga la pregunta siguiente: ¿qué pasa con la variable dependiente si la independiente crece?, la respuesta a esta pregunta determinará el tipo de relación entre ellas (siempre hay que “hacer crecer” a la variable independiente); en términos prácticos, siempre haga crecer a la variable independiente y decida qué le pasaría a la dependiente, la respuesta será el tipo de relación esperada.

Lo primero que debe hacerse en un análisis de regresión simple, es realizar un gráfico de dispersión con los datos, éste dará una **idea** para determinar si existe o no relación entre las variables y de qué tipo es ésta; hay que recordar que ningún gráfico demuestra nada, es por ello que he resaltado el término porque la única manera (por lo pronto) de comprobar el tipo de relación entre dos variables es encontrar el modelo matemático (fórmula) que relacione a las variables intervinientes.

Propongo un ejemplo para ir desarrollando cada idea y proceso en el análisis de regresión simple.

Se recabó información de 15 estudiantes en las variables “comprensión verbal” y “expresión escrita”, se desea determinar si existe o no relación entre ellas; los datos en la tabla 66 indican los resultados obtenidos.

Comprensión oral	16	49	30	20	34	23	48	17	33	42	24	30	26	32	38
Expresión escrita	29	49	38	30	15	26	47	15	49	31	20	16	43	43	42

TABLA 66. VALORES ENCONTRADOS EN LA INVESTIGACIÓN A 15 ESTUDIANTES

Primer paso: establecer cuál es la variable dependiente, este paso es importantísimo dado que de la correcta elección de cuál es la variable a estudiar dependerá todo el análisis; para ello se sugiere preguntarse: ¿qué variable depende lógicamente de la otra? O según el caso ¿qué variable es de un nivel más complejo o abarca más complejidad? En este caso se

tomará como variable objetivo la expresión escrita (hay veces que puede haber discusión en la determinación de cuál es cuál, eso dependerá de varios criterios profesionales).

Hecho esto, debemos encontrar la recta de mejor ajuste para los datos, para ello explico paso a paso el proceso en Excel.

Los datos obtenidos deben colocarse en columnas consecutivas y **obligadamente** los correspondientes a la **variable independiente a la izquierda** como se muestra en la figura 158:

Comprensión oral	Expresión escrita
Variable Independiente	Variable Dependiente
16	29
49	49
30	38
20	30
34	32
23	26
48	47
17	15
33	49
42	31
24	20
30	29
26	43
32	43
38	42

FIGURA 158: DATOS DE LA INVESTIGACIÓN

Hecho esto hay que buscar el gráfico correspondiente a estos datos para encontrar lo que se llama la recta de mejor ajuste; en Excel debe buscarse el gráfico de “dispersión”, previo a esto debe resaltar los valores de las variables (no incluya los rótulos) según se indica en la figura 159.

Comprensión oral	Expresión escrita
Variable Independiente	Variable Dependiente
16	29
49	49
30	38
20	30
34	32
23	26
48	47
17	15
33	49
42	31
24	20
30	29
26	43
32	43
38	42

FIGURA 159: RESALTAR DATOS

Luego busque en el menú principal “insertar” y dentro de este haga “click” en el ícono que representa el gráfico de Dispersión según se presenta en la figura 160

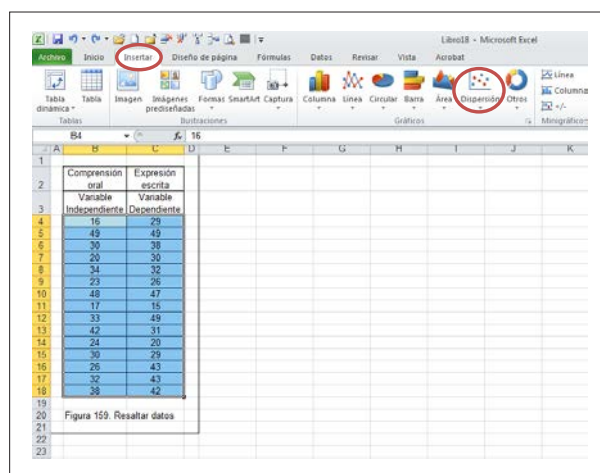


FIGURA 160: PROCESO A SEGUIR PARA OBTENER LOS DATOS DE LA RECTA DE REGRESIÓN Y OTROS

Luego escoja la primera opción dentro de las formas que ofrece el menú de gráficos de dispersión, como resultado de estas indicaciones se encontrará lo presentado en la figura 161.

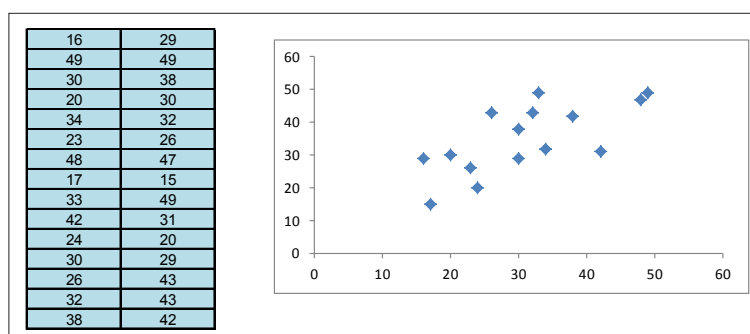


FIGURA 161: GRÁFICO DE DISPERSIÓN DE LOS DATOS

A los puntos del gráfico se les llama “nube de puntos”, no es otra cosa que la representación de los valores de las parejas en el plano Cartesiano.

En este momento debemos ya darnos una idea de lo que ocurre en la relación de las variables y a esa idea se le llama “tendencia”; esto significa que la nube de puntos nos avisa el tipo de relación que existe entre las variables; como se puede notar, esa tendencia indica que la relación entre las variables es directa; es decir que a medida que el nivel de comprensión verbal (variable independiente) aumenta, también lo hace la capacidad de expresión escrita (variable dependiente).

Preguntas:

1. ¿Es ese el tipo de relación esperada?
2. ¿Es lógica la relación?

Esto es algo que el investigador siempre debe cuestionar y responderse; en este caso la respuesta es que sí para ambas preguntas, lo cual significa que estamos por buen camino.

Lo siguiente será encontrar el modelo matemático (función) que nos indique numéricamente cómo se relacionan las variables, para esto, sobre el gráfico encontrado hay que posicionar al cursor sobre cualquiera de los puntos azules y hacer *click* con el botón de la derecha del ratón, aparecerá un cuadro de diálogo como el indicado en la figura 162

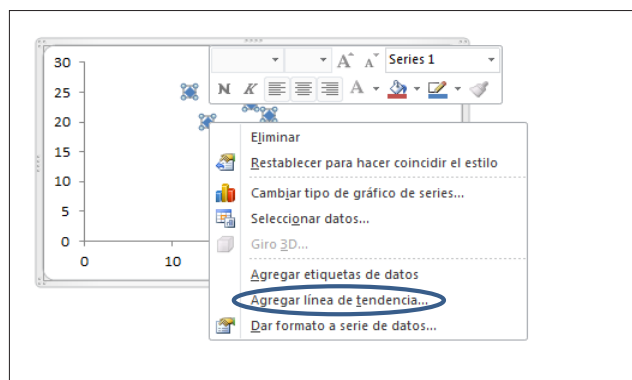


FIGURA 162: CUADRO DE DIÁLOGO INICIAL PARA ENCONTRAR EL MODELO MATEMÁTICO

Allí se debe escoger la opción “Agregar línea de tendencia...”, luego de esto aparecerá lo siguiente

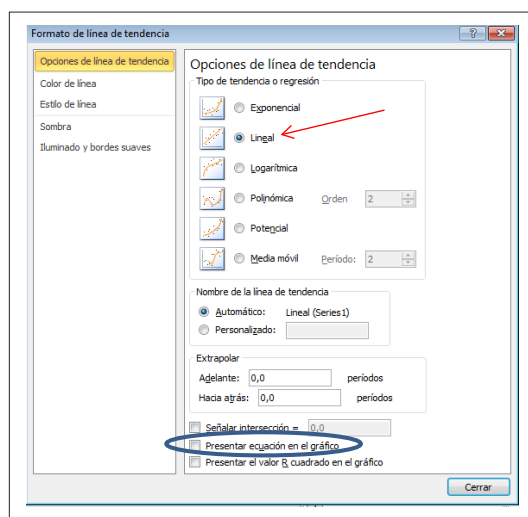


FIGURA 163: CUADRO DE DIÁLOGO PARA ESCOGER EL TIPO DE RELACIÓN Y PEDIR A EXCEL QUE ENCUENTRE LA FUNCIÓN LINEAL QUE MEJOR SE AJUSTA A ESOS PUNTOS

En este cuadro de diálogo se puede escoger el tipo de relación que al investigador le parezca el más adecuado, para el caso de regresión lineal obviamente se acepta el sugerido (por defecto) siempre por Excel, es decir la tendencia lineal; además se ofrecen otras opciones de las cuales vamos a señalar por lo pronto la que dice “Presentar ecuación en el gráfico” según se puede observar en la figura 163.

El resultado de esta acción se representa en la figura 164, en la cual ya se puede observar la ecuación de regresión entre las dos variables.

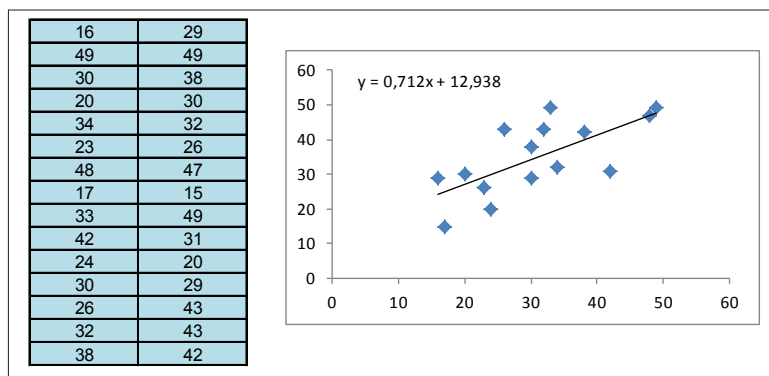


FIGURA 164: SE HA ENCONTRADO LA FUNCIÓN QUE RELACIONA A LAS DOS VARIABLES

La ecuación de una recta se representa de la siguiente forma: $y = b x + a$, en donde “b” es la pendiente de la recta y “a” es la intersección de la recta con el eje de ordenadas. Esa pendiente indica la inclinación de la recta (tema que no se abordará en este libro y que se supone se estudia en los establecimientos de nivel medio) pero es un valor muy importante que se lo debe tratar con mucho cuidado ya que provee mucha información en el proceso de la relación entre las variables.

Como cualquier valor numérico “b” puede ser positivo, negativo o cero, este último caso obviamente indicará que no existe ninguna relación entre las variables; en todo caso solo cuando se encuentra la ecuación de la recta se puede determinar el tipo de relación entre las variables (más adelante se indicará que hay otro valor numérico que confirma el tipo de relación).

Si el valor de “b” es **positivo** entonces el tipo de relación será **Directa**, es decir que a medida que crece la variable independiente, la dependiente también lo hará.

Si el valor de “b” es **negativo** entonces el tipo de relación será **Inversa**, es decir que a medida que crece la variable independiente, la dependiente irá decreciendo.

Fíjese que para cualquier caso, siempre se hace “crecer” a la variable independiente y luego de ello determinar la consecuencia en la dependiente.

Veamos unos ejemplos

Ejemplo 1

La Stork Foundation desea mostrar con estadísticas que, contrariamente a la creencia popular, las cigüeñas *sí traen bebés*. Por tanto, ha recabado datos sobre el número de cigüeñas y el número de bebés (ambos en miles) en varias ciudades grandes de Europa central (“Estadística y algo mas: Análisis de Regresión y Correlación,” n.d.).

Cigüeñas	27	38	13	24	6	19	15
Bebés	35	46	19	32	15	31	20

Determine la ecuación de mejor ajuste y ayude a los miembros de esta fundación a obtener una conclusión (primaria).

Se determina primero las variables dependiente e independiente, en este caso las cigüeñas harán las veces de variable independiente y los bebés representarán la variable dependiente; en este y todos los casos la recomendación es hacerse dos preguntas previas:

1. ¿Tienen relación las variables?
2. Si es así, ¿qué tipo de relación se espera? (directa o inversa).

Supongo que la respuesta a la primera pregunta es no; pero como este es un caso solo para ejemplificar algunas cosas, seguiré adelante; en cuanto a la segunda pregunta el tipo de relación a esperar le permitirá a usted determinar si la lógica de la relación es la que en verdad se da en entre las variables y, si esto coincide, deberá quedarse “tranquilo” porque al parecer el proceso seguido en la recolección de datos ha sido el correcto.

Lo realizado en Excel respecto a este ejercicio, se presenta en la figura 165

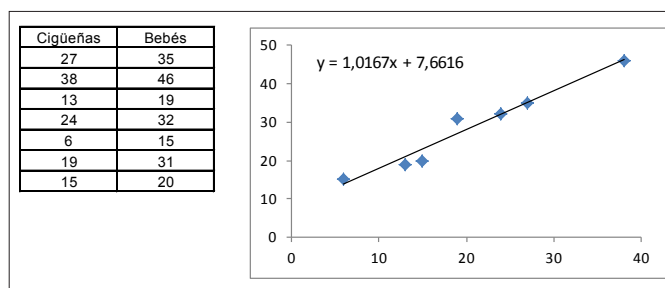


FIGURA 165: RELACIÓN ENTRE EL NÚMERO DE CIGÜEÑAS Y BEBÉS

Según se puede observar, la ecuación de regresión es: $y = 1.0167x + 7.6616$

En este caso se interpreta lo siguiente:

1. Según el gráfico, a media que el número de cigüeñas aumenta, también se incrementa el número de bebés.
2. El valor de “b” es mayor que cero (positivo) por tanto se afirma que la relación encontrada es directa, con este dato ya podemos confirmar que la relación es directa.

Como podrá observar, estadísticamente se puede encontrar relación entre variables aunque racionalmente eso sea algo absurdo. Sobre esto hay que tener mucho cuidado ya que no es cuestión de encontrar relación entre variables ya que numéricamente siempre podrá hacerlo; hay que determinar si lo que se pretende relacionar tiene alguna lógica; por ejemplo, usted puede relacionar la estatura con el rendimiento académico de los estudiantes de quinto de bachillerato y ¡va a encontrar una ecuación que relacione estas variables! pero supongo le parecerá absurdo hacerlo.

Ejemplo 2

Un visitador a médicos ha recabado los siguientes datos de sus clientes respecto al número de visitas y el número de veces que dichos médicos prescribieron sus productos

Número de visitas	1	6	4	1	3	5	7	4	2	5	6	4	3	2	7	3	6	3	5	8
Número de prescripciones	2	3	1	4	2	0	1	2	5	4	3	4	5	0	4	2	1	0	1	1

¿Tienen relación estas variables?, si la hay ¿qué tipo de relación espera?

Por lo pronto no discutiré la respuesta a la primera pregunta y supondremos que sí hay relación, respecto a la segunda digamos que la lógica (o la intuición) nos dice que mientras más visitas haga, los médicos prescribirán mayor cantidad de sus productos por tanto la relación esperada será directa según la premisa dada (recuerde que en estos ejemplos es discutible la posición adoptada).

Solución numérica

El diagrama de dispersión y recta estimada se presentan en la figura 166

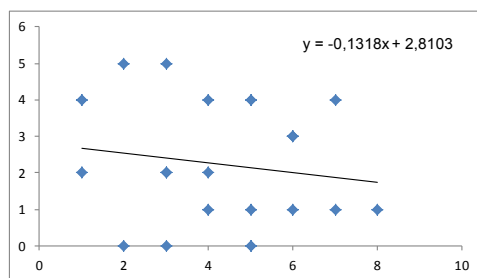


FIGURA 166: RELACIÓN ENTRE EL NÚMERO DE VISITAS Y PRESCRIPCIONES

Ecuación de la recta: $y = -0.1318 X + 2.8103$

Según el gráfico encontrado, lo que se puede observar en los puntos azules no da muchas luces sobre el tipo de relación o al menos no es muy claro o determinante, pero al encontrar la ecuación se puede ver que dado el signo de “b”, la relación dada es inversa, eso también se puede ver en la recta (al contrario de lo planteado como intuitivo), por tanto se establece que a medida que las visitas se han incrementado los médicos cada vez han prescrito menos el producto.

Le parezca o no lógica la relación, deberá aceptarse el resultado y en la práctica preguntarse qué ocurrió ¿Qué le recomendaría usted al visitador?

FUERZA DE RELACIÓN ENTRE LAS VARIABLES

Establecer el tipo de relación es importante para saber (entre otras cosas) si los resultados obtenidos son lógicos, pero hay también otros aspectos que analizar y uno de ellos es lo

relativo a la fuerza de relación o dependencia entre las variables, por tanto, debemos preguntarnos lo siguiente: dado que hay relación entre las variables, ¿cuán dependiente es la una de la otra?, ¿cuán fuerte es la relación que les une?

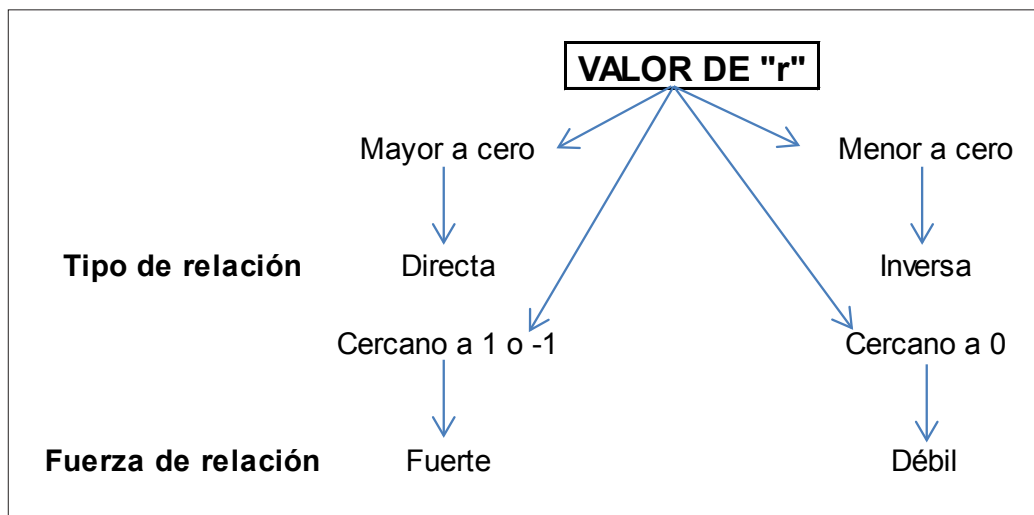
Para contestar estas interrogantes debemos estudiar y calcular lo que se conoce como Coeficientes de Correlación y Determinación; ambos miden a su manera la dependencia entre las variables.

COEFICIENTE DE CORRELACIÓN

¿Qué es entonces el Coeficiente de Correlación? Este valor indica el grado en que la variable independiente afecta o influye en la variable dependiente, el Coeficiente de Correlación “r” toma valores en el intervalo $[-1 ; 1]$ y por tanto puede ser negativo o positivo indicando también de esta manera el tipo de relación entre las variables; es decir, si “r” es menor a cero la relación será inversa y si el mayor será directa. En términos numéricos cuanto más cercano sea el valor de “r” a los extremos, significará que la variable independiente afecta de con mayor fuerza al cambio en la variable dependiente y por el contrario si se acerca mucho a cero, se deberá interpretar que existe poca o ninguna relación entre las variables estudiadas.

Se debe aclarar que no existe la relación causa efecto, la correlación entre variables no significa causalidad; en términos numéricos no se encontrará un valor de “r” igual a uno en la relación entre dos variables y si eso ocurre alguna vez, ¡dude de esa relación!

La siguiente ilustración establece algunas consideraciones a tomar en cuenta:



FUENTE: AUTOR

En cuanto al cálculo de este coeficiente en Excel se lo puede hacer de varias formas, entre ellas, se debe buscar la función “*coef.de.correl.*” según se muestra en la figura 167, en el ejemplo último el proceso será el siguiente:

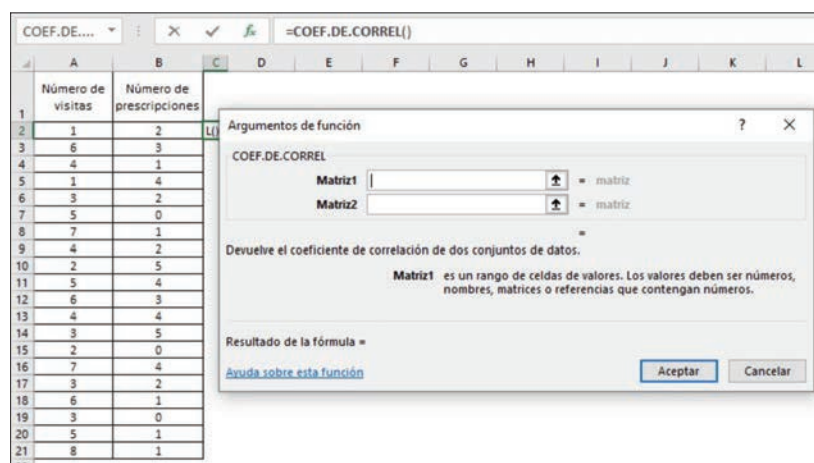


FIGURA 167: CUADRO DE DIÁLOGO PARA INICIAR EL CÁLCULO DEL COEFICIENTE DE CORRELACIÓN "r"

En el recuadro de *Matriz1* debe colocarse el rango de celdas en el cual se encuentran los valores de la variable independiente y en el recuadro correspondiente a *Matriz2* estará el rango de celdas de la variable dependiente según se muestra en la figura 168.

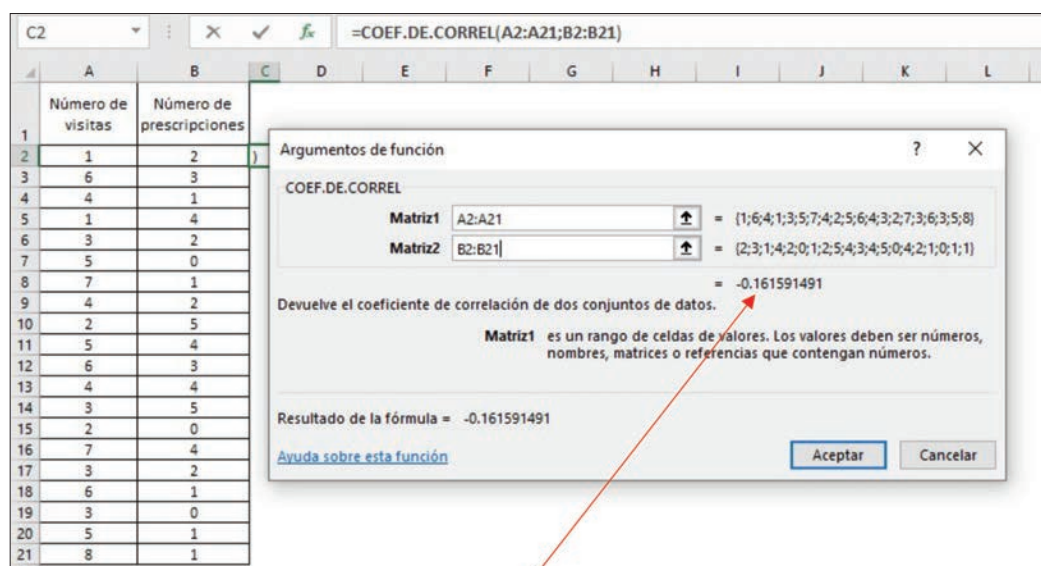


FIGURA 168: CÁLCULO DEL COEFICIENTE DE CORRELACIÓN

Como se puede apreciar al realizar esto Excel presenta el valor de "r" y bastará hacer "click" para que aparezca ese valor en la celda establecida.

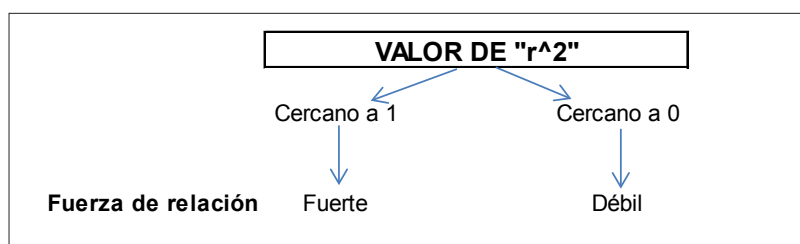
Note que el signo de "r" es negativo al igual que lo sucedido con el signo de "b" de la ecuación, esto siempre debe coincidir.

Entonces en el ejemplo estudiado podremos decir que existe una correlación de -0.1615 entre el número de visitas y la cantidad de prescripciones y como es claro, este valor es muy bajo (se acerca mucho a cero), lo que indicará que las prescripciones del producto no tienen mucha relación con la cantidad de visitas o no dependen mayormente de ellas.

COEFICIENTE DE DETERMINACIÓN

En cuanto al coeficiente de determinación “ r^2 ”, éste tendrá valores entre 0 y 1 y se dice que la variable independiente “explica” en determinado porcentaje la variación de la variable dependiente; y obviamente también determina la fuerza de relación entre las variables estudiadas. Numéricamente puede variar en el intervalo $[0, 1]$ indicando qué tanto una variable describe a la otra, suele escribirse en porcentaje $[0\%, 100\%]$.

La siguiente ilustración establece las características de interpretación general



FUENTE: AUTOR

Por tanto, se puede decir que:

si $r^2 = 0$ entonces NO EXISTE relación entre las variables.

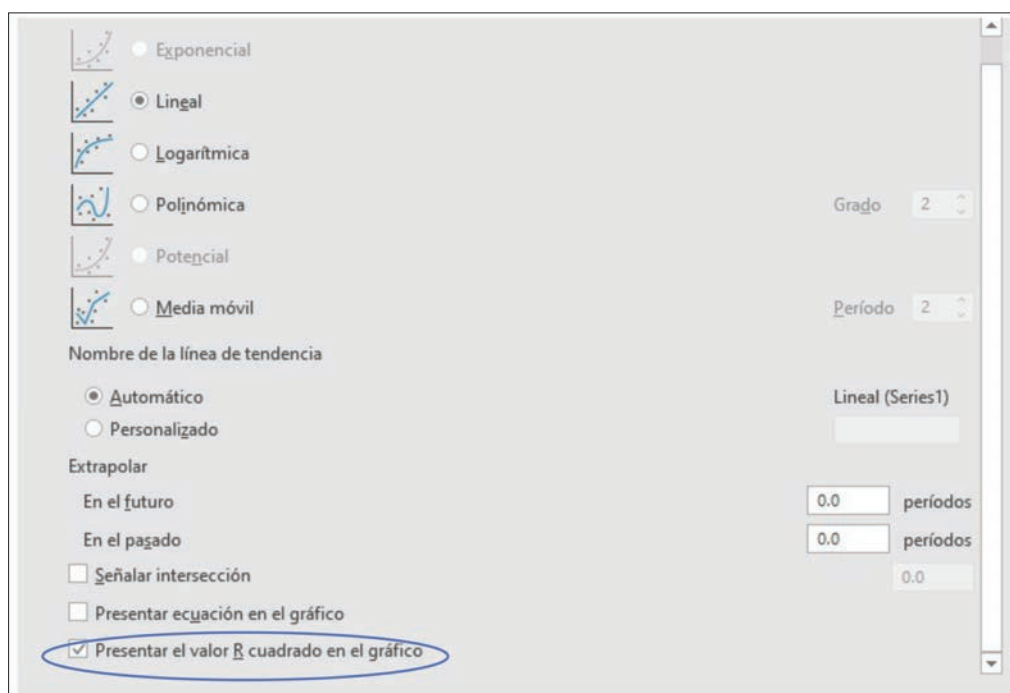
si $r^2 = 1$ entonces la relación entre las variables es “perfecta” (en realidad como dijera en líneas anteriores, esto no ocurrirá)

Como es obvio para conocer el valor del coeficiente de determinación, bastará elevar al cuadrado el valor de “ r ”; para el caso de nuestro ejemplo, $r^2 = 0.02611$ que transformado a porcentaje será: $r^2 = 26.11\%$; esto se interpreta de la siguiente manera: el número de visitas influye en un 26.11% en la prescripción del producto, en otras palabras de cada 100 prescripciones del producto, 26 de ellas se dieron por la influencia de la visita del vendedor.

En cuanto a la recomendación a dar y parafraseando un dicho popular: “no por mucho visitar se prescribirá más”.

También se puede encontrar el valor del coeficiente de determinación, realizando el proceso de manera gráfica al igual que se hizo con los ejemplos de la cigüeña y este mismo de la prescripción del producto.

Para ello se deben seguir los mismos pasos para encontrar el gráfico de dispersión y la ecuación, solo faltaría pedir, en el último paso, que Excel encuentre además de la ecuación, también “el valor de R cuadrado en el gráfico” como se muestra en la figura 169.

FIGURA 169: CUADRO DE DIÁLOGO PARA QUE EXCEL MUESTRE EL VALOR DE r^2

Con los ejemplos establecidos hasta esta parte, se han podido determinar algunas cosas muy importantes que el investigador debe preguntarse en estos procesos: ¿las variables a estudiar realmente se relacionan?, ¿qué tipo de relación espera?, ¿el tipo de relación encontrado es lógico?, ¿el valor de los coeficientes tanto de correlación como de determinación indican una importante relación entre las variables estudiadas?

Para el ejemplo del visitador a médicos las cosas parecen no funcionar bien, dado que las respuestas a cada pregunta serían (en su orden): ¡no!, dado que el valor de “ r ” es muy bajo; esperaba una relación directa y en la práctica la relación fue inversa, es decir mientras más visitas hace, menos prescriben los médicos; los valores de “ r ” y “ r^2 ” indican una relación muy baja entre las variables a estudiar; en cuanto al tipo de relación encontrado esto tal vez le diga que debe cambiar de estrategia de venta.

Estos dos ejemplos han servido como instrucción para desarrollar el proceso gráfico-numérico a realizar en Excel, pero falta mucho para terminar el análisis que debe hacerse entre dos variables.

Luego de conocer el tipo de relación entre las variables a estudiar y la fuerza que las determina, es importante conocer otros valores que permitirán establecer cuán fidedigna ha resultado la muestra estudiada comparada con la realidad de la población; debemos recordar que todo el proceso que se realizará es en base a datos muestrales y el interés es conocer si estos resultados son fiables o verdaderos representantes de la población estudiada.

Además de conocer el tipo y fuerza de relación entre las variables se debe determinar si en el proceso se han cometido errores que sean estadísticamente aceptables, también

establecer cuán confiable es el estudio que se está haciendo es decir si los cálculos y resultados encontrados en la muestra se pueden trasladar a la población.

Para explicar cada paso a seguir en adelante, revisaremos un ejemplo en el campo de las ciencias de la Educación.

Un investigador desea conocer si existirá o no relación entre la capacidad de memorizar algo [10; 50] y los niveles de aprendizaje [5; 50], para ello ha recabado datos en una población de estudiantes secundarios; los valores para ambas variables se determinan en la siguiente tabla:

MEMORIA	NIVELES DE APRENDIZAJE	MEMORIA	NIVELES DE APRENDIZAJE	MEMORIA	NIVELES DE APRENDIZAJE	MEMORIA	NIVELES DE APRENDIZAJE	MEMORIA	NIVELES DE APRENDIZAJE
11	6	19	19	44	37	11	10	19	10
36	31	21	20	50	48	30	34	49	27
33	29	10	6	37	21	31	28	27	15
46	38	40	31	40	32	36	29	32	25
21	16	37	29	20	15	17	15	36	19
45	25	23	27	35	32	12	8	28	21
32	25	41	25	38	35	37	19	14	24
43	27	17	12	14	25	23	32	46	40
13	15	35	29	22	23	26	25	38	21
20	9	38	28	23	15	50	49	17	5

Antes de desarrollar el ejercicio, es conveniente preguntarse algunas cosas: ¿tienen relación las variables?, ¿cuál es la variable dependiente?, ¿qué tipo de relación se espera entre las variables? Las respuestas deberían ser: sí tienen relación, la variable dependiente es “niveles de aprendizaje” esto significa que una de las causas que afectan al aprendizaje es la memoria, se espera una relación directa ya que la lógica dirá que a mayor capacidad de memoria el aprendizaje aumentará.

Estas respuestas no tienen que “gustar” a todos, algunos pueden responder por ejemplo que no hay relación, eso significaría que sería absurdo realizar un análisis que trate de medir la relación entre las variables; y es así, si usted considera que las variables planteadas no se relacionan, ¡no pierda el tiempo!, no haga ningún proceso que trate de establecer algo que para su lógica no tiene sentido.

Para efectos didácticos de este libro, voy a suponer que las variables planteadas sí se relacionan y por tanto que los niveles de memoria sí se ven influenciados por la capacidad memorística de los sujetos estudiados.

Según he indicado en los ejemplos anteriores hay que realizar un gráfico de dispersión mismo que se presenta en la figura 170.

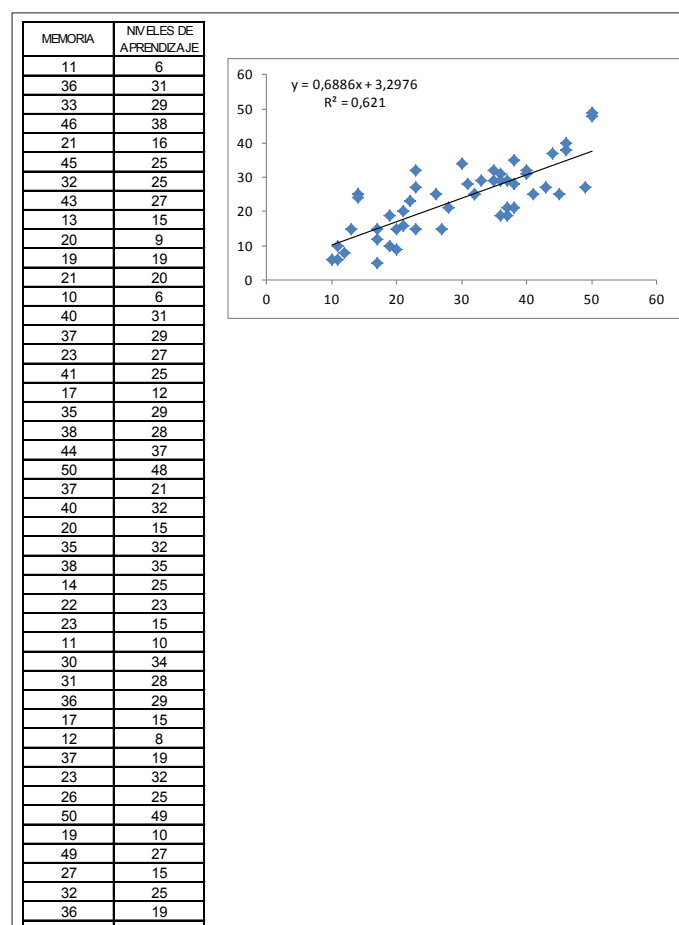


FIGURA 170: GRÁFICO DE DISPERSIÓN DE LA RELACIÓN ENTRE MEMORIA Y NIVELES DE APRENDIZAJE

Recuerde que la variable independiente debe colocarse a la izquierda, se ha realizado el proceso para que Excel grafique y encuentre el valor de la ecuación de relación entre las variables ($y = 0.6886x + 3.2976$) y también el valor del coeficiente de determinación “ r^2 ” (0.621).

Según estos primeros resultados, se confirma el tipo de relación esperada dado que el valor del coeficiente “ b ” de la ecuación es mayor a cero y que sí existe relación entre estas variables (en la muestra) dado que el valor de “ r^2 ” es alto: $r^2 = 62.1\%$; por tanto, el valor del coeficiente de correlación se puede obtener sacando la raíz cuadrada, dando como resultado que $r = 0.788$ (valor bastante cercano a 1).

Debo hacer una aclaración y alcance a la interpretación de estos dos coeficientes, se puede discrepar en que un porcentaje de 62.1% no es muy elevado, dado que esto indica que de cien casos, en “solo” 62 de ellos se establece una dependencia entre la capacidad de memoria y el nivel de aprendizaje; ante esto hay que recordar que los estudios en temas de Psicología y Educación (y en muchas ciencias) no deben hacerse con una sola variable dado que no existe la relación causa – efecto y por tanto los análisis de relación entre dos variables solo arrojan resultados parciales que si bien es cierto darán luces sobre la variable estudiada, no será definitivo y seguramente habrá que complementar con otros análisis.

Ante esto suele preguntarse: ¿qué valor de “ r ” o “ r^2 ” permite decir que la relación es fuerte?, considero que no puede darse una respuesta fija a la pregunta, esto dependerá de las variables, de las circunstancias en las que se está haciendo el estudio (es decir características de la población) entre otras razones.

ERRORES DE ESTIMACIÓN (SE) Y DE INCLINACIÓN (SB)

Luego de conocer los primeros resultados el investigador podrá imaginarse que, si esto ocurre en la muestra, en la población deberá ocurrir algo parecido, pero para comprobar esto hay que realizar algunos cálculos, entre ellos, se debe determinar el nivel de error que puede darse al analizar solo una muestra y no la población.

Estos errores son de dos tipos: error de estimación (o error de posición) y error del coeficiente “ b ” (o error de inclinación); en términos gráficos se explican estos errores de la siguiente manera: el gráfico 171 establece que si bien la línea de regresión de la muestra (en negro) indica la tendencia de los datos, las líneas en rojo sugieren que la línea de tendencia de la población puede estar más “arriba” o “abajo” pero que la tendencia se mantiene; esto se refiere al error de estimación o de posición.

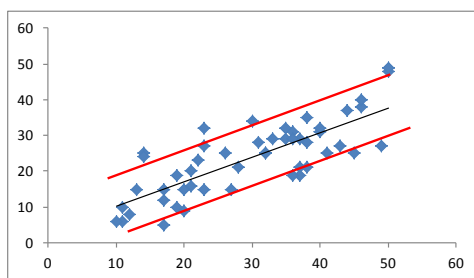


FIGURA 171: LAS LÍNEAS EN ROJO INDICAN EL INTERVALO EN EL QUE PUEDE ESTAR LA LÍNEA DE TENDENCIA DE LA POBLACIÓN

En cuanto al error del coeficiente “ b ” o error de inclinación el gráfico 172 explica que la inclinación de la recta de la muestra (línea en negro) indica la tendencia, pero quizá la tendencia de la población varíe (mucho o poco hacia arriba o abajo) (líneas en rojo).

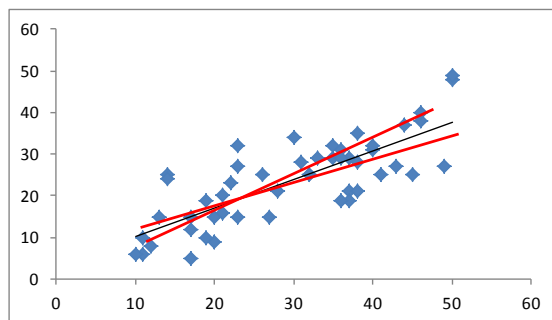


FIGURA 172: LAS LÍNEAS EN ROJO INDICAN LA POSIBLE INCLINACIÓN QUE PUEDE TENER LA LÍNEA DE TENDENCIA DE LA POBLACIÓN

Este error de inclinación es el más importante ya que si la dirección de la recta de la población varía mucho, puede ocurrir que el tipo de relación que arrojó la muestra no sea el mismo de la población y que en realidad no está reflejando lo que en verdad ocurre. En la figura 173 se detalla una situación extrema de este error.

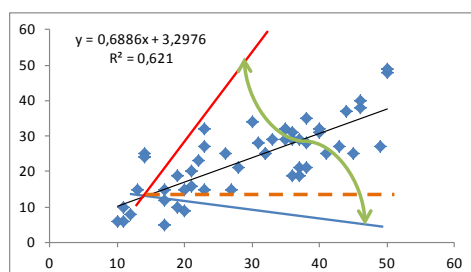


FIGURA 173: LAS LÍNEAS EN COLOR INDICAN DIRECCIONES EXTREMAS DE LA POBLACIÓN PARA EL CASO DE LA RECTA CORTADA LA DIRECCIÓN INDICARÍA QUE $b=0$

Si este es el caso, según los resultados de la muestra la relación entre las variables es directa, pero si el error es muy grande, en la población se puede dar que la relación tenga un rango entre directa e inversa y “en el camino” obviamente el valor de la pendiente “b” será cero (línea punteada), esto significaría que en verdad no exista relación en la población.

En este caso el valor de “b” de la población puede ser positivo, negativo o lo peor que puede ocurrir es que sea cero (línea en tomate) (más adelante explicaré con más detalle esta situación).

Hay que calcular entonces los valores numéricos de ambos errores para saber si los datos de la muestra son confiables.

CÁLCULO DEL ERROR DE ESTIMACIÓN (SE)

En Excel el cálculo del error estándar de estimación se hace de la siguiente manera:

En cualquier celda vacía, busque en fx la función: Error.Típico.XY, obtendrá un cuadro de diálogo como el de la figura 174.

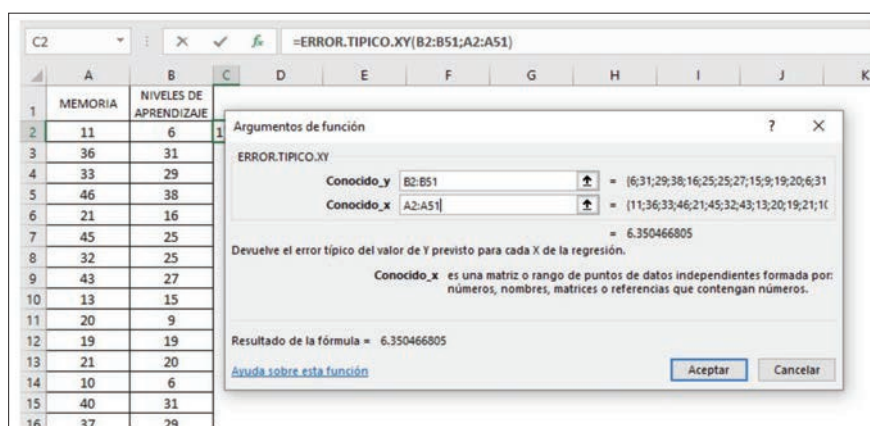


FIGURA 174: CUADRO DE DIÁLOGO PARA OBTENER EL VALOR DEL ERROR ESTÁNDAR DE ESTIMACIÓN

Para encontrar el valor del error de inclinación la manera más simple en Excel es encontrando una fórmula matricial, en los siguientes pasos se explica con detalle el proceso.

- | | A | B | C | D | E | F | G |
|----|---|---------|-------------|---|---|---|---|
| 1 | | | | | | | |
| 2 | | | NIVELES DE | | | | |
| 3 | | MEMORIA | APRENDIZAJE | | | | |
| 4 | | 11 | 6 | | | | |
| 5 | | 36 | 31 | | | | |
| 6 | | 33 | 29 | | | | |
| 7 | | 46 | 38 | | | | |
| 8 | | 21 | 16 | | | | |
| 9 | | 45 | 25 | | | | |
| 10 | | 32 | 25 | | | | |
| 11 | | 43 | 27 | | | | |
| 12 | | 13 | 15 | | | | |
| 13 | | 20 | 9 | | | | |
| 14 | | 19 | 19 | | | | |
| 15 | | 21 | 20 | | | | |
| 16 | | . | . | | | | |
| 17 | | . | . | | | | |
| 18 | | . | . | | | | |
| 19 | | | | | | | |

- Haciendo “click” en f_x busque la función “ESTIMACIÓN.LINEAL”, al aceptar se desplegará un cuadro de diálogo como se muestra en la figura 176



3. En “Conocido_y” debe resaltar los valores de la variable dependiente (y), en “Conocido_x” debe resaltar los valores de la variable independiente (x), en los casilleros “Constante” y “Estadística” digite 1 (uno); ¡NO HAGA clic EN ACEPTAR!, en este momento debe tener lo que se muestra en la figura 177

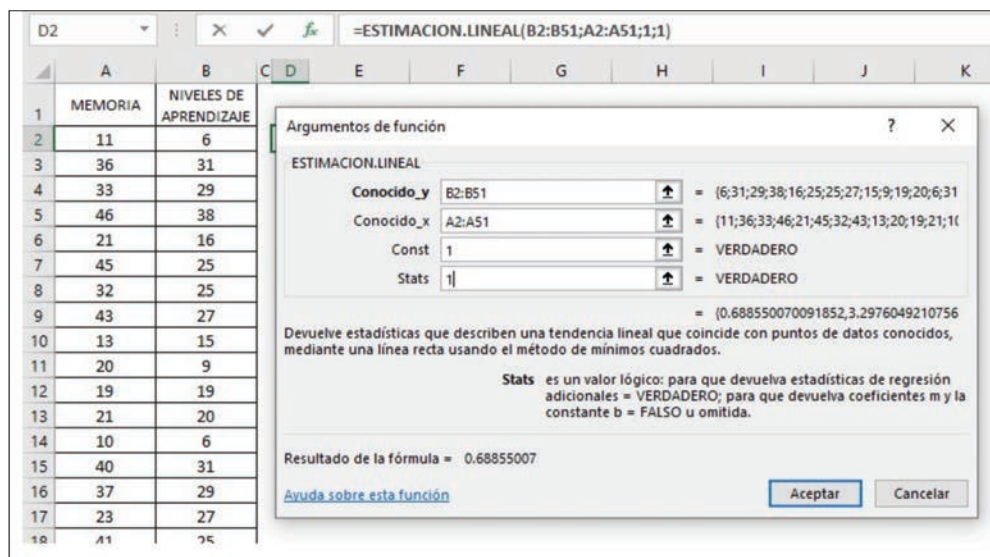


FIGURA 177: PANTALLA PREVIA A ENCONTRAR LA FÓRMULA MATRICIAL

4. En el teclado busque tres teclas que le ayudarán a completar el proceso:
“Control”



“Shift” (esta tecla se encuentra justo encima de Control)



“Enter”



5. Aplaste de manera independiente cada tecla (sin soltar) y cuando las tres estén ya aplastadas suéltelas despacio y al mismo tiempo (mi sugerencia es que este proceso lo haga con cierto “ritmo”).



6. Deberá obtener una matriz como la que se muestra en la figura 178

MEMORIA	NIVELES DE APRENDIZAJE		
11	6		Valor de "b"
36	31		Valor de "a"
33	29	0.68855007	3.29760492
46	38	Valor de "Sb"	0.07763418
21	16		2.471573
45	25		
32	25		
43	27		
13	15		
20	9		
19	19		
21	20		

FIGURA 178: FÓRMULA MATRICIAL

Cabe anotar que algunas veces al momento de soltar las tres teclas no aparece la matriz, esto es un problema común al inicio, si no lo logró, vuelva a intentarlo.

¿Qué información devuelve esta matriz?, la primera fila representa los coeficientes de la ecuación tomando en cuenta la ecuación general que es: $y = bx + a$.

Si nos fijamos en la figura 170, podremos ver que son los mismos valores (el valor de "b" en ese caso está redondeado), por lo tanto, se puede volver a escribir la ecuación con estos valores: $y = 0.688x + 3.297$.

La segunda fila arroja, de izquierda a derecha, el valor del error de inclinación, en este caso Sb sería 0.0776, el otro valor indicaría el error correspondiente al valor de "a" de la ecuación, a esto nos referiremos más adelante.

Por lo pronto conocemos entonces:

Gráfico que confirma el tipo de relación esperada: directa

Ecuación: $y = 0.688x + 3.297$

Valor del error de Estimación: $Se = 6.3504$

Valor del error de Inclinación (error del coeficiente "b"): $Sb = 0.0776$, este valor indicaría que si tomamos el valor de "b" (pendiente de la ecuación), la inclinación variará en 0.0776 hacia arriba y hacia abajo, no parece ser significativo este valor, pero esto lo comprobaremos al proseguir en el análisis.

ELABORACIÓN DE HIPÓTESIS

El proceso que se ha venido cumpliendo paso a paso es para determinar si en base a los datos de la muestra obtenida, se puede concluir que en la población ocurre lo mismo. Entre

los principales elementos para determinar esto es saber si la ecuación de la forma: $y = bx + a$, obtenida en base a los datos registrados, se puede transformar en la ecuación de la población a otra cuyo formato es el mismo pero escrito de esta manera: $Y = Bx + A$. Como parte de todo el proceso investigativo en Estadística, deben plantearse entonces hipótesis a comprobar luego de realizar los cálculos y análisis numéricos que correspondan al objeto de estudio. Para el tema que estamos tratando, estas hipótesis (Inicial y Alternativa) tratan de comprobar en principio que el valor del coeficiente “B” de la población cumple la condición básica de que en ningún momento tome el valor de cero, ya que esto anularía la posible relación entre las variables en dicha población.

El planteamiento de dichas hipótesis se resume de la siguiente manera:

Hipótesis Inicial (H_0): $B = 0$

Hipótesis Alternativa (H_1): $B \neq 0$

Una de estas hipótesis debe ser rechazada (o aceptada), si H_0 es rechazada esto significará que sí existe relación, en la población estudiada, entre las variables, caso contrario significará que, aunque en la muestra todo parecería ir bien, en realidad esto no ocurre en la población.

En estos procesos pueden ocurrir dos casos en general que harán fallar en el análisis y suele ser que la razón de la falla será una muestra que no fue bien tomada, estos son:

1. En la población sí existe relación, pero se acepta la H_0
2. En la población no existe relación, pero se rechaza la H_0

He indicado en diferentes cursos, que el hecho de suponer en primer plano (hipótesis inicial) que el valor de “B” sea cero, es una posición “pesimista” del proceso estadístico, ya que “arranca” con sospechar que entre las variables estudiadas en la población no existirá relación entre ellas, es por ello que el análisis debe ser muy exigente y riguroso para que no haya equívocos al momento de concluir sobre la investigación planteada.

Para comprobar las hipótesis, deben hacerse dos análisis numéricos basados en sendos cálculos:

1. Calcular el verdadero valor del coeficiente “B” en la población y
2. Comparar resultados entre los valores estandarizados establecidos en la distribución Normal y los valores específicos de cada investigación.

Para ello paso a explicar a continuación estos dos temas.

VALOR DE “B” Y NIVELES DE CONFIABILIDAD

Todos los cálculos anteriores hacen referencia exclusivamente a la muestra, hay que recordar que el interés de la Estadística en general y de la Inferencial en particular, es conocer el comportamiento de las variables en la población; para saber esto se debe poner una condición

a los análisis de regresión y esto no es otra cosa que determinar el nivel de confiabilidad con el que se va a realizar el proceso.

Los niveles de confianza deben ser bastante altos para garantizar que los resultados son fiables, esto exige entonces que deben hacerse cálculos con valores por lo menos a partir del 90% de confiabilidad, aunque en la práctica lo más utilizado es hacerlo con un nivel de confianza del 95%, de todas maneras, esto lo decide el investigador.

Dado que para nuestro ejemplo la ecuación de la muestra es $y = 0.688x + 3.297$, debemos suponer que la ecuación de la población no será muy distinta especialmente en cuanto al valor de “b”, más aún conociendo ya que el valor del error es muy pequeño.

Como la ecuación de la población tiene la misma estructura que la de la muestra, es decir habrá un coeficiente “B” de la variable independiente “x” y un término independiente “A” que se espera sean semejantes en valor a los coeficientes encontrados en la muestra, es por ello que la nomenclatura se expresa en mayúsculas para diferenciarlas.

De todas maneras, sabemos que existe un error y debemos calcular el valor de “B”, este valor estará determinado por tres elementos que son los siguientes: el valor original de “b”, el valor del error de “b” y el nivel de confiabilidad con el que se haga el análisis.

Para ello se determina la siguiente fórmula que no es otra cosa que establecer lo que se conoce como intervalo de confianza para el valor de “B” y utilizando los elementos antes mencionados, la fórmula para el cálculo será la siguiente:

$$B = b \pm Zt * Sb$$

Al término $Zt * Sb$ se le conoce como el error de estimación de “B” y el valor que tome dependerá del nivel de confianza con el que se realice el análisis.

Si el error de estimación sería cero significaría que el valor de “B” sería igual al de la muestra y por tanto el proceso sería perfecto; en la realidad esto no debe asumirse con tanta ligereza ya que en realidad no ocurrirá dado que la muestra, aunque obtenida con un sistema muy serio, no es totalmente fiel a los datos de la población.

El nivel de confiabilidad con el que se decida realizar la investigación se determina con los valores de la distribución normal “z”, pero esto dependerá del tamaño de la muestra (n) ya que si el valor de “n” es mayor a 30 (muestras grandes) se debe utilizar la distribución “t” (esto lo explicaré más adelante).

Según la tabla de distribución normal “z” (Anexo 1), el valor de “z” a utilizar en la fórmula dependerá del Nivel de Confianza (N.C.); como ejemplo vamos a determinar el valor de “z” en el caso de un N.C. del 95%, dado que se está utilizando una tabla no acumulada se debe dividir en dos el valor de 95% que en decimal sería 0.475; buscamos en la tabla el valor más cercano a esto y vemos que está al nivel de 1.9 (desde la izquierda) y de 0.06 (desde arriba), por tanto el valor de “z” correspondiente a un nivel de confianza del 95% será 1.96, como se muestra en la figura 179.

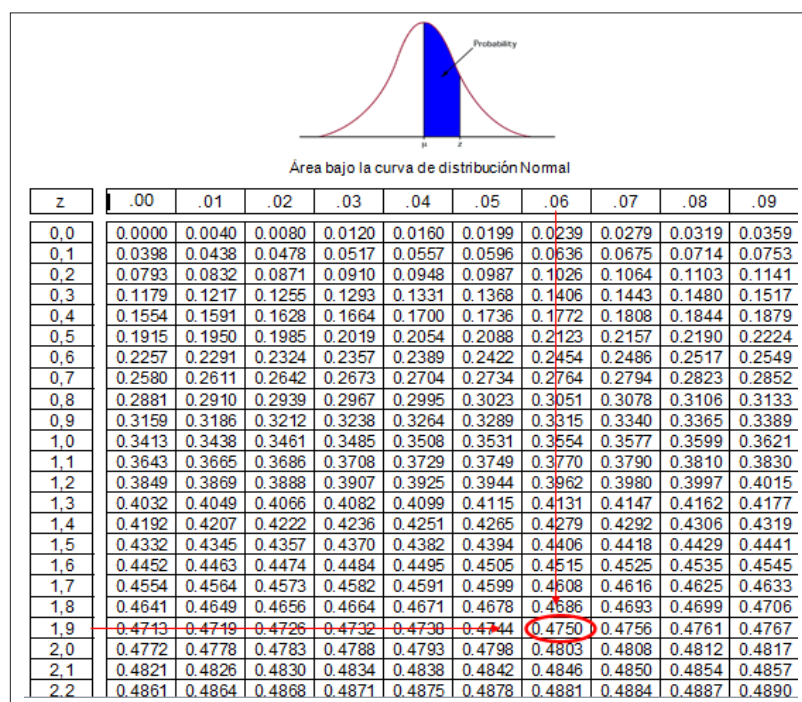


FIGURA 179: VALOR DE "z" CUANDO EL NIVEL DE CONFIANZA ES DEL 95%

Por lo tanto aplicando la fórmula propuesta, ya se puede obtener el valor del intervalo de confianza para el coeficiente "B" de la población, en este caso tendríamos lo siguiente:

$$B = 0.688 \pm 1.96 * 0.0776$$

Esto quiere decir que el error de estimación de "B" ($Z_t * s_b$), respecto al valor de "b" de la muestra es de 0.152 y realizando las operaciones, el intervalo para el valor de "B" estará dado por los valores siguientes: [0.535 ; 0.840], esto significa que con una confiabilidad el 95%, el valor del coeficiente "B" de la población no pasa por cero y por tanto se debe RECHAZAR H_0 .

Esto garantiza entonces, a un alto nivel de confianza, que la relación de estas variables tiene igual comportamiento tanto en la muestra como en la población y que los indicadores resultantes en el análisis de la muestra se mantienen también para la población.

Esto debe comprobarse aún más con otros estadígrafos que permitan asegurar lo que en cada paso se ha venido dando.

CONTRASTE ENTRE EL VALOR ESTÁNDAR Y EL ESPECÍFICO

Como se habrá notado en la fórmula para calcular el valor de "B", se utilizó el error de estimación con la expresión $Z_t * s_b$, en ella hay un subíndice en el valor de "Z" y que se representa por "t", el mismo que indica que debe usarse el valor de "Z" encontrado en la tabla, entonces surge la pregunta: ¿y qué otro valor de "Z" puede haber?

Pues bien, hay que calcular el valor de "Z" (Z calculado: " Z_c ") que corresponde a cada análisis en particular para cada caso ya que el valor de la tabla es un valor estandarizado, es

decir para una distribución con media igual a 0 (cero) y desviación estándar igual a 1 (uno); al comparar el valor estandarizado (tabla) y el específico (datos de la investigación) se establece la siguiente condición:

Si $Z_c > Z_t$ entonces debe RECHAZAR H_0 , esto significa que en la población **SÍ** existe relación entre las variables.

Para calcular el valor de Z_c , basta con encontrar el valor absoluto del valor de “b” dividido entre el valor de “Sb”

$$Z_c = |b|/Sb$$

Para el ejemplo que estamos analizando tendríamos

$$Z_c = \frac{0.688}{0.0776}$$

$$Z_c = 8.86$$

Como el análisis se hizo al 95% de confianza, el valor de $Z_t = 1.96$ y comparando con el valor de Z_c , se puede ver que $Z_c > Z_t$, es decir se debe RECHAZAR H_0 y por tanto se confirma que en la población **sí** existe relación entre las variables.

La conclusión dada cuando se calcula el intervalo del valor de “B” y cuando se comparan los valores de “ Z_c ” y “ Z_t ”, **DEBE** ser la misma, no puede haber contradicción en esto, es decir ambas deben o bien rechazar o bien aceptar H_0 .

Como parte del análisis general respecto a la ecuación obtenida en la muestra, falta interpretar el valor de “a” de la ecuación, en este caso $a = 3.2976$, ¿qué significa esto?

El valor de “a” en la ecuación de regresión es un valor “asociado” a la variable dependiente y se interpreta como un máximo o un mínimo esperado de dicha variable en la población según el tipo de relación obtenida.

Si la relación de las variables es directa, el valor de “a” deberá interpretarse como un mínimo esperado en la población y si la relación es inversa el valor de “a” deberá suponer un valor máximo de la variable dependiente.

Hay que recordar que la ecuación encontrada y el gráfico, son una proyección en base a los datos de la muestra, por tanto, en cada ocasión deberá hacerse un análisis específico para interpretar este valor.

Para el ejemplo que estamos analizando, el valor de 3.2976 estrictamente hablando debería decirse que es el mínimo valor de Niveles de aprendizaje que se espera en los estudiantes; como podremos suponer esta interpretación “a secas” no es muy lógica debido a que la variable dependiente se mide en el intervalo [5 ; 50] por lo tanto no es posible esperar que algún estudiante tenga valores menores a 5; en este caso entonces deberá tomarse como valor mínimo a esperar, el dato más bajo que se conoce de la variable dependiente en la muestra, en este caso sería 5.

La conclusión en este caso entonces sería que en la población objetivo de estudio (estudiantes secundarios), sí existe relación entre la Capacidad de memoria y los Niveles de aprendizaje y por tanto para analizar los niveles de aprendizaje de los estudiantes es válido relacionarlos con su respectiva capacidad de memorizar.

En calidad de resumen y como sugerencia para realizar los cálculos de los ejercicios, presento la figura 180 en donde se puede ver el proceso realizado de manera ordenada y sistemática; esta recomendación la hago porque si bien es cierto el proceso numérico termina aquí, faltan algunos detalles que servirán para el análisis final.

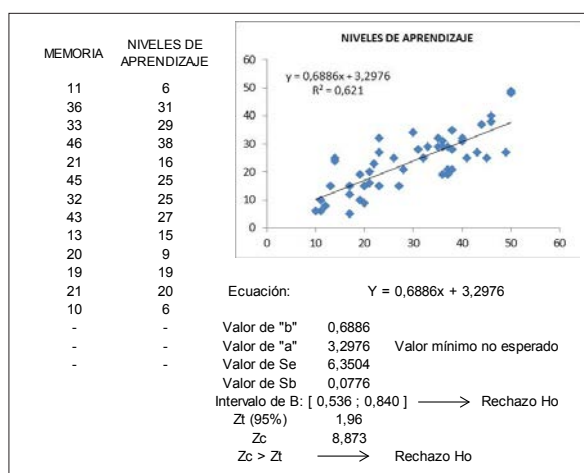


FIGURA 180: RESUMEN DEL PROCESO DE CÁLCULO DE LA RELACIÓN DE DOS VARIABLES

CÁLCULO DE VALORES DE ESTIMACIÓN

¿Para qué sirve y cómo se utiliza la ecuación de regresión?

La respuesta a esta pregunta es la siguiente: cuando se ha demostrado (hay que hacer énfasis en esto) que en la población sí existe relación entre las variables, la ecuación encontrada pasa a ser “predictiva” es decir nos permitirá inferir o predecir resultados de la variable dependiente, para la población de estudio, si conocemos valores específicos de la variable independiente; esto se hace remplazando dichos valores en la ecuación.

Para poder hacer esto existe una condición: los valores a remplazar deben estar dentro del intervalo en el que se mide la variable independiente, para este caso la variable Capacidad de memorizar se mide en el intervalo [10 ; 50], por lo tanto solo valores que estén dentro de este intervalo podrán ser remplazados.

Por ejemplo, si conocemos que un estudiante de la población tiene un nivel de memoria de 30 y dado que la ecuación sí es predictiva, al remplazar este valor (x) en la ecuación, podemos inferir que el estudiante tendrá un Nivel de Aprendizaje de 23.95.

¿Será esto exacto?

Lógicamente no, recuerde que esta ecuación indica una tendencia y que existen errores, por lo tanto será imprescindible realizar un cálculo adicional.

Al valor de 23.95 se le conoce como **estimación puntual**, es decir estamos casi “asegurando” que esto es verdad, para que la estimación sea más certera, debemos utilizar el valor del Error Estándar de Estimación (Se) y por eso su nombre. Para ello debemos restar y añadir este error al valor encontrado en la estimación puntual.

Entonces, si conocemos que un estudiante tiene una capacidad de memoria de 30, podemos decirle que con un nivel de confianza del 95% su Nivel de Aprendizaje estará en el intervalo [17.59 ; 30.30].

Y es en este momento que podemos interpretar si el valor de Se es grande o aceptable, ¿qué diría usted?

EJERCICIOS DE APLICACIÓN DEL CAPÍTULO

Ejemplo 1

1. Un alumno de la Facultad de Psicología ha recabado datos de algunos estudiantes y desea analizar si las variables estudiadas están o no relacionadas. Para ello tomó un test que mide Aptitud verbal y por otro lado ha obtenido datos sobre el Cociente Intelectual de dichos estudiantes. Realizar todo el proceso de Regresión.

Aptitud verbal	Cociente intelectual	Aptitud verbal	Cociente intelectual	Aptitud verbal	Cociente intelectual	Aptitud verbal	Cociente intelectual
125	135	109	111	123	135	114	115
127	140	105	110	125	135	98	97
105	104	109	116	103	107	114	120
102	114	109	113	120	130	114	117
114	118	100	115	112	121	100	115
120	125	100	105	113	110	105	115
104	106	105	110	111	115	110	112
110	115	112	120	100	109	99	105

Etapas de resolución sugeridas:

Primer paso: determine cuál será la variable dependiente

Segundo paso: pregúntese si en verdad para usted existe o no relación entre las variables planteadas

Tercer paso: plantee la pregunta: ¿qué tipo de relación espero entre estas variables?

Cuarto paso: elabore el gráfico de dispersión

Quinto paso, y siguientes: realice los cálculos de cada uno de los elementos para el completo desarrollo de regresión.

Variable Dependiente: Cociente intelectual

Variable Independiente: Aptitud verbal

Hipótesis inicial H_0 : No existe relación entre las variables planteadas ($B = 0$), recuerde que el planteamiento de la hipótesis es “pesimista”

Se esperaría una relación directa entre las variables, es decir mientras mejor sea la capacidad de aptitud verbal, se espera que el cociente intelectual también sea mayor (recuerde que se está estudiando el nivel de cociente intelectual en base a una variable).

El gráfico de dispersión, la ecuación y el valor de r^2 se presentan en la figura 181.

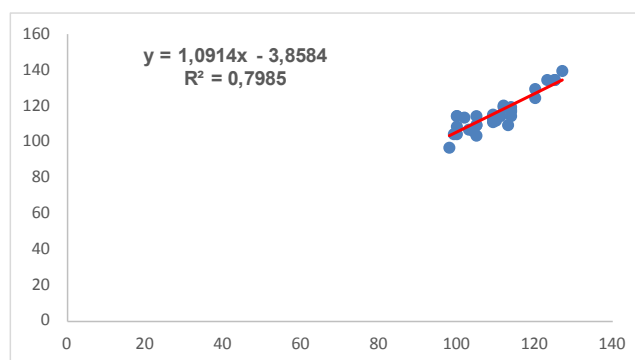


FIGURA 181: GRÁFICO, ECUACIÓN Y VALOR DE r^2 DE LA RELACIÓN ENTRE LAS VARIABLES

Se confirma entonces que existe una relación directa entre las variables, además, el valor de r^2 nos dice que la relación en la muestra es muy buena. Intuitivamente entonces podemos decir que en la población también se mantendrá esta relación y la fuerza de la misma.

Al realizar los cálculos se obtienen los siguientes resultados:

Valor de “b”: 1.0914

Valor de “a”: - 3.8584

Coefficiente de Correlación “r”: 0.8936

Coefficiente de Determinación “ r^2 ”: 0.7985

Error estándar de Estimación “Se”: 4.5945

Error de inclinación de la pendiente “Sb”: 0.1

Valor de “ Z_t ” (al 90%): 1.645

Valor de “ Z_c ”: 10.903

Intervalo de “B”: [0.8582 ; 1.3246]

¿Qué se puede decir de todo esto entonces?

Conclusión respecto a la hipótesis: con un nivel de confianza de 90% se rechaza la hipótesis inicial dado que según el intervalo encontrado el valor de “B” no pasa por cero, por tanto, sí existe relación entre las variables.

El estudiante que está investigando sobre esto deberá colegir que para tener una idea clara sobre el Cociente Intelectual (C.I.) de los estudiantes, sí es importante conocer la Aptitud verbal de ellos.

Dado que el valor de “B” no pasa por cero y $Z_c > Z_t$, entonces en la población se mantiene lo que intuitivamente se indicaba, sí existe relación entre estas variables.

Según el valor de “a” (- 3.8584) se puede decir categóricamente que este valor sería absurdo interpretarlo como mínimo esperado de un Cociente Intelectual, por tanto, deberemos tomar el valor más bajo conocido de la muestra, en este caso será: 97, por tanto, entre los estudiantes de esta población se espera como mínimo un valor de 97 en cociente intelectual.

La ecuación por tanto sí es predictiva y nos permitirá (según esto) determinar el Cociente Intelectual de un estudiante de esta población si es que conociéramos su nivel de Aptitud verbal.

Dado que la ecuación es predictiva, propongo los siguientes valores de Aptitud Verbal para inferir valores de Cociente Intelectual en esta población: $x_1 = 100$; $x_2 = 95$; $x_3 = 105$.

Para realizar el cálculo, estos valores deberían remplazarse en la ecuación de regresión y así estimar el valor del Cociente Intelectual para cada caso.

¡¡ALTO!!

Recuerde la condición para hacer estos cálculos, los valores a remplazar DEBEN estar en el intervalo de los valores conocidos de la variable independiente; en este caso la muestra indica que el menor valor de Aptitud Verbal conocido es 98 y el máximo es 102, es decir que “x” fluctúa en el intervalo: [98; 102].

Si nos solicitan realizar los cálculos sugeridos, debemos decir que para esta población sólo será posible hacer cálculos para el valor de 100 ya que los otros dos están fuera del intervalo.

Entonces, si $y = 1.0914x - 3.8584$,

$y = 1.0914 * 100 - 3.8584$

$y = 105.28$

Se puede decir que para una persona con un nivel de 100 en Aptitud verbal, su C.I. será de 105.28.

Pero esto es una estimación puntual y no es aconsejable quedarse con dicho valor ya que sabemos que siempre hay un error de estimación, por tanto hay que calcular el valor del intervalo de estimación restando y sumando el valor de Se.

$y = 105.28 \pm 4.5945$

$y_1 = 100.68$

$y_2 = 109.87$

La predicción será entonces que cualquier persona de la población estudiada con un nivel de Aptitud Verbal de 100, tendrá un Cociente Intelectual dentro del intervalo [100.68 ; 109.87].

El resumen gráfico de estos cálculos se presenta a continuación en la figura 182

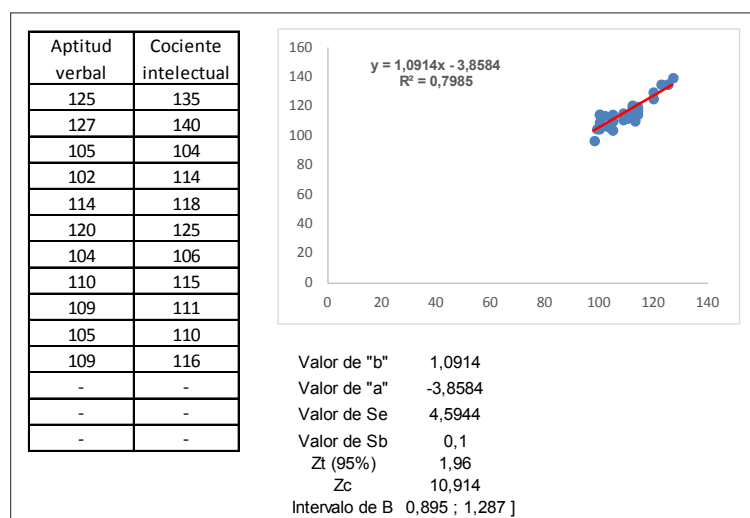


FIGURA 182: RESUMEN DE LOS CÁLCULOS DEL EJEMPLO

Hay que anotar que este es tan solo un ejercicio que sirve para ejemplificar el proceso a seguir en problemas de Regresión lineal, no estoy insinuando que se puede determinar el cociente intelectual de una persona en base a una ecuación de este tipo, aunque en temas estrictamente estadísticos esto sea lo que se pueda realizar.

Ejemplo 2

Los siguientes datos corresponden a una muestra de 28 personas que se han sometido a una investigación referente a la capacidad de retención. Identifique la variable dependiente y la independiente, determine si las variables tienen o no relación, el tipo de relación, diga cuán buena es la relación, determine si la ecuación es predictiva y si lo es encuentre valores para la variable dependiente conociendo que $x_1 = 15$ y $x_2 = 25$.

Capacidad de retención	Prueba de atención	Capacidad de retención	Prueba de atención
8	7	8	11
4	14	9	8
5	11	6	17
6	8	5	10
9	7	8	10
7	15	9	15
5	8	6	10
2	14	5	13
3	21	7	12
6	10	8	7
5	8	6	12
5	18	6	8
4	9	7	14
7	12	4	12

Variable dependiente: Capacidad de retención.

Variable independiente: Prueba de atención.

¿Tienen relación?: sí en teoría. Para responder numéricamente a esto, hay que encontrar los estadígrafos de regresión para comprobar en el proceso.

Tipo de relación esperada: directa.

Para contestar las siguientes preguntas se debe desarrollar el ejercicio.

Al realizar el gráfico llama la atención que el tipo de relación esperada es contraria al resultado encontrado ya que según la ecuación y el gráfico (figura 183) la relación es inversa (?).

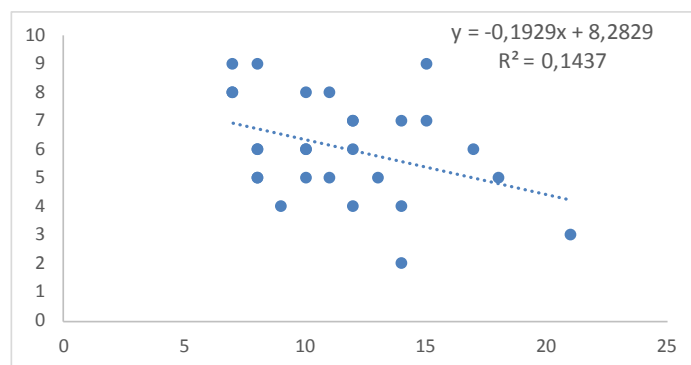


FIGURA 183: GRÁFICO, ECUACIÓN Y VALOR DE r^2 DEL EJEMPLO 2

En este momento el investigador debe cuestionarse sobre el proceso seguido ya que en la lógica no puede haber contradicción entre el tipo de relación esperada y lo hallado en la muestra (no es que no pueda ocurrir); pero para fines didácticos de este libro, seguiremos el proceso.

Este ejemplo tiene dos particularidades respecto a los otros que se han desarrollado previamente, estas son:

1. El tipo de relación es inversa y,
2. El número de elementos estudiados en la muestra es pequeño (28)

Respecto al primer punto solo se debe tener cuidado en los cálculos debido al signo negativo de “b”.

En cuanto al segundo punto es momento de introducir un concepto adicional en el proceso de Regresión Lineal, y es el que permite distinguir entre muestras grandes y pequeñas.

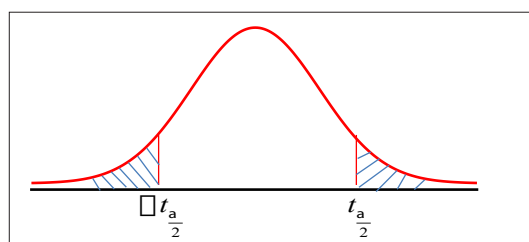
Además de la distribución Normal “z” ya conocida, también hay un tipo de distribución de características muy similares llamada Distribución “t” de Student (Anexo 2). Esta también es una distribución normal pero que se utiliza cuando el tamaño de la muestra es pequeño.

Entonces se establece una diferencia de uso entre “z” y “t” debido al número de elementos de análisis en cada caso, así:

Tamaño de muestra		Uso del tipo de distribución
Si	$n > 30$	→ "z"
	$n \leq 30$	→ "t"

La distribución "t" tiene algunas características matemáticas y su uso en Estadística es muy importante en varios temas, pero por el alcance de este libro, me limitaré a explicar la forma de encontrar valores "t" cuando la muestra es pequeña.

El gráfico de la distribución es el siguiente:



Y se hace notar que la región a "usar" se encuentra en las "colas" de la curva.

Paso a poner un ejemplo para determinar el uso de esta tabla. Supongamos que el tamaño de una muestra es 14 ($n = 14$) y se desea realizar el análisis con un nivel de confiabilidad del 95%, para encontrar el valor "t" correspondiente se deben seguir los siguientes pasos:

1. En la columna izquierda (ver figura 184), identifique "gl" (grados de libertad) y allí busquemos el valor que nos permitirá "entrar" por izquierda.

Pero ¿qué significa "grados de libertad"? Los grados de libertad se establecen por dos parámetros: tamaño de la muestra y número de variables en el análisis, como en regresión lineal estamos trabajando con dos variables, tenemos lo siguiente: $n = 14$ y número de variables $\# = 2$.

Para encontrar los grados de libertad utilizamos la expresión: $gl = n - \#$, entonces para el ejemplo $gl = 14 - 2$, por tanto, al igual que lo hicimos para encontrar un valor "z", debemos "entrar" desde la izquierda en la fila correspondiente a "gl" = 12.

2. Arriba en azul (figura 184), se encuentran los valores que nos permiten escoger el nivel de confiabilidad con el que se va a realizar el análisis, para la tabla de distribución "t" a usar tenemos: 0.2 , 0.1 , 0.05 , 0.02 , 0.01 y 0.001; estos valores corresponden, en su orden, a los siguientes niveles de confianza: 80% , 90% , 95% , 98% , 99% y 99.9%. Para el ejemplo se escogió realizar el análisis al 95%, es decir que en las colas hay un total de 5% distribuido simétricamente en porcentajes de 2.5%; por tanto, debemos "entrar" desde arriba por la columna correspondiente a 0.05.

El resultado de esta combinación nos da el valor de "t", en este caso $t = 2.179$, que será el valor a utilizar en la fórmula para encontrar "B".

gl	0,2	0,1	0,05	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,695	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015

FIGURA 184: DETERMINACIÓN DEL VALOR DE t_c SI $n=14$ Y AL 95% DE CONFIABILIDAD

Sigamos entonces con nuestro ejercicio.

Ecuación: $y = -0.1929x + 8.2829$

Hipótesis inicial H_0 : no existe relación entre las variables, $B = 0$

Al realizar los cálculos se obtienen los siguientes resultados:

Valor de “b”: -0.1929

Valor de “a”: 8.2829

Coefficiente de Correlación “r”: -0.379

Coefficiente de Determinación “ r^2 ”: 0.1437 (14.37%)

Error estándar de Estimación “Se”: 1.7203

Error de inclinación de la pendiente “Sb”: 0.0923

Valor de “t” (al 95%): 2.056 (Anexo 2)

Valor de “ t_c ”: 2.088

Para encontrar el valor de “B”, en la fórmula se cambia el símbolo de “z” por el de “t”

Intervalo de “B”: [-0.382 ; -0.003]

¿Qué se puede decir de todo esto entonces?

Conclusión respecto a la hipótesis: con un nivel de confianza de 95% se rechaza la hipótesis inicial dado que según el intervalo encontrado el valor de “B” no pasa por cero, por tanto, sí existe relación entre las variables. Además, $t_c > t$, entonces en la población se mantiene lo que intuitivamente se indicaba, sí existe relación entre estas variables.

¡¡OJO!!

Estrictamente según las condiciones del ejercicio, es verdad que se concluye que sí existe relación entre las variables, pero dado que el valor del extremo del intervalo de “B” (-0.003) es muy cercano a cero y que r^2 indica una fuerza de relación muy baja (14.37%); se puede decir que esta relación es muy frágil, a tal punto que si se cambia el nivel de confiabilidad por ejemplo al 98%, el intervalo para “B” será [-0.421 , 0.036] y como se puede observar “B” **sí** pasa por cero y por tanto no existe relación en la población.

Según el valor de “a” (8.2829) se interpretaría que el valor máximo esperado en esta población respecto a Niveles de retención es de 8.28 puntos, lo cual concuerda con los valores de la muestra y esa sería la interpretación según la teoría; pero dado que hay una contradicción en cuanto al tipo de relación esperado, esto no debería ser aceptado estrictamente.

Respecto a si la ecuación es predictiva, según las condiciones iniciales debe decirse que sí, pero particularmente no aconsejaría se hagan inferencias por las razones antes expuestas.

Ejemplo 3

El vicerrector de una institución de educación media recaba la siguiente información referente al tiempo (en minutos) en que ejecutan una tarea de análisis numérico los docentes del plantel luego de haber recibido una capacitación (en horas) de análisis estadístico en Excel. Con ello quiere saber si la inversión en dicho proceso ha valido la pena. Según la oferta del contrato de capacitación, se supone que el tiempo que deben utilizar los docentes en el análisis luego de recibir la capacitación deberá estar en promedio entre 12 y 15 minutos. Se seleccionó una muestra de 44 profesores con los siguientes resultados:

Tiempo en HORAS	Tiempo en MINUTOS	Tiempo en HORAS	Tiempo en MINUTOS	Tiempo en HORAS	Tiempo en MINUTOS	Tiempo en HORAS	Tiempo en MINUTOS
27	16	25	15	25	20	24	15
24	16	23	16	20	16	25	18
12	25	18	15	21	18	20	19
22	17	20	21	20	14	20	16
13	26	15	22	23	17	25	17
29	14	13	24	15	23	28	15
14	20	20	20	18	16	30	11
20	14	25	16	21	18	25	20
16	18	20	20	25	16	23	17
21	16	16	18	23	18	15	24
22	18	18	16	30	10	16	17

Aunque el ejercicio no indica específicamente qué hacer, la pregunta de si la inversión ha valido la pena, nos lleva a pensar que debemos saber si existe alguna relación entre el tiempo invertido en la capacitación y sus resultados en la práctica; porque de existir, es obvio pensar que sí será una buena inversión capacitar a los docentes.

No estaría por demás también realizar un análisis descriptivo del comportamiento de las variables para tener un panorama más amplio y claro. Desarrollaré entonces el proceso seguido en los ejemplos anteriores.

Variable Independiente: capacitación en análisis estadístico en Excel a profesores de una institución de educación media.

Variable Dependiente: ejecución de una tarea de análisis numérico.

¿Existe relación? Sí

Tipo de relación esperada: inversa; a mayor tiempo de capacitación, mejor destreza esperada y por tanto menor tiempo de ejecución.

En la figura 185 se presenta el gráfico, la ecuación y el valor de r^2

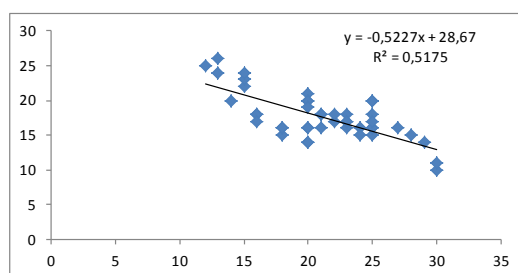


FIGURA 185: GRÁFICO DE DISPERSIÓN, RECTA DE TENDENCIA, ECUACIÓN Y VALOR DE r^2

Según los valores encontrados hasta el momento, parece que sí se cumplen las expectativas y la relación es bastante buena.

En la tabla 68 se presenta el resumen de los cálculos de “r”, Se, Sb, Zt (95%), Zc, comparación de Zt y Zc e intervalo de “B”.

r	0,7194
Se	2,3811
Sb	0,077
Zt	1,96
Zc	6,7119
Zc > Zt	
B	[-0,673 ; -0,371]

TABLA 68. RESUMEN DE ESTADÍGRAFOS

Con los datos presentados se puede afirmar que con un 95% de confiabilidad se rechaza la hipótesis inicial (H_0) dado que $B \neq 0$ y que $Z_c > Z_t$, por tanto, sabemos ya que en la población sí existe relación y que la ecuación es predictiva.

Por lo tanto, ante la inquietud del vicerrector que quiere saber si la inversión en dicho proceso ha valido la pena, puede quedar tranquilo dado que la capacitación dio frutos; pero debe realizar un análisis más exhaustivo en términos descriptivos para tener una idea más clara de lo que pasó con cada variable individualmente.

En la tabla 69, se presenta el resumen de los cálculos estadísticos

	Tiempo en HORAS	Tiempo en MINUTOS
Media	21,02	17,68
Mediana	21	17
Moda	20	16
Desviación estándar	4,66	3,39
Coefficiente de asimetría	-0,05	0,49
Rango	18	16
Mínimo	12	10
Máximo	30	26
Cuenta	44	44

TABLA 69. RESUMEN DE ESTADÍGRAFOS DESCRIPTIVOS

Según estos resultados se puede decir lo siguiente:

Tomando en cuenta los valores del coeficiente de asimetría en ambos casos estos están dentro del intervalo $[-0.5 ; 0.5]$ considerado como distribución es normal; lo mismo ocurre si se comparan los valores de las tres medidas de tendencia central ya que son muy cercanos entre sí.

Para el caso de la variable independiente no se puede emitir un criterio en cuanto al tipo de distribución ya que no hay un dato con el cual comparar, pero en lo que se refiere a la variable dependiente ya hay un primer dato que debe llamar la atención al vicerrector ¿cuál es?

Si se fija en los valores de las medidas de tendencia central de la variable dependiente, encontrará que la relación entre ellas es la siguiente: $Mo < Md < \bar{X}$, lo cual indica un sesgo positivo; y revisando el signo del coeficiente de asimetría este indica también un sesgo positivo; en ambos casos se puede calificar a la distribución como buena ya que se evidencia una tendencia hacia valores bajos en cuanto al tiempo de demora en el análisis de datos.

Pero compare los valores de tendencia central con el intervalo esperado según la expectativa en las condiciones del ejercicio; se puede determinar entonces que, cotejando la oferta y la realidad, esta última está entre 1 y 5 puntos por encima de lo que debería ser y esto dista mucho en términos de ejecución y esperanza de los resultados de la capacitación.

Lo ofrecido en referencia a la demora en el análisis de datos no concuerda entonces con los resultados de las medidas de tendencia central, por tanto, en general el grupo no cumple lo que se esperaba, esto obliga (como generalmente ocurre) a realizar grupos para conocer en detalle el cumplimiento de la condición y oferta dada. Para ello se usará el proceso de histograma estudiado en el capítulo 6.

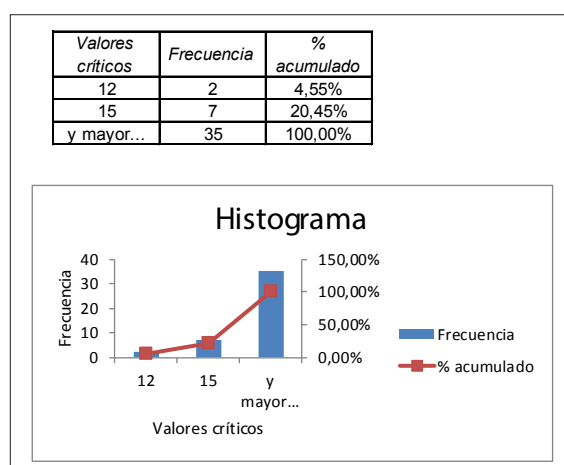


FIGURA 186: DISTRIBUCIÓN DE RESULTADOS SEGÚN LA CONDICIÓN DADA

Recuerde que la lectura de esto se hace así: hasta 12 minutos hay 2 personas, entre 12 y 15 minutos hay 7 y en un tiempo mayor a 15 minutos de demora en el análisis de datos están ¡35 personas!

¿Qué cree que diga ahora el vicerrector sobre estos resultados? La conclusión inmediata es que ¡la capacitación en realidad fue un fracaso! O tal vez que hace falta más práctica, de todas maneras, hay una alerta **roja** según estos resultados. ¿Qué opina usted?

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

1. Un estudiante de Psicología quiere desarrollar un modelo para predecir el tiempo que se necesita para armar un laberinto en base a las horas de entrenamiento. Se seleccionó una muestra de 40 personas y los resultados fueron los siguientes

HORAS	MINUTOS	HORAS	MINUTOS
14	27	8	31
10	30	14	15
12	29	18	20
30	13	25	16
24	20	24	16
15	30	33	10
17	26	45	8
25	16	15	13
20	19	12	31
13	27	10	31
16	25	12	30
10	30	8	32
15	26	5	32
30	12	10	29
35	8	25	18
24	17	20	19
20	18	24	17
20	19	10	30
15	13	15	12
18	21	10	29

- a) Realice TODO el análisis de regresión al 90% (Variables Dependiente e Independiente, tipo de relación esperada, coeficientes de Correlación y Determinación, errores de Estimación e Inclinación, Hipótesis, Intervalo de B, valor de Z_c o T_c según corresponda, Aceptación o rechazo de H_0 .)
 - b) ¿Considera que la relación obtenida es lógica?
 - c) ¿A qué conclusiones cree usted que haya llegado el estudiante?
 - d) Interprete el valor de “a”
 - e) Encuentre una estimación de intervalo para los siguientes valores de la variable independiente: $x = 50$; $x = 10$
2. El director de un centro de salud intuye que quizá haya relación entre el ausentismo y la edad y quisiera tomar la edad de los trabajadores para desarrollar un modelo (al 95% de confianza) de predicción de días de ausencia durante un año laboral. Se seleccionó una muestra aleatoria de 15 trabajadores con los resultados siguientes:
¿Cree usted que el director tiene razón? ¿Considera que la relación es lógica? Según su edad ¿cuántos días de ausentismo estimaría tener? (ojo, pregunta capciosa)

trabajador	edad	días de ausentismo
1	27	25
2	61	6
3	37	10
4	23	18
5	46	9
6	58	7
7	29	22
8	36	11
9	64	5
10	40	8
11	35	9
12	28	20
13	26	25
14	32	20
15	31	21

Para los dos siguientes ejercicios determine:

- ¿Cree que existe relación entre las variables?
 - Variable Dependiente e Independiente
 - Tipo de relación que espera entre las variables
 - Gráfico y ecuación
 - ¿Confirma la ecuación lo que se esperaba en el numeral tercero?
 - Coefficientes de Correlación y Determinación
 - Error estándar de Estimación
 - Error del coeficiente de regresión (error de inclinación)
 - Establezca la hipótesis inicial
 - Encuentre el intervalo para B
 - Encuentre el valor de z_c o t_c según corresponda
 - Compare los resultados de “B” y “z” e indique si se acepta o rechaza la H_0
 - Interprete el valor de “a”
3. En un grupo de jóvenes con propensión al alcoholismo, se realizó un estudio entre las variables agresividad (escala 30 a 90) y adaptación social (escala 35 a 95). Los resultados de dicha investigación se presentan a continuación.

Adaptación social	Agresividad	Adaptación social	Agresividad	Adaptación social	Agresividad	Adaptación social	Agresividad
75	53	81	41	83	60	70	85
54	88	70	76	42	66	88	38
56	73	87	57	86	49	77	80
47	80	74	64	85	42	40	72
36	87	89	50	35	89	80	44
85	54	38	77	48	77	83	72
43	60	42	69	72	53	85	49
81	46	50	65	70	68	65	78
37	81	80	87	85	90	48	62
66	78	51	71	61	71	81	57
64	71	40	90	36	88	71	46
78	76	39	77	95	30	87	41
38	90	52	61	89	47	88	41
35	82	84	55	50	64	88	48
37	81	35	89	71	70	83	83
60	53	59	65	89	46	85	45
						70	55

4. Los datos a continuación se refieren a las variables capacidad de comunicación (medida sobre 50 pts.) y liderazgo (sobre 85 pts.)

Comunicación	Liderazgo	Comunicación	Liderazgo	Comunicación	Liderazgo	Comunicación	Liderazgo	Comunicación	Liderazgo
27	58	10	25	18	32	45	58	21	40
23	42	24	33	20	45	14	25	46	70
23	41	42	65	16	34	48	75	46	50
15	30	33	55	38	64	35	52	49	65
33	50	43	65	49	60	49	76	14	29
40	51	19	42	10	25	49	62	20	39
32	41	44	60	36	72	20	44	49	62
33	59	17	46	21	45	27	42	46	51
25	51	11	26	24	40	46	65	50	85
10	28	31	41	38	52	36	55	16	35
32	49	37	52	38	55	50	84	21	33
28	51	48	70	22	53	13	48	48	67
32	49	17	25	30	46	30	44	29	44
28	50	13	28	13	31	15	33	28	50
48	80	30	45	11	25	21	25	15	31

5. Una institución educativa decidió organizar un módulo especial para los alumnos de último nivel para prepararlos en sus exámenes de ingreso a la universidad, al final del módulo se realizó una evaluación sobre 100 puntos para que los estudiantes tuvieran una idea respecto a los posibles resultados, esto se contrastó con el número de horas que cada uno indicó había trabajado. ¿Cree usted que haya relación entre las variables estudiadas?, ¿qué tipo de relación espera encontrar? A continuación, los datos obtenidos de los 30 alumnos.

HORAS DE ESTUDIO DURANTE EL MÓDULO	CALIFICACIÓN FINAL DE LA EVALUACIÓN	HORAS DE ESTUDIO DURANTE EL MÓDULO	CALIFICACIÓN FINAL DE LA EVALUACIÓN
20	64	28	71
16	61	24	80
34	84	17	60
23	70	22	69
27	88	26	87
32	92	26	83
18	72	25	70
22	77	30	80
38	89	22	75
20	70	25	85
39	80	17	65
38	85	38	87
36	89	19	65
40	90	38	90
28	87	40	91

6. La profesora de tercero de básica intuye que los alumnos pierden puntos debido al exceso de tareas asignadas por los profesores durante cada semana, dado que no alcanzan a realizar todo lo solicitado en cada materia y por tanto entregan algunos trabajos sin terminar o definitivamente mal realizados. Ante esto ha recabado información (cuadro adjunto) de las tareas asignadas y el promedio de notas de las mismas; ayude a la profesora a realizar un análisis completo sobre esto y verificar si tiene o no razón.

Nota: se sugiere realizar un análisis descriptivo y también hacer intervalos en cada variable, (para la variable independiente con amplitud 3)

Número de tareas asignadas	Promedio de notas / 5 pts.	Número de tareas asignadas	Promedio de notas / 5 pts.	Número de tareas asignadas	Promedio de notas / 5 pts.
18	2	8	4	19	1
18	2	17	2	5	5
6	4	18	2	20	1
12	3	16	3	14	2
20	1	18	1	9	4
12	3	16	2	19	1
10	3	13	3	14	2
5	4	16	3	16	2
9	3	18	2	15	4
18	1	12	3	11	3

7. En un establecimiento educativo de nivel medio se tomaron dos pruebas como parte del proceso de ingreso a nuevos alumnos, estas pruebas establecían los niveles de Comprensión oral y Expresión escrita. Los resultados de todos los candidatos se muestran en la tabla siguiente. Si considera que las variables sí tienen relación ¿cuál considera es la variable dependiente? Realice todo el proceso y exprese recomendaciones a los directivos de dicha institución.

Comprensión oral	Expresión escrita	Comprensión oral	Expresión escrita
16	29	26	30
49	49	32	43
30	38	38	42
20	30	45	39
34	31	33	29
23	26	22	25
48	47	48	47
17	15	17	16
33	41	30	40
42	40	42	40
24	31	24	32
30	30		

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1

Obviamente las variables no pueden ser las horas o los minutos, por tanto, la pregunta es ¿qué pretende este ejercicio en cuanto a relación?, la respuesta determinará las variables dependiente e independiente.

El tiempo en armar el laberinto entonces dependerá del entrenamiento dedicado a ello, es decir:

Variable independiente: capacitación para armar laberintos (medido en horas)

Variable dependiente: destreza para armar un laberinto (medido en minutos)

La relación sí existe y se espera que sea inversa, a mayor entrenamiento menor tiempo se demorará en armar el laberinto.

NOTA: aquí cabe una aclaración en cuanto al tipo de relación, se puede decir también que a mayor entrenamiento (capacitación) mayor destreza en armar el laberinto lo cual indicaría que la relación será directa; pero debe tomarse en cuenta que, si adquiere mayor destreza, significará que se demorará menos y la destreza se está midiendo en tiempo; menor tiempo, mayor destreza.

a	35,127
b	-0,7439
r	-0,8307
r ²	0,6901
Se	4,2828
Sb	0,0809
Zt	1,64
Zc	9,1994
B-	-0,8765
B+	-0,6113

El valor de "r" es alto lo cual indica una muy buena relación entre las variables

Tanto por la comparación de los valores "z" como por el intervalo de "B", la conclusión es rechazar Ho. Por tanto sí existe relación entre las variables estudiadas en la población

TABLA 70. CÁLCULOS EJERCICIO 1

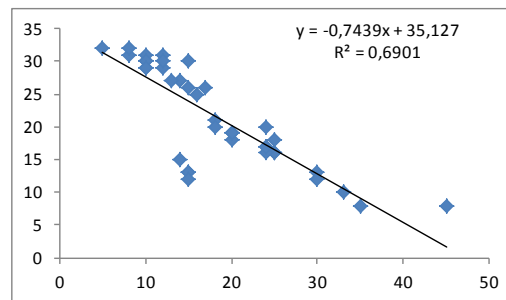


FIGURA 187: GRÁFICO DE DISPERSIÓN DEL EJERCICIO 1

El valor de "a" es 35.127, esto significa que este es el máximo tiempo que se espera de alguna persona de la población para que pueda armar el laberinto. Aunque este valor no está como dato en la muestra, tampoco está muy lejano del valor observado (32), por lo tanto, se considera un valor lógico esperado.

La relación obtenida sí es lógica ya que se esperaba fuese inversa y con el signo de "b" se comprueba

Dada que la ecuación sí es predictiva (se rechazó Ho), sí se pueden realizar cálculos de estimación, pero el valor solicitado de $x = 50$ no se puede remplazar ya que está fuera del

intervalo de los valores de la variable independiente; por tanto, solo se puede hacer una predicción para el valor de $x = 10$.

Estimación puntual: $Y = 27.688$ minutos

Estimación de intervalo: $Y: [23.4052 ; 31.9708]$ minutos

Ejercicio 3

El tipo de relación esperada sí se da entre las variables, es decir que a medida que la agresividad crezca, la adaptación social será cada vez menor. La tabla 71 resume lo solicitado en el ejercicio.

En cuanto al valor del término independiente “a” se puede decir que no es un valor lógico ya que la variable dependiente puede tomar valores de hasta 95, en este caso entonces se debe decir que sí es un máximo esperado, pero no lógico según los datos de la muestra y para fines de estudio y análisis se debe tomar el valor máximo de 95 que consta entre los valores muestrales.

a	117,14	
b	-0,7912	
r	0,6652	El valor de "r" indica una buena relación entre las variables
r ²	0,4425	
Se	14,4982	
Sb	0,1119	
Zt (95%)	1,96	Tanto por la comparación de los valores "z" como por el intervalo de "B", la conclusión es rechazar Ho. Por tanto sí existe relación entre las variables estudiadas en la población
Zc	7,0717	
B-	-1,0105	
B+	-0,5719	

TABLA 71. CÁLCULOS EJERCICIO 3

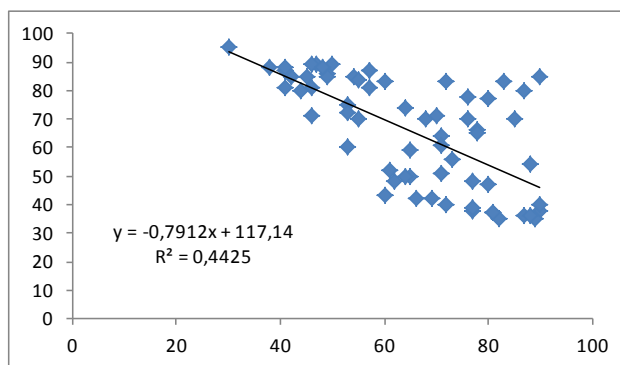


FIGURA 188: GRÁFICO DE DISPERSIÓN DEL EJERCICIO 3

Ejercicio 5

a	49,7	
b	1,0447	
r	0,8198	El valor de "r" indica una excelente relación entre las variables
r ²	0,672	
Se	5,8441	
Sb	0,1379	
Tt (95%)	2,048	Tanto por la comparación de los valores "z" como por el intervalo de "B", la conclusión es rechazar Ho. Por tanto sí existe relación entre las variables estudiadas en la población
Zc	7,5741	
B-	0,7622	
B+	1,3272	

TABLA 72. CÁLCULOS EJERCICIO 5

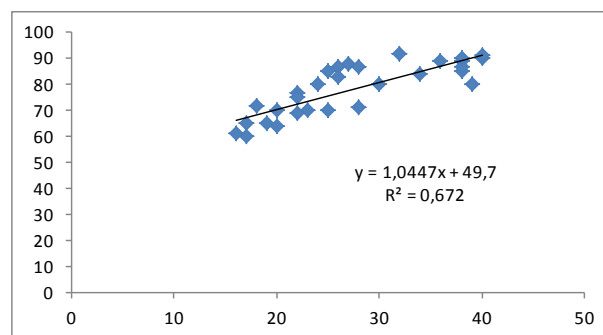


FIGURA 189: GRÁFICO DE DISPERSIÓN DEL EJERCICIO 5

Se puede indicar estos resultados a los estudiantes e insistir en que en realidad la dedicación de horas de estudio sí influye en el rendimiento esperado dado que la correlación es muy alta.

El valor de "a" (49.7) debe interpretarse como un valor mínimo a esperarse (la relación es directa) aunque el valor más bajo conocido sea de 60.

Ejercicio 7

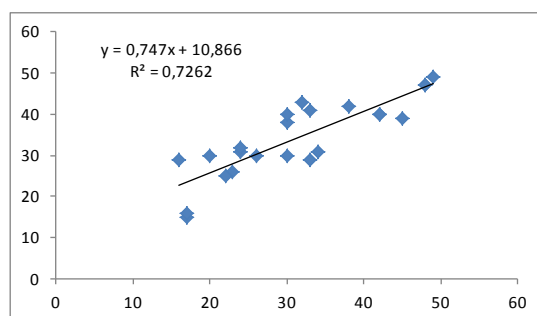


FIGURA 190: GRÁFICO DE DISPERSIÓN DEL EJERCICIO 7

a	10,866	
b	0,747	
r	0,8522	Los valores de "r" y "r ² " indican una excelente relación entre las variables
r ²	0,7262	
Se	4,9158	
Sb	0,1001	
tt (95%)	2,08	tc > tt, se rechaza entonces la Ho
tc	7,4625	
B-	0,5388	Según estos resultados, en la población sí existe relación entre las variables
B+	0,9552	

TABLA 73. CÁLCULOS EJERCICIO 7

Dado que la relación entre las variables es muy alta, los directivos pueden tener la seguridad de que si desean evaluar la capacidad de escritura para cuando ya sean alumnos regulares, pueden por ejemplo contarles alguna historia y que luego de ello los aspirantes escriban sus ideas sobre lo que escucharon; esto también estimulará la comprensión verbal.

Dados estos resultados, al parecer las pruebas de comprensión verbal sí son buenas predictoras de la capacidad de expresión escrita y si se desea evaluar esta última competencia deberán mantener este sistema.

CAPÍTULO 10:

REGRESIÓN LINEAL MULTIVARIABLE (TRES O MÁS VARIABLES)

EL ARTE CULINARIO COMO PREÁMBULO ILUSTRATIVO A LA REGRESIÓN MULTIVARIABLE

En mis clases siempre he iniciado este tema con un proceso culinario para que se pueda entender mejor lo que ocurre cuando se combinan varias variables. Aunque no es algo ortodoxo (muchas veces procuro salir de los esquemas) y además parezca raro ¡y en verdad lo es!, el proponer como ejemplo el proceso que se sigue al preparar un locro o más aun la fanesca, o de pronto una pizza o una paella; es muy ilustrativo para entender qué les ocurre a las variables intervinientes en un análisis correlacional multivariable.

Las variables independientes – en este caso ingredientes – deben tener relación con el objetivo (variable dependiente que en este caso será una fanesca); estas variables serán: leche, agua, refrito (cebolla, perejil, culantro, achiote, aceite), pescado, sal y cada uno de los granos que la componen; cada uno de estos ingredientes tienen sus respectivas y particulares características, por ejemplo los granos son duros; la sal es granulada; el perejil es verde, etc.

Por tanto, tenemos lo siguiente:

Variable dependiente (objetivo a lograr): fanesca

Variables independientes: ingredientes

Variables que mejor se relacionan con el objetivo (por ejemplo): leche, agua, granos.

Tipo de relación de estas variables con el objetivo: directa (mejor calidad de los granos implica mejor calidad de la fanesca, mayor cantidad de leche o agua tiene como consecuencia mayor cantidad de fanesca, claro que alguien puede decir que “se aguó el producto”, pero...)

Ya en el proceso de elaborar la fanesca en cuestión, la receta puede decir algo como: ponga a hervir el agua (recuerde que dije era una variable básica), es decir debe tener una fuerte relación con el objetivo, en términos de Estadística, el valor de “ r ” del agua en relación a la fanesca debe ser alto, pero si en este momento toma una cuchara y se toma un poco de agua, usted no podrá decir que el agua tiene una alta relación con la fanesca porque está ¡muy lejos de parecerse a ese sabroso potaje!

Esto pasará con cada una de las variables que escogió como independientes, aunque sean muy importantes la relación individual “ r ” de cada una de ellas con el objetivo es realmente

cero. Esto es lo que ocurre cuando se realizan análisis de regresión simple y, especialmente en temas de Psicología y Educación, no es recomendable hacer procesos correlacionales simples, ya que una variable por más importante que sea si la relacionamos sola con cualquier objetivo a analizar, siempre dará una información incompleta.

Volviendo al ejemplo, digamos que ya está preparada la fanesca y usted prueba un primer bocado que contiene por ejemplo una haba, el haba antes de ingresar a la olla es dura pero cuando usted la come está blanda, es decir este ingrediente perdió una característica muy particular, tampoco podrá esperar que la fanesca se vuelva verde por la característica de color del perejil y culantro y así pasará con cada variable; pero es obvio que cada variable independiente aportó con lo suyo para lograr el objetivo; ¿puede entonces determinar ahora con exactitud el tipo de relación de cada variable con la dependiente? ¿Y la fuerza de relación de cada una? ¿Qué pasó con la leche y el agua que son fundamentales, llega a sentir las en realidad?

Esto es lo que sucede cuando se hacen análisis de regresión multivariable, podemos interpretar el tipo y fuerza de relación de cada variable independiente con la dependiente, pero cuando todas han sido “mezcladas” las cosas pueden cambiar, pero el resultado es el más idóneo ya que la ecuación resultante representa el objetivo logrado con mayor confiabilidad que cuando se hacen análisis solo con una variable independiente.

Cuando se realizan estudios de relación entre dos variables siempre queda flotando la idea de que no es completo, esto debido a que en realidad ninguna variable es causal absoluta de otra, en otras palabras cuando se quiere analizar un comportamiento humano, por ejemplo, evidentemente habrá muchas razones para determinar dicho proceder aunque seguramente haya una o más variables que afecten más que otras, pero de ninguna manera se puede pensar que solo hay una causa que determine dicho comportamiento.

Lo mismo ocurre en el ámbito de la Educación, por ejemplo, las notas de una materia no pueden ser consecuencia solo de la falta de estudio (aunque esta sea una variable muy fuerte), hay que revisar otras variables que estén influenciando en esos resultados.

Esto significa que, aunque el estudio de relación entre dos variables produce resultados importantes, deberá siempre preguntarse qué otras variables influyen en otra que será el objetivo de estudio y recabar datos que permitan tener un panorama más amplio sobre la influencia real y la consecuencia de esa influencia.

PROCESO DEL ANÁLISIS MULTIVARIABLE

El estudio de regresión conceptualmente mantiene la misma teoría y lógica de lo estudiado en el tema anterior, la diferencia es que se añaden más variables independientes que puedan determinar con mayor exactitud la variación de una variable objetivo de estudio (variable dependiente).

Por tanto, este análisis no es otra cosa que la extrapolación del análisis de regresión simple, por ende cumple todos los principios y condicionantes de éste y por ello nos limitaremos a establecer lo siguiente:

La ecuación de los valores de “y” (variable dependiente) está determinada por una ecuación como la siguiente:

$$y = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 \dots + B_nX_n + a$$

Esto significa que la variable dependiente puede estar afectada por “n” variables ($X_1, X_2, X_3, \dots, X_n$) cuya influencia particular podrá ser determinada en un análisis más profundo según avancemos en este proceso.

Las condiciones en cuanto al tipo de relación entonces se mantienen, es decir el signo del coeficiente de cada variable determinará si la relación de cada una será directa o inversa y el valor del término independiente “a” también deberá interpretarse como un valor mínimo o máximo de la variable dependiente (esto lo determinará la lógica en base al valor sobre el cual se mida).

Recordemos el tema de Evaluación Previa de Proyectos (EPP) presentado al inicio del estudio de Estadística Inferencial; allí se decía que una combinación bien planteada de varias variables permitiría determinar con mayor exactitud el éxito de un proyecto.

Trasladando esto al análisis de relación de varias variables, el principio es el mismo, se deben escoger las variables precisas (ahora determinadas como variables independientes), es decir aquellas que en verdad tienen relación con el objetivo de estudio (ahora llamada variable dependiente); obviamente el hecho de encontrar una ecuación multivariable es más preciso que establecer una combinación de premisas con los conectores lógicos pero en ambos casos se trata de determinar si esa combinación entre variables o premisas, permitirá lograr el objetivo propuesto.

Esto es lo que se pretende al encontrar una ecuación multivariable: establecer todos los elementos lógicos que influyen en otro específico, determinar cuál de ellos es el que afecta más, conocer si la población tiene el mismo comportamiento que el encontrado en la muestra y poder inferir resultados de cualquier elemento que pertenezca a la población con altos índices de confiabilidad.

Los estadígrafos a determinar en un proceso de regresión multivariable serán los mismos que los encontrados para la relación entre dos variables, es decir también habrá por ejemplo un error de estándar de estimación, un error del valor del coeficiente “b” de cada variable, un valor de la fuerza de relación al igual que se encontraba en los análisis de regresión simple; el proceso a seguir en Excel se explica con el siguiente ejemplo:

PROCESO PARA ENCONTRAR EL MODELO Y ESTADÍGRAFOS

En un estudio de comportamiento humano, se han determinado algunas variables que al parecer tendrían relación con los niveles de autoestima; para ello se aplicaron pruebas a

20 sujetos y se encontraron datos de cada una de las variables analizadas. Se trata entonces de determinar si existe o no suficiente información como para establecer que dichas variables se relacionan en alto grado o no con el objeto de estudio. A continuación, se presenta el cuadro resumen con los datos obtenidos.

Conducta autodescriptiva / 50	Aceptación en relación a sus pares / 20	Aceptación condiscípulos / 20	Relación con familiares / 20	Confiabilidad en respuestas / 10	Nivel de autoestima / 100
30	8	10	8	4	56
36	12	12	6	8	66
30	10	10	16	6	66
28	12	8	12	3	60
22	10	4	6	4	42
28	10	12	8	6	58
34	16	14	8	4	72
30	8	10	8	5	60
40	12	6	12	6	70
36	16	12	12	8	76
40	10	12	12	7	74
44	8	16	12	6	80
18	6	8	6	6	38
32	10	14	10	8	66
22	12	8	8	8	50
44	12	8	10	2	74
30	6	14	8	2	58
26	16	6	8	2	56
22	12	8	8	6	50
20	12	2	8	2	42

En la redacción de este ejemplo se ha facilitado ya que la variable objetivo (variable dependiente) se refiere al nivel de autoestima y según lo indicado este dependerá de otras variables (independientes) que en este caso son: Conducta auto descriptiva, Aceptación en relación a sus pares, Aceptación condiscípulos, Relaciones familiares y Confiabilidad en respuestas.

La primera pregunta a resolver será si para usted las variables planteadas tienen o no relación con la variable dependiente; además intuitivamente (o por conocimiento del tema) usted podrá adelantarse a decir cuál de estas variables es la que mejor se relaciona con la Autoestima; pero esto deberá ser comprobado ya que la teoría no siempre concuerda con la realidad de una población a otra.

Otro tema a plantearse, previo a la resolución del ejemplo, es que usted determine qué tipo de relación espera que cada variable tenga con la variable dependiente.

A continuación entonces paso a explicar el proceso a seguir.

A diferencia de lo exigido para resolver ejercicios de regresión lineal simple en cuanto a la “posición” que debe ocupar la variable dependiente en la hoja de Excel, en este caso no importa si se la ubica a la izquierda o derecha de todas las variables independientes; para este caso se encuentra a la derecha.

Utilizamos el mismo proceso seguido para encontrar la ecuación de Regresión Simple, es decir hay que buscar en Excel la función: “ESTIMACIÓN.LINEAL” y seguir los mismos pasos ya detallados con anterioridad para el uso de esta función.

1. Resalte seis columnas hacia la derecha (hay 6 variables en total) y tres filas hacia abajo según se indica en la figura 191

	A	B	C	D	E	F
1	Conducta autodescriptiva / 50	Aceptación en relación a sus pares / 20	Aceptación condiscípulos / 20	Relación con familiares / 20	Confiableidad en respuestas / 10	Nivel de autoestima / 100
2	30	8	10	8	4	56
3	36	12	12	6	8	66
4	30	10	10	16	6	66
5	28	12	8	12	3	60
6	22	10	4	6	4	42
7	28	10	12	8	6	58
8	34	16	14	8	4	72
9	30	8	10	8	5	60
10	40	12	6	12	6	70
11	36	16	12	12	8	76
12	40	10	12	12	7	74
13	44	8	16	12	6	80
14	18	6	8	6	6	38
15	32	10	14	10	8	66
16	22	12	8	8	8	50
17	44	12	8	10	2	74
18	30	6	14	8	2	58
19	26	16	6	8	2	56
20	22	12	8	8	6	50
21	20	12	2	8	2	42

FIGURA 191: CELDAS RESALTADAS PARA OBTENER LA MATRIZ CON LOS ESTADÍGRAFOS

2. Busque la función “ESTIMACIÓN.LINEAL”, deberá obtener el cuadro de diálogo que se presenta en la figura 192.

	A	B	C	D	E	F
1	Conducta autodescriptiva / 50	Aceptación en relación a sus pares / 20	Aceptación condiscípulos / 20	Relación con familiares / 20	Confiableidad en respuestas / 10	Nivel de autoestima / 100
2	30	8	10	8	4	56
3	36	12	12	6	8	66
4	30	10	10	16	6	66
5	28	12	8	12	3	60
6	22	10	4	6	4	42
7	28	10	12	8	6	58
8	34	16	14	8	4	72
9	30	8	10	8	5	60
10	40	12	6	12	6	70
11	36	16	12	12	8	76
12	40	10	12	12	7	74
13	44	8	16	12	6	80
14	18	6	8	6	6	38
15	32	10	14	10	8	66
16	22	12	8	8	8	50
17	44	12	8	10	2	74
18	30	6	14	8	2	58
19	26	16	6	8	2	56
20	22	12	8	8	6	50
21	20	12	2	8	2	42

FIGURA 192: CUADRO DE DIÁLOGO PARA ENCONTRAR VARIOS ESTADÍGRAFOS DE REGRESIÓN MULTIVARIABLE

En “Conocido_Y” resalte los valores correspondientes a la variable dependiente, en “Conocido_X” resalte los valores correspondientes a las variables independientes, en “Constante” y en “Estadística” digite el valor 1 para cada caso, ¡¡ALTO!!, recuerde que va a encontrar lo que se conoce como fórmula matricial, por lo tanto no debe hacer *click* en aceptar, su proceso debe mostrarse según se indica en la figura 193.

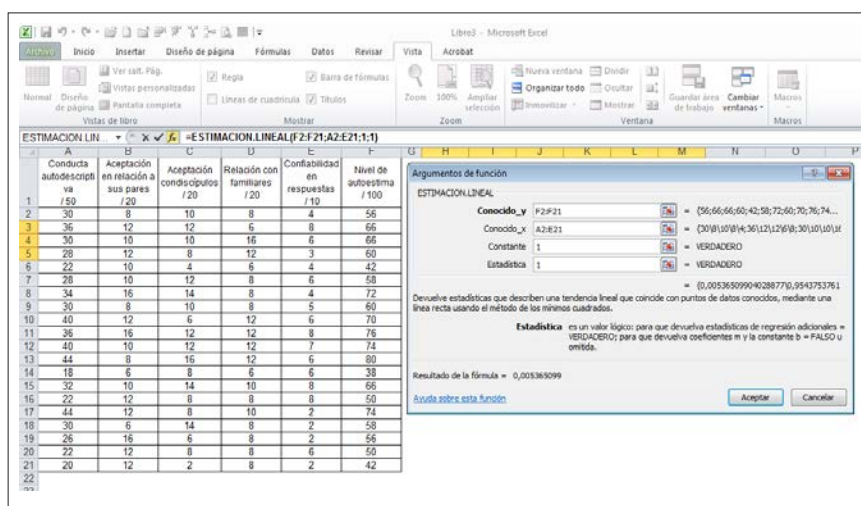


FIGURA 193: ARGUMENTOS INGRESADOS PARA OBTENER LOS DATOS DE REGRESIÓN MULTIVARIABLE

- En este momento debe apastar de manera independiente y sin soltar las teclas: Control – Shift – Enter y cuando las tres estén ya apastadas suéltelas despacio, con cierto “ritmo” y al mismo tiempo.
- Deberá obtener una matriz como la que se muestra en la figura 194.

FIGURA 194: MATRIZ CON LOS RESULTADOS DEL PROCESO

Nota: no se preocupe por los símbolos “#N/A” significan “no aplica”, no es un error de proceso.

La matriz que arroja Excel es la que se presenta a continuación en la tabla 73:

COEFICIENTES DE LAS VARIABLES INDEPENDIENTES					
b_5	b_4	b_3	b_2	b_1	a
0.0053651	0.95437538	0.98508667	0.92655827	1.01250459	1.15921253
0.11877378	0.10404405	0.08663444	0.08344834	0.04183159	1.35201382
0.99487294	1.00576744				

→ Coeficientes “x” y valor de “a”
 → Valores Sb
 → Valores de r^2 y Se

TABLA 73. RESULTADOS OBTENIDOS DEL EJEMPLO DE REGRESIÓN MULTIVARIABLE

La información obtenida en cada fila de la matriz encontrada es la siguiente:

Primera fila: coeficientes de las variables independientes de la ecuación multivariable y valor de “a” (término independiente de la ecuación).

Segunda fila: valores del error de inclinación (Sb) de cada uno de los coeficientes

Tercera fila (de izquierda a derecha): valor del coeficiente de determinación r^2 y error de estimación (Se).

En cuanto a los valores de la primera y segunda filas, hay que indicar que Excel devuelve los valores de los coeficientes y errores de inclinación de cada variable independiente de manera inversa a lo que supuestamente están ordenados; es decir, para la lectura de los datos, se supone que las variables se numerarían de izquierda derecha desde 1 a 5 (en las independientes), pero los valores de la ecuación encontrada van de derecha a izquierda de manera correspondiente.

Para este caso tenemos que la variable X_1 sería “Conducta auto descriptiva” y su coeficiente es 1.012505, la variable X_2 es “Aceptación en relación a sus pares” y su coeficiente es 0.926558 y así sucesivamente avanzando de derecha a izquierda. Al último valor de la matriz en el extremo derecho (1.159213 para este caso) siempre corresponderá al valor del término independiente “a”.

Por lo tanto la ecuación multivariable deberá escribirse de la siguiente manera:

$$Y = 1.0125 X_1 + 0.9265 X_2 + 0.9850 X_3 + 0.9543 X_4 + 0.0053 X_5 + 1.1592$$

Según esto todas las variables tienen una relación directa con la variable dependiente, ¿es esto lo que esperaba?

Supongo que la variable 5 (Confiabilidad en respuestas) le llame la atención. Esta variable que no es común en los análisis de regresión se establece realizando un análisis de fiabilidad para las respuestas de aquellas personas a las que se les aplicó el test de autoestima para determinar cuán confiables son las respuestas a dicha prueba. Por lo tanto, el comportamiento de esta variable y la influencia en los resultados puede ser algo distorsionador ya que en realidad no mide un comportamiento humano como las otras variables.

Tenemos entonces la ecuación y el tipo de relación que cada variable tiene con la dependiente, el valor del término independiente “a” estrictamente debe interpretarse en este caso como el valor mínimo esperado en autoestima, pero supongo le parecerá algo absurdo que se pueda esperar que una persona tenga valores tan bajos especialmente si la variable se mide sobre 100.

En estos casos debe escogerse el valor más bajo de los datos procesados y decir que ese sería el mínimo puntaje de autoestima que se espera encontrar en la población estudiada; para el caso del ejemplo ese valor sería 38, que es un valor más lógico (aunque algo preocupante también).

En cualquier análisis de Regresión ya sea Simple o Multivariable, es importante conocer los valores sobre los cuales se mide cada variable, por ejemplo, para poder interpretar el valor

de “a” sabemos que la variable de autoestima se mide sobre 100, es obvio entonces que el valor de 1.1592 (a interpretarse como mínimo o máximo) por lógica en base al valor sobre el cual se ha medido la autoestima, no puede ser un máximo esperado. Pero por ejemplo si una variable cualquiera se mide sobre 10 y el valor de “a” es 6, la interpretación directa es que ese valor es lo máximo esperado para esa población, esto será bueno, aceptable o malo según las condiciones de esa variable o estudio realizado.

CORRELACIÓN DE CADA VARIABLE CON LA VARIABLE DEPENDIENTE

Continuando con el análisis hay que interpretar los valores de los coeficientes de las variables independientes no solo por el signo, es decir por el tipo de relación, si no que se debe conocer cuál de las variables es la que mejor se relaciona con la variable dependiente.

Cuando se estudia el tema solo con una variable independiente, Excel encuentra el valor de “r” que indica la correlación (fuerza de relación) que tiene esa variable en específico con la variable dependiente; para el caso de regresión multivariable, la matriz que arroja Excel solo se conoce el valor de “r²” (y por tanto “r”) pero este valor representa cuan fuerte es la relación que tienen **todas las variables juntas con la variable dependiente**.

Lastimosamente con este proceso en Excel no se puede saber el valor de “r” de cada variable con la variable dependiente que sería la única forma objetiva para saber qué variable o variables se relacionan mejor con la dependiente.

Si no se hace un proceso de cálculo que permita conocer los valores de la correlación de cada variable con la variable dependiente, puede (no es lo ideal) tomarse en cuenta los coeficientes de la ecuación y determinar la mejor o menor relación de cada variable en base al **valor absoluto** de esos coeficientes, es decir mientras más alto es el coeficiente mejor será la relación con la variable dependiente.

En el caso del ejemplo desarrollado tendríamos que la variable 1 (Conducta auto descriptiva) es la que mejor se relaciona ya que es el valor más alto de todos los coeficientes, luego será la variable 3 (Aceptación condiscípulos), luego, en su orden, las variables cuatro, dos y cinco.

La razón matemática de esta forma de determinar qué variable es la que mejor o menor relación tiene con la variable dependiente es que cada coeficiente representa la “rapidez” de cambio que tiene la variable dependiente; por tanto un coeficiente más alto significará un mayor impacto, esto se explica también indicando que los coeficientes de cada variable representan de manera individual la pendiente (inclinación) de la recta para el caso de relaciones individuales.

Dada la ecuación del ejemplo

$$Y = 1.0125 X_1 + 0.9265 X_2 + 0.9850 X_3 + 0.9543 X_4 + 0.0053 X_5 + 1.1592$$

Será fácil determinar que el valor de 1.0125, que es mayor a cualquiera de los demás, tendrá más influencia que los otros y el valor de 0.0053 indicará que esa variable (confiabilidad en respuestas) será la que menos influya en la variable dependiente.

Pero como dijera antes, esta forma de determinar la mejor o menor relación que cada variable independiente tiene con la dependiente no es la idónea en términos de análisis de regresión ya que teóricamente el único valor que puede determinar esto es el del coeficiente de correlación.

Por lo tanto, es mejor hacer el cálculo de correlación entre las variables y para esto se debe utilizar la función de “Análisis de datos” para encontrar la matriz de correlación de todas las variables entre sí y esos valores indicarán con certeza qué variable se relaciona mejor con la variable dependiente.

El proceso será el siguiente:

1. En el menú principal de Excel haga click en “Datos”, se desplegarán las funciones que se relacionan con este menú y entre ellas al final encontrará **“Análisis de datos”**.
2. Aparecerá un listado con varias funciones, busque la función **“coeficiente de correlación”**, según se indica en la figura 195.

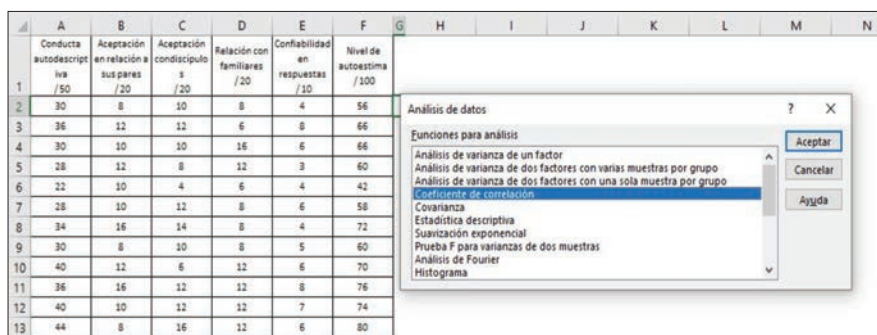


FIGURA 195: PANTALLA QUE PERMITE OBTENER EL COEFICIENTE ENTRE VARIABLES

3. Haga “click” en el nombre de la función y aparecerá el siguiente cuadro de diálogo según se ve en la figura 196.

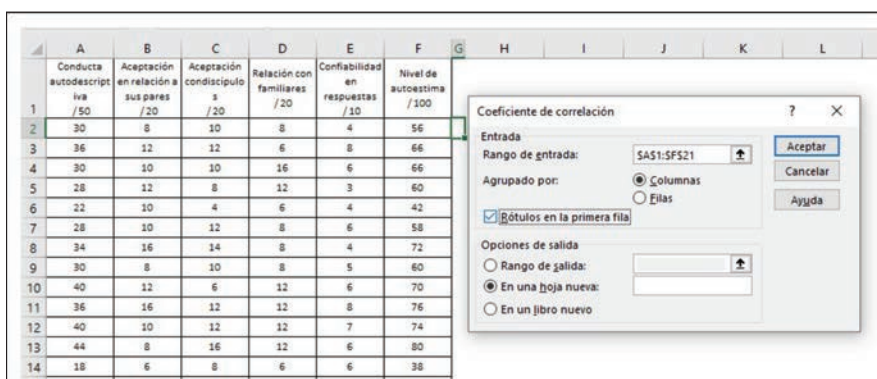


FIGURA 196: PANTALLA QUE PERMITE OBTENER EL COEFICIENTE DE CORRELACIÓN ENTRE VARIABLES

4. En “Rango de entrada” resalte los datos de todas las variables incluidos los nombres de cada columna, sugiero que las demás opciones se las deje según lo previsto en el cuadro de diálogo, luego de esto acepte; obtendrá la matriz con los valores de correlación de todas las variables, según se muestra en la tabla 74.

	Conducta autodescriptiva / 50	Aceptación en relación a sus pares / 20	Aceptación condiscípulos / 20	Relación con familiares / 20	Confiabilidad en respuestas / 10	Nivel de autoestima / 100
Conducta autodescriptiva	1					
Aceptación en relación a sus pares	0,114016084	1				
Aceptación condiscípulos	0,539266195	-0,1906719	1			
Relación con familiares	0,510497899	0,11366994	0,220512863	1		
Confiabilidad en respuestas	0,145614848	-0,00581228	0,395710304	0,16856942	1	
Nivel de autoestima	0,94107079	0,26670916	0,649805476	0,6299831	0,246480101	1

TABLA 74. VALORES DE LOS COEFICIENTES DE CORRELACIÓN ENTRE TODAS LAS VARIABLES

Los valores de la última fila indican la correlación de cada variable con la variable dependiente (Nivel de autoestima).

Como se podrá notar, para este ejemplo ha coincidido totalmente el orden de mayor a menor en que las variables independientes se relacionan con la dependiente, ya que el máximo valor de “r” corresponde a la primera variable (Conducta auto descriptiva), luego será la tercera variable (Aceptación condiscípulos), luego, en su orden, las variables cuatro, dos y cinco.

ANÁLISIS DEL ERROR DE LOS COEFICIENTES

Siguiendo con el proceso analizaré lo que ocurre con los valores de los errores estándar de los coeficientes, es decir los valores de “Sb”.

Estos valores indican si las variables son “consistentes” es decir que mantienen el tipo de relación aunque se reste y sume el error a cada valor de sus coeficientes, para ello basta comparar el valor de cada coeficiente “b” con el valor absoluto de su respectivo error “Sb”, si el valor de este último es mayor al del coeficiente, significa que esa variable tiene “problemas de comportamiento”, es decir que el tipo de relación manifestado no se mantiene al aplicarle el error.

Para nuestro ejemplo podemos observar en la tabla 75 los valores de los coeficientes y sus respectivos errores “Sb”.

X_5	X_4	X_3	X_2	X_1	
0,0053651	0,95437538	0,98508667	0,92655827	1,01250459	Coefficientes
0,11877378	0,10404405	0,08663444	0,08344834	0,04183159	Valores Sb

TABLA 75. COEFICIENTES DE CADA VARIABLE Y SUS RESPECTIVOS ERRORES

Al realizar la comparación, se observa que solo en la quinta variable (Confiabilidad en respuestas) el error es mayor que el valor del coeficiente, esto significa que en realidad en la población ese coeficiente puede variar entre $[-0.1134$ y $0.1241]$; estos valores se obtienen restando y sumando el valor de “Sb” al valor del coeficiente de la variable correspondiente.

Por tanto, aunque el valor de 0.00536 indique una relación directa con la variable dependiente (en la muestra), en la población este tipo de relación puede cambiar, es por ello que a esa variable se la puede calificar como inconsistente.

Esto ya se podía intuir al plantear el ejercicio ya que como se había indicado, esta variable era la única que no se refería a un comportamiento humano como lo son las otras y esto podía distorsionar el análisis.

AJUSTE DEL MODELO

Suele ocurrir que si el valor del coeficiente de una variable es muy cercano a cero (como lo que ocurre en este caso), seguramente esa variable tendrá un comportamiento ambivalente; esto puede deberse a que se ha utilizado una variable que en realidad no tiene mucha relación o influencia en el comportamiento de la variable, lo cual desde el punto de vista lógico y teórico es explicable dado que todas las variables tienen que ver con un comportamiento humano y la quinta variable es más bien un tema del instrumento con el cual se ha evaluado.

Cuando ocurren estos casos en los que por error se tomó en cuenta una variable que demuestra poca o ninguna relación con la variable objetivo, se recomienda eliminar del modelo a dicha variable y rehacer el análisis.

Debo reconocer que Excel en esto sí es limitado ya que utilizando un software específico de Estadística este procedimiento se haría automáticamente entregando como resultado final el modelo ya reducido de mejor ajuste. Para este caso el SPSS por ejemplo es una herramienta muy eficaz, ¡punto para el SPSS!

Eliminando entonces la variable que se demostró no tener relación con la variable dependiente, al realizar el nuevo proceso se obtiene la siguiente matriz expresada en la tabla 76:

b_4	b_3	b_2	b_1	a	
0,95506952	0,98665999	0,92693068	1,01218855	1,17073751	→ Nuevo modelo
0,09942092	0,07664023	0,08023017	0,03984678	1,28279405	→ Valores Sb
0,9948722	0,97173446				→ Valores de r^2 y Se

TABLA 76. NUEVO MODELO LUEGO DE ELIMINAR LA VARIABLE INCONSISTENTE

Este nuevo modelo ajustado es muy similar al inicial, pero refleja un mejor comportamiento de las variables y ya no existe esa variable que pueda distorsionar el análisis, lo cual demuestra que esa variable fue mal escogida.

Tomando el ejemplo de la fanesca, sería como agregar ¡papa! o ¡espinaca! como uno de los elementos de la receta, indudablemente estos ingredientes distorsionarán el sabor o la consistencia del objetivo.

ANÁLISIS DEL COEFICIENTE DE DETERMINACIÓN R^2

El valor del coeficiente r^2 es de 0.9948 que leído como porcentaje significa que en el 99.48% de los casos, las cuatro variables escogidas sí se relacionan con la variable objetivo, en otras palabras las variables independientes sí son predictoras de la autoestima, se puede decir también que la autoestima puede ser explicada en función de: la conducta auto descriptiva, la aceptación en relación a sus pares, la aceptación de los condiscípulos y la relación con familiares en un 99.48%.

Tomando en cuenta que el valor de “ r ” es la raíz cuadrada del coeficiente de determinación, podemos determinar que el valor de “ r ” es igual a 0.997 que es un valor muy cercano a 1 y nos indica que las variables independientes están relacionadas casi en su totalidad con la Autoestima.

Cuando en el capítulo anterior se hizo el estudio de los coeficientes de correlación y determinación, se indicaba que es complejo determinar qué valor de “ r ” o “ r^2 ” sería “aceptable” para saber si en verdad dos variables se relacionan o decidir si la fuerza de relación es alta, mediana o baja.

En el caso de regresión multivariable, ya se puede “exigir” valores para el coeficiente de correlación del modelo general ya que a diferencia de la regresión simple en la cual se indicó no puede haber causalidad, este modelo multivariable ya pretende que, al tener las variables correctas, estas sí predigan el comportamiento del objetivo de estudio.

Comparando con nuestra fanesca, aunque el agua sea una variable fundamental y supondríamos una alta relación, por sí sola no produce el resultado esperado por tanto el valor de r^2 entre agua y fanesca en realidad será mínimo, pero al agregarse los demás ingredientes el coeficiente de determinación será cada vez más alto hasta alcanzar el valor máximo de 1.

Lo mismo ocurre con variables de comportamiento humano y es por ello que en un proceso de diagnóstico, el psicólogo o el educador necesitan conocer todas las variables (síntomas) que se relacionen con la patología o problema académico a tratar, cada una de ellas se relacionará de una u otra forma y con distinto grado, pero si al presentarse varios síntomas que tengan relación concreta con el problema, el diagnóstico será más preciso. En el ejemplo que estamos revisando, se puede concluir entonces que, dado que el valor de r^2 es muy alto el valor del nivel de autoestima estará determinado con mucha fuerza por las variables escogidas.

Por tanto los valores del coeficiente de determinación a esperarse en un análisis multivariable, deben superar el 90% y en términos del coeficiente de correlación, se espera sean de 0.95 o más.

Esto concuerda con los porcentajes de confiabilidad esperados en cualquier análisis estadístico y recomendados ya, por ejemplo, en el tema correspondiente a comprobación de hipótesis.

ANÁLISIS DEL ERROR ESTÁNDAR DE ESTIMACIÓN S_e

En el capítulo anterior se estableció que el error estándar de estimación establece de alguna manera el grado de confiabilidad de la predicción de la ecuación de estimación encontrada.

Recordemos entonces que los errores calculados en Estadística son desviaciones estándar, en este caso el valor de “ S_e ” nos está diciendo que hay una diferencia mínima al estimar valores predictivos de la variable dependiente.

Tomando en cuenta que la variable a analizar es la autoestima y que esta se mide sobre 100, es innegable que un error de estimación de 0.9717 realmente es mínimo. Esto quiere decir que el modelo predictivo es muy confiable y que cualquier estimación que se pueda hacer en esta población será muy certera.

PREDICCIÓN CON LA ECUACIÓN MULTIVARIABLE

Por ejemplo, si quisiéramos predecir los niveles de autoestima de alguna persona de la población en ese estudio de comportamiento humano que hemos ejemplificado, y llegáramos a conocer sus valores en cada una de las variables independientes, al pronosticar su nivel de autoestima según el modelo encontrado este será muy certero.

Supongamos entonces los siguientes valores para cada variable independiente:

Conducta auto descriptiva: 40

Aceptación en relación a sus pares: 12

Aceptación condiscípulos: 14

Relación con familiares: 13

Al remplazar estos valores en la ecuación de estimación, el resultado sería el siguiente: 66.59; pero como sabemos hay un error de estimación de 0.9717, se debe restar y sumar dicho error al valor calculado, haciendo esto se obtiene el siguiente intervalo: [65.62 ; 67.56].

Por tanto con un alto nivel de confiabilidad se puede afirmar que si una persona ha presentado esos valores en cada variable independiente, se podrá pronosticar casi con certeza (recuerde que “ r ” = 0.997) que su nivel de autoestima estará entre 65.62 y 67.56.

OJO: Para el cálculo de estimación, debe utilizarse el valor de “ a ” obtenido en la ecuación sea o no lógico.

De manera estricta en el proceso de regresión multivariable estos pasos son los que se deben realizar, pero siempre he recomendado que también se haga un análisis descriptivo de

las variables intervinientes ya que una cosa es determinar si existe relación entre variables y otra saber el comportamiento de las mismas, esto siempre dará más luces al análisis general de cualquier investigación a realizar.

Por lo tanto y para terminar con el ejemplo propuesto, a continuación (en la tabla 77) se presenta el resumen numérico en cuanto a lo descriptivo de las variables planteadas.

	<i>Conducta autodescriptiva / 50</i>	<i>Aceptación en relación a sus pares / 20</i>	<i>Aceptación condiscipulos / 20</i>	<i>Relación con familiares / 20</i>	<i>Nivel de autoestima / 100</i>
Media	30,6	10,9	9,7	9,3	60,7
Error típico	1,730	0,657	0,811	0,585	2,696
Mediana	30	11	10	8	60
Moda	30	12	8	8	66
Desviación estándar	7,735	2,936	3,629	2,618	12,057
Varianza de la muestra	59,832	8,621	13,168	6,853	145,379
Curtosis	-0,780	-0,227	-0,331	0,557	-0,738
Coefficiente de asimetría	0,202	0,220	-0,282	0,888	-0,310
Rango	26	10	14	10	42
Mínimo	18	6	2	6	38
Máximo	44	16	16	16	80
Suma	612	218	194	186	1214
Cuenta	20	20	20	20	20

TABLA 77. RESUMEN DESCRIPTIVO DE LAS VARIABLES

En términos generales se puede decir lo siguiente:

1. Según las medidas de tendencia central, todas las variables presentan una distribución normal.
2. Este tipo de distribución no es lo esperado, ya que idealmente sería mejor que todas las variables presenten sesgo negativo.
3. Según el coeficiente de asimetría solo la cuarta variable presentaría un sesgo positivo, lo que indicaría que es algo preocupante a tomar en cuenta.
4. La desviación estándar indica mucha variabilidad en los datos, es decir hay mucha dispersión, esto se confirma con el signo de los valores de la curtosis y con los valores del rango.
5. En cuanto a la variable dependiente, los resultados son realmente preocupantes ya que los valores de tendencia central indican estar en el percentil 60, valor bastante bajo para una variable de este tipo.
6. Lo mismo ocurre con las variables independientes, así
 - i. En cuanto a Conducta auto descriptiva, esta indicaría una mala autopercepción del comportamiento personal.
 - ii. Los valores de tendencia central de las otras variables prácticamente no llegan a la mitad del máximo valor esperado, esta información descriptiva de las variables independientes, debe analizarse con más detalle por parte del investigador.

EJERCICIOS DE APLICACIÓN DEL CAPÍTULO

Ejemplo 1

Luego de una investigación referente a la memoria en personas adultas, se han obtenido los siguientes datos, determinar si en la población se puede inferir el nivel de memoria en base a las variables planteadas ¿cuáles serían sus conclusiones y recomendaciones?

Número de palabras con sentido recordadas	Número de palabras sin sentido recordadas	Nivel de memoria / 50	Número de palabras con sentido recordadas	Número de palabras sin sentido recordadas	Nivel de memoria / 50
15	12	25	19	10	32
18	11	42	16	15	31
14	10	21	17	12	35
15	8	26	18	10	32
16	9	32	19	10	30
21	7	43	21	12	45
14	4	22	25	11	45
18	12	48	20	15	40
20	15	41	15	12	28
19	16	40	21	10	42
18	10	39	20	11	39
18	13	38	16	10	31
16	10	31	19	14	41
12	10	25	20	16	40
15	14	34	18	10	39
16	11	33	17	13	33
16	10	32	19	10	37
17	9	36	21	8	36
15	8	29	18	9	40
12	8	34	16	9	30
19	9	30	14	10	26
18	9	35	18	15	35
			22	12	47

Variable dependiente: Nivel de memoria

Variables independientes:

Número de palabras con sentido recordadas

Número de palabras sin sentido recordadas

Tipo de relación esperado:

directa para ambas variables.

Siguiendo el mismo proceso sugerido en Excel, se obtienen los resultados siguientes:

Matriz con los coeficientes de la ecuación, valores de “Sb”, “ r^2 ” y Se, según se muestra en la tabla 78.

x2	x1	a
0,4561972	1,79610523	-1,63999261
0,25154709	0,24004975	4,43011038
0,62900545	4,08640193	

→ Valores de “Sb”
→ Valores de r^2 y Se

TABLA 78. MATRIZ CON ESTADÍGRAFOS DEL EJEMPLO 1

La ecuación entonces será: $y = 1.796 x_1 + 0.456 x_2 - 1.639$

En cuanto al tipo de relación esperada se cumple para ambas variables y sus respectivos errores “Sb” indican que estas son consistentes.

Tomando en cuenta los coeficientes, la variable que mejor explica el nivel de memoria es la primera (número de palabras con sentido recordadas. A continuación (en la tabla 79) se presentan los valores de “r” para ver si concuerda con esto.

	Número de palabras con sentido recordadas	Número de palabras sin sentido recordadas	Nivel de memoria / 50
Número de palabras con sentido recordadas	1		
Número de palabras sin sentido recordadas	0,2610379	1	
Nivel de memoria / 50	0,77456624	0,36672966	1

TABLA 79. MATRIZ DE CORRELACIÓN ENTRE LAS VARIABLES

De acuerdo a estos resultados se comprueba que la variable Número de palabras con sentido recordadas es la que mejor se relaciona con la variable Nivel de memoria.

El valor de “a” (-1.639) es absurdo – es obvio que no puede existir un valor negativo de la variable memoria – en este caso al igual que en el ejemplo desarrollado se debe buscar el valor más pequeño de la muestra e interpretarlo como el mínimo esperado en la población, por lo tanto, se dirá que el mínimo nivel de memoria en la población se espera sea de 21.

Según el valor del coeficiente de determinación (0.629) estas variables parecen no ser muy confiables para predecir el nivel de memoria en la población estudiada, además el valor del error de estimación es casi el 10% respecto al máximo sobre el que se mide la variable dependiente (50), este error, aunque dentro de lo permitido, está en el límite.

Con esto podríamos concluir que la ecuación no es predictiva y si se desea estimar los niveles de memoria en estas personas adultas, se escojan otras o más variables independientes que expliquen de mejor manera la variable a estudiar.

En cuanto al análisis descriptivo, el cuadro resumen se presenta en la tabla 80.

	Número de palabras con sentido recordadas	Número de palabras sin sentido recordadas	Nivel de memoria / 50
Media	17,578	10,867	34,889
Error típico	0,396	0,378	0,977
Mediana	18	10	35
Moda	18	10	32
Desviación estándar	2,659	2,537	6,555
Varianza de la muestra	7,068	6,436	42,965
Curtosis	0,341	0,304	-0,528
Coefficiente de asimetría	0,167	0,174	-0,062
Rango	13	12	27
Mínimo	12	4	21
Máximo	25	16	48
Suma	791	489	1570
Cuenta	45	45	45

TABLA 80. RESUMEN DESCRIPTIVO DE LAS VARIABLES ESTUDIADAS

Tanto por las medidas de tendencia central como por los valores de los coeficientes de asimetría, estas variables tienen un comportamiento normal. Para el caso de estas variables y dado que no existe un parámetro para poder contrastar con los resultados, el tipo de distribución será muy difícil de calificar, pero al menos es lo esperado teóricamente.

En cuanto a la variable dependiente que es la única con parámetro a comparar, indica un nivel de memoria no muy optimista, tomando en cuenta los valores representativos del grupo (MTC), pero dado que hay mucha dispersión sería conveniente establecer grupos para conocer el peso de cada uno en cuanto a niveles de memoria y establecer estrategias que permitan mejorar a quienes lo necesiten.

Ejemplo 2

En una organización pretenden contratar un seguro de vida y desean conocer ciertos datos de estado de salud y costumbres de sus empleados (muestra) para determinar si el costo por riesgo (en cuanto al nivel de colesterol) será o no muy alto. Entre lo analizado se han establecido algunas variables que se presentan en el cuadro siguiente. Determine usted: ecuación, variable que tiene mayor y menor relación con la variable estudiada, variables que no sean consistentes, **multicolinealidad**, si la ecuación es predictiva, interprete el error de estimación y el valor de “a”, haga un análisis descriptivo de cada variable e indique sus conclusiones. Si la ecuación es predictiva, encuentre qué valores, referentes al nivel de colesterolemia, deberían predecirse para los siguientes valores de las variables independientes: 30, 20, 8 y 35 respectivamente desde x_1 hasta x_4

IMC	Sedentarismo	Actividad física semanal	Edad	Colesterolemia
23,0	24	12,0	38	146
28,2	16	9,0	42	148
26,0	17	4,3	40	153
24,8	27	6,0	33	156
29,4	19	6,5	39	164
27,2	23	3,0	42	215
27,5	23	7,7	44	170
28,1	24	3,8	49	175
28,9	32	4,9	45	176
29,8	35	6,3	33	184
29,9	29	4,0	45	185
30,3	32	5,2	37	195
30,6	33	4,8	51	212
31,7	34	5,7	45	214
31,7	40	3,0	44	216
33,1	42	2,0	50	225
31,9	21	9,0	41	166
29,7	25	10,0	44	170
30,0	25	10,0	46	172
30,6	26	10,0	51	177
31,4	34	7,0	47	178
32,3	37	8,0	35	186
32,4	31	9,0	47	187
32,8	34	7,0	39	197
33,1	35	7,0	53	214
34,2	36	4,0	47	216
34,2	42	9,0	46	218
35,6	44	2,0	52	227
34,4	23	11,0	43	168
32,2	27	12,0	46	172
32,5	27	12,0	48	174
33,1	28	11,0	53	179
33,9	36	9,0	49	180
34,8	39	10,0	37	188
34,9	33	12,0	49	189

La matriz que arroja Excel con los datos presentados a continuación es la tabla 81:

x4	x3	x2	x1	a
0,73007831	-3,12637923	1,41137254	1,91661371	74,294728
0,37607239	0,69065728	0,3738901	0,96240459	20,9617486
0,80048106	10,652606			

TABLA 81. ESTADÍSTGRAFOS EJEMPLO 2

La ecuación entonces será:

$$Y = 1.9166 x_1 + 1.411 x_2 - 3.126 x_3 + 0.730 x_4 + 74.294$$

Variable con mejor relación (en función de los coeficientes): Actividad física semanal (x_3).

Variable con menor relación (en función de los coeficientes): Edad (x_4).

Todas las variables son consistentes, los valores “Sb” de cada variable son menores a sus respectivos coeficientes.

En cuanto a si la ecuación es o no predictiva, los criterios seguramente estarán divididos ya que al tomar en cuenta el valor de " r^2 " (80.04%) se puede decir que la relación no es muy fuerte; pero se puede argumentar también que tampoco es un valor muy bajo y además el valor de "Se" es muy pequeño por tanto el error a cometer en la estimación estará dentro de lo aceptable.

Si se toma en cuenta la recomendación de que el porcentaje que determina la fuerza de relación sea mayor o igual al 90%, entonces deberá decidirse que el modelo no permite predecir, más allá de que el error de estimación sea bajo y de que no existan variables inconsistentes.

En cuanto al valor de "a" (74.294) deberá interpretarse como un mínimo esperado del nivel de colesterolemia en la población por las siguientes razones:

1. Todos tenemos colesterol en la sangre.
2. Según los datos, un nivel de colesterol inferior a 200 se considera bueno.

En negrilla he resaltado el término **multicolinealidad**, este se refiere a la contradicción que se presenta entre el tipo de relación lógica esperado y el tipo de relación encontrado en el modelo; es decir que si se espera que alguna variable tenga una relación directa o inversa y en el modelo se determina que es inversa o directa; esa contradicción se traduce en un problema de multicolinealidad.

Esto no puede "arreglarse" numéricamente y tampoco es algo que deba preocupar en la investigación; deberá interpretarse según las circunstancias del modelo y de cada variable.

Trataré de explicar con el ejemplo de la fanesca expuesto al inicio del capítulo. Uno de los ingredientes de la fanesca es el haba, una de las características de este grano es que es dura antes de ingresar en la mezcla, pero luego de la cocción el haba se vuelve blanda, es decir cambió radicalmente, esto también ocurre con las variables, cuando se "mezclan" con las demás, pueden cambiar sus características según unas variables influyen en las otras, para el caso del haba, las variables independientes como el agua, la leche y el calor hacen que cambie una de sus propiedades.

Un ejemplo que suelo utilizar para explicar el extraño comportamiento de alguna variable es en referencia a las actitudes de las personas en general y muy especialmente en los adolescentes. La pregunta que suelo hacer a los estudiantes es la siguiente: ¿usted tiene el mismo comportamiento en su casa que cuando está con sus amigos? Luego de algunas cómplices risas, normalmente la respuesta es que no y en algunos casos empiezan a señalarse entre ellos con mayor o menor sentido de culpabilidad.

¿Por qué sucede esto? Pues por la influencia de los otros, el comportamiento de una persona a veces puede cambiar tanto entre un grupo y otro que parecería otra persona; solo piense en aquellas personas que al tomar licor cambian radicalmente; la variable externa influyente en el comportamiento de esa persona hace verle muy distinta al "original".

Para el caso de este ejemplo puede también haber controversia respecto a la cuarta variable (edad), analicemos: mientras mayor IMC (índice de masa corporal, x_1) supongo estará conmigo en que se esperaría mayor nivel de colesterol, mientras mayor Sedentarismo (x_2) es obvio pensar que el nivel de colesterol también subirá, a mayor cantidad de Actividad física (x_3) el nivel de colesterol disminuirá y eso nos dice el signo de esta variable, en cuanto a la edad dependerá del criterio de cada uno, unos dirán que sí y otros que no hay multicolinealidad (¿qué opina usted?).

A continuación, en la tabla 82, se presenta el cuadro de correlación entre las variables:

	IMC	Sedentarismo	Actividad física semanal	Edad	Colesterolemia
IMC	1				
Sedentarismo	0,63218573	1			
Actividad física semanal	0,12114242	-0,291221801	1		
Edad	0,44683047	0,212545509	-0,00642954	1	
Colesterolemia	0,57216226	0,781112824	-0,53091229	0,39316463	1

TABLA 82. COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES ESTUDIADAS

Según los resultados presentados, el valor de la variable que mejor se relaciona no coincide con lo indicado en base a los coeficientes ya que el Sedentarismo (x_2) es la variable más “fuerte”, en lo que sí coincide es en que la Edad (x_4) es la de menor relación.

En cuanto al tipo de relación y tomando en cuenta el signo de “r” de cada variable con la variable dependiente, se determina que **no** existe multicolinealidad (salvo el criterio particular respecto a la variable Edad).

En cuanto al análisis descriptivo (tabla 83) se presenta el resumen de los valores encontrados para cada variable.

	IMC	Sedentarismo	Actividad física semanal	Edad	Colesterolemia
Media	30,977	30,086	7,349	44,286	185,486
Error típico	0,503	1,228	0,517	0,931	3,787
Mediana	31,7	31	7	45	180
Moda	33,1	27	9	44	170
Desviación estándar	2,973	7,265	3,062	5,507	22,402
Varianza de la muestra	8,838	52,787	9,373	30,328	501,845
Curtosis	0,332	-0,723	-1,118	-0,567	-0,801
Coefficiente de asimetría	-0,751	-0,009	-0,077	-0,395	0,287
Rango	12,6	28	10	20	81
Mínimo	23	16	2	33	146
Máximo	35,6	44	12	53	227
Suma	1084,2	1053	257,2	1550	6492
Cuenta	35	35	35	35	35

TABLA 83. RESUMEN ESTADÍSTICOS DESCRIPTIVOS

El hecho de que las variables se distribuyan simétricamente (distribución normal) no significa necesariamente que sea una buena distribución, por ejemplo según la información dada en el caso del índice de masa corporal, un valor entre 30 y 35 indica ya obesidad leve y en esos valores se encuentra el promedio, además si tomáramos en cuenta el signo y valor del coeficiente de asimetría la distribución tendrá un sesgo negativo, lo cual es bastante malo porque indica que la tendencia en esta variable es hacia valores altos (lo cual por el valor de la moda podíamos suponer).

En el caso de la edad se podría decir que los valores encontrados de colesterolemia son acordes a lo que se espera, ya que los niveles de colesterol no son altos y eso se debería esperar en ese rango de edad.

Por último, ya que el modelo encontrado no es predictivo, no se puede resolver la última parte solicitada en el ejemplo.

Para hacer una comparación, y tal vez por curiosidad, voy a encontrar un nuevo modelo eliminando la variable edad (x_4), ya que según la ecuación inicial es la variable que presenta el menor valor de correlación con la variable dependiente; en la tabla 84 se presenta la matriz con los resultados:

x3	x2	x1	a
-3,30506381	1,30077749	2,71406615	86,5643771
0,71441059	0,3856735	0,90837129	20,8596702
0,77541657	11,1181476		

TABLA 84. ESTADÍGRAFOS DEL NUEVO MODELO

Comparando los resultados los valores en general se mantienen, los estadígrafos que más variación presentan son los que corresponden al coeficiente de la primera variable y al valor de “a”; se mantiene que la variable x_3 es la que presenta mejor relación y el valor de “ r^2 ” subió tres puntos. Esto puede interpretarse también diciendo que la variable edad no aporta mucho como elemento importante y que tal vez se puede buscar una variable de mejor impacto para los niveles de colesterolemia y tratar de encontrar un modelo nuevo de mejor ajuste aún.

Ejemplo 3

Se han obtenido datos que permitirían predecir el probable éxito de los estudiantes en programas de posgrado o de extensión. Si se demuestra que la predicción es buena, la ecuación podrá servir para seleccionar nuevos estudiantes prometedores. Las variables utilizadas son:

X_1 : Calificaciones en introducción a la Estadística / 100

X_2 : Calificaciones en introducción a la Psicología / 100

X_3 : Ansiedad / 10

X_4 : Aptitud para usar computadoras / 100

Y: Datos obtenidos en una prueba piloto que mide destrezas de estudio / 100

Y	X_1	X_2	X_3	X_4
97	89	78	5	42
92	86	74	2	40
90	91	77	5	29
85	76	80	3	36
84	87	81	4	32
82	87	63	5	29
81	73	76	4	31
80	75	74	4	41
77	70	68	5	39
75	67	69	3	35
75	74	73	4	33
74	75	54	3	32
73	73	62	2	36
73	72	64	2	30
71	73	53	4	32
70	70	74	7	30
70	73	66	8	35
70	69	62	5	33
70	72	57	6	40
67	73	68	4	30
66	71	60	6	27
66	68	64	3	25
63	68	61	3	18
61	73	65	6	11
61	72	56	5	24
60	68	55	7	30
59	69	54	6	22
57	66	60	3	15
54	67	56	6	16
50	69	56	6	15

Determine usted: ecuación, variable que tiene mayor y menor relación con la variable estudiada, variables que no sean consistentes, multicolinealidad, valor mínimo y máximo de la variable dependiente, si la ecuación es predictiva, interprete el error de estimación y el valor de “a”. Encuentre también la relación individual de cada variable con la variable dependiente y haga un análisis descriptivo de cada variable e indique sus conclusiones. Si la ecuación es predictiva, identifique cuán bien le iría a un candidato en la prueba piloto si obtuviese respectivamente en cada variable independiente los siguientes valores: 75, 60, 8 y 40 para las variables x_1 a x_4 respectivamente.

La matriz con los resultados del modelo buscado según los datos es la que se presenta en la tabla 85:

x_4	x_3	x_2	x_1	a
0,56793328	-1,08731282	0,35999875	0,76017513	-19,7865281
0,07235287	0,33518399	0,0784252	0,09315479	6,10034672
0,94769978	2,79617673			

TABLA 85. ESTADÍGRAFOS PARA EL EJEMPLO 3

Según los coeficientes encontrados en el modelo, las variables en orden de mejor a menor relación serán: x_3 , x_1 , x_4 y x_2 ; según los signos que indican el tipo de relación, no existe multicolinealidad; el signo de “a” es absurdo (algunas veces el valor del término independiente se ve afectado en cuanto al signo) por tanto deberá buscarse el menor valor de “y” en la muestra para indicar que ese será el mínimo esperado en la población, en este caso es 50.

Los valores de error de inclinación (Sb) en todos los casos son menores a sus respectivos coeficientes; el coeficiente de determinación “ r^2 ” indica una alta relación de las variables escogidas con la variable independiente y el valor del error de estimación es muy pequeño comparado con el parámetro de medida en la prueba piloto; todo esto nos indica entonces que la ecuación sí es predictiva, por tanto se podrá inferir el nivel de destreza de cualquier persona que pertenezca a la población.

En referencia a la correlación de cada una de las variables con la dependiente, en la tabla 86 se presentan los valores correspondientes.

	Y	X1	X2	X3	X4
Y	1				
X1	0,80853331	1			
X2	0,76236274	0,5778299	1		
X3	-0,33829771	-0,11784241	-0,20780794	1	
X4	0,75380844	0,40415051	0,45735884	-0,17981039	1

TABLA 86. COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES ESTUDIADAS

Según la tabla anterior, el orden de mejor a menor relación que tienen las variables independientes hacia la dependiente es: x_1 , x_2 , x_4 y x_3 .

En cuanto al análisis descriptivo, en la tabla 87 se presentan los resultados numéricos correspondientes a cada variable.

	Y	X1	X2	X3	X4
Media	71,767	73,867	65,333	4,533	29,600
Error típico	2,073	1,271	1,567	0,291	1,509
Mediana	70,5	72,5	64	4,5	30,5
Moda	70	73	74	5	30
Desviación estándar	11,352	6,962	8,584	1,592	8,265
Varianza de la muestra	128,875	48,464	73,678	2,533	68,317
Curtosis	-0,238	0,887	-1,136	-0,669	-0,203
Coefficiente de asimetría	0,269	1,377	0,289	0,182	-0,657
Rango	47	25	28	6	31
Mínimo	50	66	53	2	11
Máximo	97	91	81	8	42
Suma	2153	2216	1960	136	888
Cuenta	30	30	30	30	30

TABLA 87. RESUMEN ESTADIGRAFOS DESCRIPTIVOS

Según esto, se puede decir lo siguiente:

1. Según los valores de las medidas de tendencia central (MTC), las variables x_1 , x_3 y x_4 presentan una distribución normal; sobre esto hay que decir también que siendo el tipo de distribución esperada además los valores medios no están muy lejos del máximo de cada variable. Este tipo de observación es importante hacerla ya que aunque una variable se distribuya de forma simétrica, sus valores pueden estar lejos del máximo y eso realmente no sería bueno.
2. De acuerdo al valor del coeficiente de asimetría, las variables x_2 , x_3 y x_4 son las que se distribuyen normalmente al igual que la variable dependiente.
3. Según las MTC, las variables x_1 y x_3 están cerca del percentil 75, esto implica que las calificaciones del grupo tienden a ser altas.
4. La variable x_4 indica un grave problema en cuanto al manejo de computadoras.
5. Todas las variables presentan alta dispersión, por tanto, se recomienda formar grupos para realizar un análisis más específico según casos más cercanos.
6. La variable dependiente presenta una distribución normal y con valores altos en MTC, por tanto la tendencia del grupo es alta en lo que se refiere a destrezas de estudio.

En referencia al cálculo solicitado para un candidato que haya obtenido los siguientes valores: 75, 60, 8 y 40 para las variables x_1 a x_4 respectivamente, el resultado como predicción puntual será: 72.84 y como intervalo [70.04 ; 75.64].

EJERCICIOS PROPUESTOS PARA EL CAPÍTULO

Para los ejercicios propuestos a continuación, encuentre lo siguiente:

- i. La ecuación que relacione todas las variables con la variable dependiente
- ii. Los valores del error de inclinación de los coeficientes de cada variable independiente en el modelo general
- iii. El valor del coeficiente de determinación en el modelo general
- iv. El valor del error estándar de estimación en el modelo general
- v. Interprete el valor de “a”
- vi. Indique cuál es el orden desde mayor a menor relación de las variables con la variable dependiente, utilizando los coeficientes de la ecuación y encontrando los coeficientes de correlación. Compare los resultados.
- vii. Indique si la ecuación es predictiva
- viii. Determine si existe o no multicolinealidad
- ix. Estime el valor para la variable dependiente para algunos casos que usted crea se puedan dar
- x. Haga un análisis descriptivo de todas las variables

1. Para determinar el nivel de alfabetización en un grupo de personas, se han aplicado algunas pruebas que tienen relación con este tema. Los resultados se encuentran en la tabla siguiente.

Nivel de alfabetización inicial / 20	Conciencia de lo impreso. Identificación de la simbología gráfica / 10	Conciencia fonológica. Toma de conciencia y capacidad de manejo de fonemas / 20	Conocimiento del alfabeto. Identificar las letras (símbolos) del alfabeto por medio de tarjetas / 30	Lectura. Reconocimiento visual de palabras y frases simples / 15	Escritura. Se evalúa que sepa escribir el nombre, dictados palabras simples y otras / 30
15	10	10	25	7	22
13	7	15	21	7	13
7	7	2	6	6	14
11	6	15	13	5	18
10	9	13	6	3	21
10	10	10	8	5	18
9	6	14	11	3	11
13	9	13	24	6	12
15	10	19	25	7	12
10	3	13	12	5	18
8	8	9	11	5	7
9	6	10	12	5	11
13	6	16	22	8	11
19	8	17	27	11	30
8	7	11	10	1	11
18	10	18	27	12	22
17	8	14	26	10	27
18	9	15	27	11	29
13	7	16	14	8	18
19	10	17	27	12	30

2. Las autoridades de una institución educativa de nivel medio están preocupadas porque han detectado una sensible baja en la capacidad de socialización de algunos de los estudiantes especialmente de segundo y tercero de bachillerato. El departamento de orientación hizo entonces una investigación para la cual aplicó cuatro pruebas de temas que a su parecer creía se relacionaban con la sociabilidad de las personas y otra que medía la capacidad de socializar de los estudiantes. Los resultados se presentan a continuación.

Sociabilidad / 100	Radio de confianza. Sentimiento entrañable que una persona obtiene de la confianza en uno mismo y de sentir que los demás confían en esa persona / 33	Conexiones interpersonales. La inteligencia interpersonal se construye a partir de una capacidad nuclear para sentir distinciones entre los demás, en particular contrastes en sus estados de ánimo, temperamentos, motivaciones / 30	Conciencia emocional de uno mismo. Conocimiento de nuestras propias emociones y cómo nos afectan / 30	Atención a Normas y Valores. Capacidad del individuo a aceptar conductas / 30
70	20	16	25	19
81	23	25	13	20
60	16	17	18	15
90	26	30	30	13
75	22	24	22	16
48	16	18	21	23
95	28	30	30	19
46	20	14	13	16
80	23	27	23	28
80	22	29	21	18
59	21	19	15	13
79	22	25	18	24
62	19	23	13	19
56	17	23	18	16
76	21	25	18	15
66	26	23	13	22
79	28	21	18	14
41	19	15	12	20
51	14	23	16	15
78	22	25	22	9

3. En un proyecto de desarrollo comunitario fue necesario determinar las cualidades de ciertas personas para establecer su capacidad de liderazgo. El cuadro siguiente presenta

la valoración a una de ellas dada por 17 personas. Indicar si existe o no relación de estas variables con el objetivo de estudio y contestar las preguntas planteadas.

Planificación / 5	Seguimiento /5	Iniciativa /5	Solución de Problemas /5	Desarrollo del Equipo /5	Capacidad Motivacional /5	Desarrollo de Personas /5	Autocontrol /5	Receptividad /5	Confianza en el líder /50
4	5	2	4	3,5	4	2,5	3	3	39,5
4	3	4	4	4	5	4	3	5	44,5
4	4	5	4	4	4	4	5	3	45,5
4	4	4	4	5	4	4	5	4	43,5
4,5	5	4	4	4,5	5	4	5	5	47,5
4	4	2	3	4	3	5	5	5	42,5
3,5	4	4	4	4	4	4	5	4	41,5
3	3	3	3	2	2	2	4	4	33,5
4,5	3,5	4,5	4	4,5	4,5	5	4	4,5	46,5
5	5	4	4	3	3,5	4	4	5	46
4	4	4	4	4	5	4	5	5	48,5
5	5	5	5	5	5	5	4	5	49
4	3	4	4	2	3	4	3	4	39,5
2	3	5	4	2	3	2	4,5	3	42,5
4	4	4	4	4	4	4	4	4	44
4	5	5	3	2	2	2	2	3	39
2	3	3	3	2	2	4	1	2	36,5

4. Se pretende conocer si las variables estudiadas (sociabilidad, adaptación al medio, edad, motivación) tienen o no relación con la variable capacidad en formación de grupos; para ello se pide realizar un análisis general de la matriz dada.

Interés en formación de grupos / 35	edad	adaptación al medio / 35	sociabilidad / 40	motivación / 40
22	23	21	16	16
33	23	34	17	30
18	30	24	23	15
30	17	30	30	24
20	17	35	16	27
28	28	35	25	24
33	22	17	29	34
21	31	33	17	27
26	23	16	34	31
30	21	24	34	28
22	19	26	17	24
18	29	33	28	28
31	35	16	35	15
20	19	29	24	19
17	20	21	24	31

5. En una institución educativa de nivel medio, se desea determinar si el rendimiento escolar estará o no influenciado por las tres variables escogidas, si la respuesta es positiva la institución tendrá en cuenta estas variables para próximos estudios. Los resultados se encuentran en la tabla siguiente

SOCIABILIDAD <25 DESADAPTACION; 25 - 52 NORMALIDAD; > 52 SUPERADAPTACION	MOTIVACION 10 - 20 NULA; 21 - 33 BAJA; 34 - 42 MEDIA 44 - 50 ALTA	EPISODIOS DEPRESIVOS <36 NO EXISTE DEPRESION; 36 - 39 LEVE; 40-47 MODERADA; 48-55 MARCADA > 56 GRAVE	RENDIMIENTO ESCOLAR / 10
35	50	34	9,5
40	44	39	8
37	50	33	7,5
20	32	54	7,25
30	40	38	9,5
38	35	31	10
34	40	33	8
34	40	38	8
44	42	34	10
30	38	44	7,25
45	48	33	10
33	40	43	9,25
20	34	50	6,75
30	40	42	9,5
40	44	38	5,75
38	42	39	8,25
29	32	40	8,25
30	38	41	9,5
34	34	30	10
42	34	35	9,25

6. En una institución de educación media han decidido contratar a 5 personas que estén capacitadas para resolver problemas, de cualquier índole, que se presenten con los estudiantes. En el proceso de selección se tomaron varias pruebas que miden capacidades específicas relacionadas con el objetivo planteado. Identifique si estas variables son predictoras o no de la variable dependiente; sea cual fuere el caso, sugiera a las autoridades de la institución cuáles serían las personas que pueden ayudar a alcanzar el objetivo propuesto.

Resolución de problemas / 25	Habilidades sociales / 15	Creatividad / 20	Habilidades afectivas / 15	Planificación y capacidad de abstracción / 25
16	3	7	10	5
21	2	8	15	14
16	10	14	15	10
7	3	5	14	8
12	5	6	15	7
22	3	12	14	10
16	8	19	14	3
12	10	13	11	9
8	6	14	14	14
7	10	5	8	19
16	10	8	14	9
25	8	7	11	18
23	4	16	12	12
25	6	9	8	3
5	8	7	13	18
25	4	19	11	14
17	3	14	11	2
9	8	5	11	7
21	9	16	9	18
22	10	20	14	14
5	9	16	9	18
5	3	7	8	15
7	10	11	12	18
23	10	12	11	20
16	5	6	15	4
7	6	15	15	10
21	10	14	11	19
8	6	12	10	13
5	6	7	12	9
6	3	13	13	16
14	2	19	12	2
6	9	10	14	10
14	6	13	14	9
13	10	12	11	8
11	9	11	10	15
14	8	9	15	17
20	6	5	9	11
5	4	5	8	20
9	5	9	10	19
13	5	5	13	11

7. La directora del nivel inicial de una institución educativa, quiere desarrollar un modelo para predecir el tiempo que se demorarían los niños entre 7 y 12 años para armar un laberinto en base a las horas de entrenamiento. Se seleccionó una muestra de 80 niños y los resultados fueron los siguientes.

EDAD	HORAS	MINUTOS	EDAD	HORAS	MINUTOS	EDAD	HORAS	MINUTOS	EDAD	HORAS	MINUTOS
12	27	19	11	23	16	10	24	20	10	25	14
8	24	20	9	18	23	8	15	30	12	30	13
8	12	30	11	20	19	10	17	26	9	18	20
12	22	20	8	15	27	11	25	16	12	25	16
7	13	29	7	13	28	9	20	19	11	24	16
10	29	18	10	20	19	11	13	27	11	33	10
7	14	27	11	18	22	7	16	25	12	45	8
10	20	21	12	25	16	9	10	30	9	15	13
8	16	26	8	15	27	10	15	26	7	12	31
9	21	22	12	24	17	12	30	12	8	10	31
9	10	31	10	15	27	10	35	8	7	12	30
8	15	26	10	16	25	12	24	17	9	8	32
9	18	24	8	10	29	12	20	18	8	5	32
12	36	10	7	14	28	9	20	19	7	10	29
12	30	17	12	25	15	8	15	13	12	25	18
11	20	22	11	30	12	9	18	21	11	20	19
9	15	26	8	8	31	7	8	31	12	24	17
11	20	21	7	12	16	8	14	15	12	10	30
8	10	30	7	14	27	11	22	18	8	15	12
9	15	27	7	12	29	8	10	30	11	10	29

8. El siguiente grupo de datos se ha obtenido para analizar la capacidad administrativa del personal docente y administrativo con el fin de conocer la posibilidad de creación de nuevos cargos que requieran determinadas habilidades, se tomó una muestra a 17 trabajadores. ¿Qué variables, luego del análisis, serían las más adecuadas tomar en cuenta para esto?

Para esto, debe realizar:

- Análisis descriptivo de cada variable
- Análisis de regresión simple de cada variable independiente con la variable dependiente
- Análisis total de regresión multivariable
- Exponga conclusiones y recomendaciones

NOTA: todas las variables independientes se miden sobre 5 puntos y la variable dependiente sobre 50 pts. Un valor inferior a 37.5 se considera de baja capacidad administrativa

Trabajador	Planificación	Seguimiento	Iniciativa	Solución de Problemas	desarrollo del Equipo	Capacidad Motivacional	Desarrollo de personas	Auto control	Receptividad	Capacidad administrativa
1	4	5	2	4	3.5	4	2.5	3	3	39.5
2	4	3	4	4	4	5	4	3	5	44.5
3	4	4	5	4	4	4	4	5	3	45.5
4	4	4	4	4	5	4	4	5	4	43.5
5	4.5	5	4	4	4.5	5	4	5	5	47.5
6	4	4	2	3	4	3	5	5	5	42.5
7	3.5	4	4	4	4	4	4	5	4	41.5
8	3	3	3	3	2	2	2	4	4	33.5
9	4.5	3.5	4.5	4	4.5	4.5	5	4	4.5	46.5
10	5	5	4	4	3	3.5	4	4	5	46
11	4	4	4	4	4	5	4	5	5	48.5
12	5	5	5	5	5	5	5	4	5	49
13	4	3	4	4	2	3	4	3	4	39.5
14	2	3	5	4	2	3	2	4.5	3	42.5
15	4	4	4	4	4	4	4	4	4	44
16	4	5	5	3	2	2	2	2	3	39
17	2	3	3	3	2	2	4	1	2	36.5

SOLUCIÓN EJERCICIOS IMPARES

Ejercicio 1

En la tabla 88 se presentan los resultados de la ecuación general, los valores de “Sb” de cada variable independiente, “Se” y “r²” del modelo.

x5	x4	x3	x2	x1	a
0,16921228	0,23989567	0,217141	0,20021967	0,1908862	0,09795592
0,00984121	0,03278963	0,01249527	0,01639715	0,0294993	0,26156024
0,99788554	0,21063211				

TABLA 88. ESTADÍSTICOS EJERCICIO 1

Ecuación: $Y = 0.1908x_1 + 0.2002x_2 + 0.2171x_3 + 0.2398x_4 + 0.1692x_5 + 0.0979$

Por los signos de los coeficientes de las variables, no existe multicolinealidad.

Los valores de “Sb” de cada coeficiente están resaltados en verde, según estos resultados todas las variables son consistentes.

El valor de “r²” está resaltado en azul e indica una relación casi perfecta, esto significa que las variables escogidas para analizar el nivel de alfabetización son precisas.

De acuerdo al valor de “Se” resaltado en mostaza, se puede decir que cuando se realice alguna predicción sobre la variable dependiente en la población estudiada, el error de estimación será muy pequeño y por tanto la predicción muy precisa.

Todo lo indicado anteriormente, implica que la ecuación sí es predictiva y con un error de pronóstico mínimo (0.2106).

Como ejemplo de predicción se propone lo siguiente: supongamos que un individuo de la población ha obtenido los siguientes resultados en las pruebas de las distintas variables:

Conciencia de lo impreso: 7

Conciencia fonológica: 16

Conocimiento del alfabeto: 27

Lectura: 12

Escritura: 28

Al aplicar estos valores en la ecuación, tendríamos lo siguiente:

$$Y = 0.1908 * 7 + 0.2002 * 16 + 0.2171 * 27 + 0.2398 * 15 + 0.1692 * 28 + 0.0979$$

Y el resultado será: $Y = 16.911$, esta sería entonces la estimación puntual, aplicando el valor de “Se”, los valores de estimación de intervalo en cuanto al nivel de alfabetización serán: $Y [16.700 ; 17.122]$. Fíjese que el intervalo de predicción es muy cercano al puntual, por tanto, se confirma que los pronósticos para cualquier persona de la población serán muy certeros.

En cuanto al valor de “a” será difícil aceptarlo, aunque no estaría descartado para la población ya que eso significaría que puede existir alguna persona analfabeta (¡nivel de alfabetización de 0.097!), pero como el nivel mínimo de alfabetización inicial según los datos

es 7, entonces se debe aceptar ese valor como el mínimo esperado y no el encontrado en la ecuación (0.0979).

Si se toma en cuenta los coeficientes de cada una de las variables independientes para determinar el orden de mayor a menor en el que éstas inciden en la variable dependiente, el resultado se presenta en la tabla 89.

x4	x3	x2	x1	x5
0,2398	0,2171	0,202	0,1908	0,1692

TABLA 89. ORDEN DE MAYOR A MENOR RELACIÓN DE LAS VARIABLES INDEPENDIENTES SOBRE LA VARIABLE DEPENDIENTE SEGÚN COEFICIENTES

Los valores de los coeficientes de correlación se obtienen con la función correspondiente que se encuentra en la pestaña de “Análisis de datos”; los resultados se presentan a continuación en la tabla 90.

	Nivel de alfabetización inicial	Conciencia de lo impreso	Conciencia fonológica	Conocimiento del alfabeto	Lectura	Escritura
Nivel de alfabetización inicial	1					
Conciencia de lo impreso	0,505341278	1				
Conciencia fonológica	0,69742617	0,16746374	1			
Conocimiento del alfabeto	0,922378866	0,43798631	0,63984375	1		
Lectura	0,89930116	0,40087356	0,49521516	0,815740111	1	
Escritura	0,773587732	0,36620077	0,35175422	0,524222274	0,70537478	1

TABLA 90. COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES ESTUDIADAS

Según el detalle en la tabla anterior, el orden de mayor a menor en el que las variables independientes se relacionan con la variable dependiente se establece en la tabla 91.

x3	x4	x5	x2	x1
0,9224	0,8993	0,7736	0,6974	0,5053

TABLA 91. ORDEN DE MAYOR A MENOR RELACIÓN DE LAS VARIABLES INDEPENDIENTES SOBRE LA VARIABLE DEPENDIENTE SEGÚN COEFICIENTES DE CORRELACIÓN

Como se puede comparar sí difiere el orden según el camino que se quiera seguir. La recomendación siempre será que para decidir se tome en cuenta los coeficientes de correlación y no los valores de los coeficientes de la ecuación; estos últimos determinarán la fuerza de relación solo para el caso en que por alguna razón no se puedan obtener los valores de “r”.

Por último, procedo a encontrar los valores descriptivos de todas las variables.

	<i>Nivel de alfabetización inicial. / 20</i>	<i>Conciencia de lo impreso / 10</i>	<i>Conciencia fonológica / 20</i>	<i>Conocimiento del alfabeto / 30</i>	<i>Lectura / 15</i>	<i>Escritura. / 30</i>
Media	12,75	7,8	13,35	17,7	6,85	17,75
Mediana	13	8	14	17,5	6,5	18
Moda	13	10	10	27	5	18
Desviación estándar	3,9320	1,8806	3,9105	7,9743	3,0997	7,1073
Coefficiente de asimetría	0,2882	-0,6815	-1,2135	-0,0883	0,2027	0,4996
Rango	12	7	17	21	11	23
Mínimo	7	3	2	6	1	7
Máximo	19	10	19	27	12	30
Cuenta	20	20	20	20	20	20

TABLA 92. VALORES DESCRIPTIVOS DE TODAS LAS VARIABLES

Según estos datos se puede decir lo siguiente:

- Si se toma en cuenta los valores de las medidas de tendencia central, salvo las variables “Conciencia de lo impreso” y “Conocimiento del alfabeto” todas presentan una distribución normal, este tipo de distribución es lo esperado, pero hay que anotar lo siguiente
 - Salvo en la variable “Conciencia de lo impreso” las demás presentan valores medios muy bajos comparados con el valor máximo sobre el que se mide cada una; es decir, aunque su distribución es normal, la “cresta” de la curva está más alejada de los valores máximos.
 - Las variables de menor puntaje comparando con sus respectivos parámetros son “Lectura” (menos de la mitad del máximo esperado) y “Conocimiento del alfabeto” y “Escritura”, todas con valores muy cercanos a la mitad del máximo esperado; esto no es muy bueno para lo que se debería esperar.
 - El valor de la moda en la variable “Conocimiento del alfabeto” no es muy representativo ya que solo 4 sujetos de 20 presentan este valor.
 - Según los resultados de los coeficientes de asimetría, las variables “Conciencia de lo impreso” y “Conciencia fonológica” presentarían asimetría negativa, lo cual en términos de tendencia es bueno.
- Tomando en cuenta el valor de la desviación estándar, todas las variables presentan dispersión, se puede decir sin embargo que la más homogénea es la variable “Conciencia de lo impreso”; hay que tomar en cuenta también que, según el rango, en las variables “Conciencia de lo impreso”, “Conciencia fonológica” y “Escritura” los valores mínimos encontrados son atípicos e influenciarán en la decisión en cuanto a la dispersión.

Ejercicio 3

En la tabla 93 se presentan los resultados de la ecuación general *, los valores de “Sb” de cada variable independiente, “Se” y “r²” del modelo.

x9	x8	x7	x6	x5	x4	x3	x2	x1	a
0.42525	0.68387	1.92670	3.12034	-1.38111	-1.07756	1.58676	2.03276	-0.96056	18.13935
0.95035	0.57427	0.74295	1.20065	1.12609	1.64932	0.56456	0.92563	1.24012	4.71515
0.92709	1.74728								

TABLA 93. ESTADÍGRAFOS EJERCICIO 3

* Por un tema de espacio, los resultados se redujeron a 5 decimales

Ecuación:

$$Y = -0.960x_1 + 2.032x_2 + 1.586x_3 - 1.077x_4 - 1.381x_5 + 3.120x_6 + 1.926x_7 + 0.683x_8 + 0.425x_9 + 18.139$$

Por los signos de los coeficientes de las variables, sí existe multicolinealidad en las variables 1, 4 y 5; es decir estas variables se han visto afectadas en el proceso de “mezcla”. Sería importante realizar el análisis individual de estas variables para saber qué comportamiento (tipo de relación) tienen con la variable dependiente.

Los valores de “Sb” de cada coeficiente están resaltados en verde, según estos resultados existen tres variables (1, 4 y 9) que no son consistentes. Fíjese que las variables 1 y 4 se repiten, sería importante hacer un análisis especial con estas variables.

El valor de “r²” está resaltado en azul e indica un nivel de relación muy bueno, esto significa que, para analizar el nivel de alfabetización, están bien escogidas las variables.

De acuerdo al valor de “Se” resaltado en mostaza, se puede decir que cuando se realice alguna predicción sobre la variable dependiente en la población estudiada, el error será aceptable tomando en cuenta que representa solo un 3% respecto al valor sobre el que se mide la variable dependiente.

En general se puede decir que la ecuación sí es predictiva. Al igual que en el ejercicio anterior, puede remplazar valores para las variables independientes y obtener una estimación de la confianza de liderazgo; recuerde que los valores a remplazar deben estar dentro del intervalo entre menor y mayor de los datos conocidos en la muestra.

En cuanto al valor de “a” (18.139) este indica por concepto el puntaje mínimo esperado en cuanto a la confianza en el líder, pero dado que este valor está muy lejano al valor mínimo de la muestra, se sugeriría que se tome el puntaje de 33.5 como el valor mínimo a esperar en esta población; pero de todas maneras es un valor lógico dado que la variable se mide sobre 50 y alguien sí puede obtener un valor muy bajo como el de la ecuación.

Si se toma en cuenta los coeficientes de cada una de las variables independientes para determinar el orden de mayor a menor en el que éstas inciden en la variable dependiente, el resultado se presenta en la tabla 94.

x6	x2	x7	x3	x5	x4	x1	x8	x9
3,12	2,032	1,926	1,586	-1,381	-1,077	-0,96	0,683	0,425

TABLA 94. ORDEN DE MAYOR A MENOR RELACIÓN DE LAS VARIABLES INDEPENDIENTES SOBRE LA VARIABLE DEPENDIENTE

Los valores de los coeficientes de correlación de todas las variables se presentan a continuación en la tabla 95.

	Planificación	Seguimiento	Iniciativa	Solución de Problemas	Desarrollo del Equipo	Capacidad Motivacional	Desarrollo de Personas	Autocontrol	Receptividad	Confianza en el líder
Planificación	1									
Seguimiento	0,66	1,00								
Iniciativa	0,16	0,06	1,00							
Solución de Problemas	0,50	0,28	0,47	1,00						
Desarrollo del Equipo	0,62	0,39	0,09	0,59	1,00					
Capacidad Motivacional	0,59	0,30	0,24	0,79	0,84	1,00				
Desarrollo de Personas	0,51	0,07	0,01	0,38	0,68	0,54	1,00			
Autocontrol	0,30	0,16	0,13	0,38	0,59	0,50	0,27	1,00		
Receptividad	0,69	0,23	0,02	0,39	0,57	0,62	0,54	0,54	1,00	
Confianza en el líder	0,64	0,41	0,46	0,72	0,75	0,85	0,61	0,54	0,62	1,00

TABLA 95. COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES ESTUDIADAS

Según el detalle en la tabla anterior, el orden de mayor a menor en el que las variables independientes se relacionan con la variable dependiente se establece en la tabla 96.

x6	x5	x4	x1	x9	x7	x8	x3	x2
0,85	0,75	0,72	0,64	0,62	0,61	0,54	0,46	0,41

TABLA 96. ORDEN DE MAYOR A MENOR RELACIÓN DE LAS VARIABLES INDEPENDIENTES SOBRE LA VARIABLE DEPENDIENTE SEGÚN LOS COEFICIENTES DE CORRELACIÓN

Como se puede comparar, el orden según el camino que se quiera seguir (coeficientes de la ecuación o coeficientes de correlación), es muy distinto y como siempre la recomendación será la segunda opción.

Para este ejercicio voy a proceder a encontrar los valores de “r” y “r² de las variables 1 y 4 debido a lo dicho al principio de este análisis.

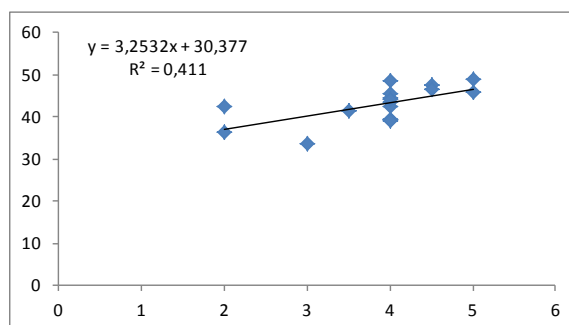


FIGURA 197: GRÁFICO DE RELACIÓN DE LA VARIABLE 1 CON LA VARIABLE DEPENDIENTE

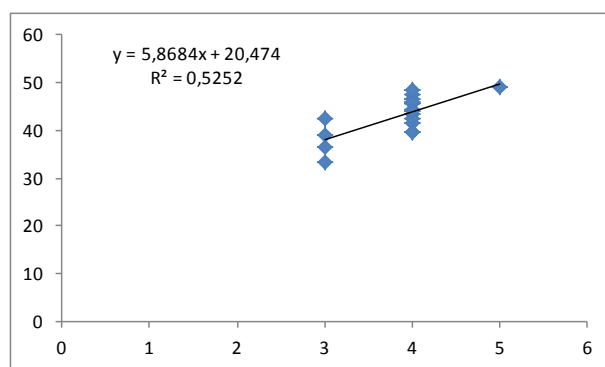


FIGURA 198: GRÁFICO DE RELACIÓN DE LA VARIABLE 4 CON LA VARIABLE DEPENDIENTE

Al observar para cada caso, el valor del coeficiente de “x” es positivo; esto indica que las variables cambiaron su “comportamiento” en el proceso de unión con las demás. El valor de “r²” es bastante bueno según lo dicho en la teoría referente a la correlación simple. Todo esto nos hace ver que no debe preocuparnos ni la multicolinealidad ni la inconsistencia de estas variables hallada en la ecuación general multivariable.

Los resultados de las variables en cuanto a las medidas descriptivas, se presentan en la tabla 97.

	Planifi cación	Segui miento	Inicia tiva	Solución de Problemas	Desarrollo del Equipo	Capacidad Motivacional	Desarrollo de Personas	Auto control	Recepti vidad	Confianza en el líder
Media	3,853	3,971	3,912	3,824	3,500	3,706	3,735	3,912	4,029	42,912
Mediana	4	4	4	4	4	4	4	4	4	43,5
Moda	4	4	4	4	4	4	4	5	5	39,5
Desviación estándar	0,843	0,800	0,939	0,529	1,104	1,062	1,002	1,176	0,943	4,280
Coefficiente de asimetría	-1,157	0,108	-0,865	-0,259	-0,395	-0,394	-0,797	-1,134	-0,602	-0,582
Rango	3	2	3	2	3	3	3	4	3	15,5
Mínimo	2	3	2	3	2	2	2	1	2	33,5
Máximo	5	5	5	5	5	5	5	5	5	49
Cuenta	17	17	17	17	17	17	17	17	17	17

TABLA 97. VALORES DESCRIPTIVOS DE TODAS LAS VARIABLES ANALIZADAS

Dado que todas las variables independientes se miden sobre 5 puntos, los valores de las medidas de tendencia central indican tener un criterio de alto desempeño del líder respecto en todas las competencias medidas, esto parece repercutir por tanto en la evaluación de la variable dependiente (medida sobre 50 puntos) y confirma entonces la alta correlación obtenida en el modelo general.

En cuanto a la dispersión todas las variables coinciden en el valor del rango salvo la variable “autocontrol”, pero si se revisan los valores simples de dicha variable, solo una persona obtuvo un valor de uno en dicha variable, es decir no es tendencia y por tanto se puede considerar igual que las demás.

En referencia a la asimetría, si bien es cierto que según las medidas de tendencia central todas presentan una distribución normal, las variables “planificación”, “seguimiento” y

“autocontrol” indicarían un sesgo negativo, que para efectos de la variable objetivo, este tipo de distribución revelaría una tendencia hacia valores más altos de estas variables y por tanto esta distribución se calificaría como buena; por ello se debe tomar en cuenta que, para quien ejerce el liderazgo, estas competencias pueden considerarse como su “fuerte”.

Ejercicio 5

En la tabla 98 se presentan los resultados de la ecuación general, los valores de “Sb” de cada variable independiente, “Se” y “r²” del modelo.

x3	x2	x1	a
-0,12125339	-0,05010162	0,00113843	15,194865
0,06739129	0,05685329	0,06620893	4,59736216
0,30035328	1,13841391		

TABLA 98. ESTADÍSTICOS EJERCICIO 5

$$\text{Ecuación: } Y = 0.00113 x_1 - 0.050 x_2 - 0.1212 x_3 + 15.194$$

Por los signos de los coeficientes de las variables, sí existe multicolinealidad en la segunda variable (motivación); al hacer el análisis individual de esta variable con la dependiente el resultado sí es lógico, es decir se presenta una relación directa según se aprecia a continuación en la figura 199.

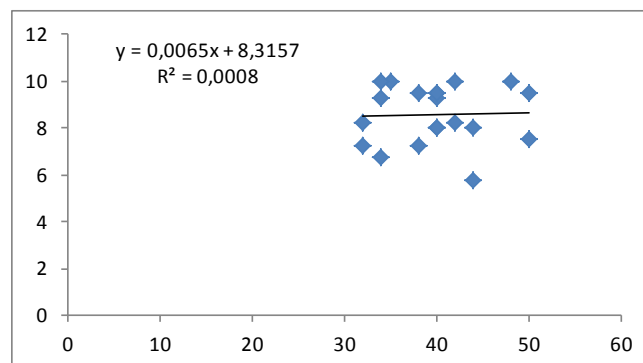


FIGURA 199: GRÁFICO DE RELACIÓN ENTRE LA VARIABLE MOTIVACIÓN Y RENDIMIENTO

Aunque según esto se puede decir que en verdad no existe relación (“r²” = 0.08%) pero el signo de “b” es positivo y eso indica que el comportamiento individual de esta variable (tipo de relación directa) sí responde a lo esperado.

Los valores de “Sb” de cada coeficiente están resaltados en verde, según estos resultados sólo la variable 3 es consistente; algo que ya se podía suponer dado que estos coeficientes están muy cerca de cero.

De acuerdo al valor de “Se” resaltado en mostaza, se puede ver que, aunque el valor numérico es de 1.13; en términos porcentuales supera el 10% respecto al valor sobre el cual se mide la variable y esto ya es un error significativo.

El valor de “r²” está resaltado en azul e indica un nivel de relación que no es aceptable en estudios de regresión multivariable, ya que se espera que supere un valor de 90%.

Con todo lo analizado anteriormente se deduce que las variables escogidas para determinar el rendimiento escolar en esta población no son las idóneas y se recomendaría utilizar otras o hacer un análisis respecto a que si los instrumentos utilizados para medir las variables independientes son idóneos para la población; de otro lado también hay que revisar si los valores de la variable rendimiento escolar realmente están reflejando esto, dado que según se observa, los resultados son muy altos y no es muy común que esto ocurra.

En general se puede decir entonces que la ecuación no es predictiva y por tanto no será posible inferir nada sobre el rendimiento escolar en la población estudiada.

En cuanto al valor de “a” (15.194) este valor debe considerarse como absurdo dado que la variable dependiente se mide sobre 10 puntos, esto aporta más al criterio que el modelo general y las variables no son confiables para analizar el rendimiento escolar.

En este caso ya no hace falta hacer más análisis debido a la inconsistencia del modelo.

Ejercicio 7

En este ejercicio hay que determinar con claridad las variables que intervienen, ante todo se debe decir que la variable dependiente **no son los minutos** y que una de las variables independientes **no son las horas**; esto aclaro porque suele confundirse dado que los datos están expresados como valores de tiempo.

Para este ejercicio se tiene lo siguiente:

Variables independientes: Edad y Entrenamiento en armar laberinto (medido en horas)

Variable dependiente: Armado de laberinto (medido en minutos).

El modelo matemático con los datos dados se presenta en la tabla 99.

x2	x1	a
-0,73537441	-0,10012918	36,5698524
0,07428968	0,31474201	2,3270486
0,71531055	3,59007905	

TABLA 99. ESTADÍGRAFOS EJERCICIO 7

Ecuación: $Y = -0.1001 x_1 - 0.7353 x_2 + 36.569$, donde x_1 es la edad y x_2 es el tiempo de entrenamiento de cada niño.

Por los signos de los coeficientes de las variables, no existe multicolinealidad dado que la lógica dirá que a mayor edad (x_1) y más tiempo de entrenamiento (x_2), se demorarán menos.

Los valores de “Sb” de cada coeficiente están resaltados en verde, según estos resultados solo la variable tiempo de entrenamiento es consistente.

El valor de “ r^2 ” está resaltado en azul e indica una relación baja como concepto de un proceso multivariable, pero solo son dos variables independientes, por tanto, se puede aceptar que un valor de r^2 equivalente al 71.53% no es tan malo tampoco.

De acuerdo al valor de “Se” resaltado en mostaza, este nos indica que, si se pronosticase el tiempo de resolución que un niño de esa población necesitaría para armar el laberinto, los cálculos tendrían un error de 3.59 minutos; que resulta ya ser un valor considerable, por tanto un error importante.

Aunque he expuesto razones que minimizan los problemas de relación con la variable dependiente, es muy riesgoso decir que el modelo es predictivo, por tanto, no se podrán realizar cálculos de inferencia.

En cuanto al valor de “a” habrá que suponer que es el tiempo máximo en el que se espera que un niño de la población se demore en armar el laberinto será de 36.56 minutos, valor que no está muy lejos del máximo dato encontrado en el variable dependiente por tanto debe aceptarse como legítimo.

Según los coeficientes de las dos variables independientes, la variable x_2 es la que mejor se relaciona con el tiempo de armar el laberinto y según los coeficientes de correlación ocurre lo mismo, según se observa en la tabla 100.

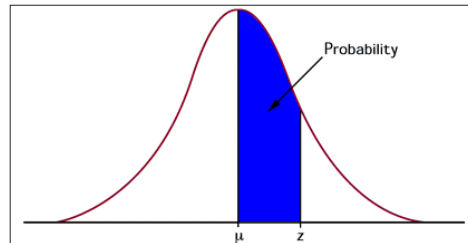
	Edad	Entrenamiento	Resolución laberinto
Edad	1		
Entrenamiento	0,68606263	1	
Resolución laberinto	-0,59416636	-0,845539097	1

TABLA 100. COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES ESTUDIADAS

BIBLIOGRAFÍA

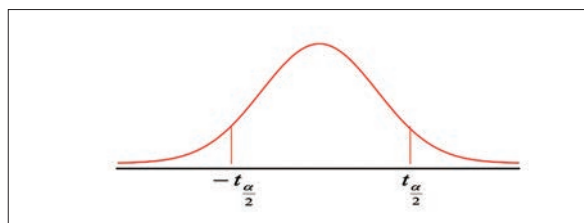
- Biblia Reina Valera 1960 (RVR 1960)*. (n.d.). American Bible Society.
- Botella, J., León, O., & San Martín, R. (1997). *Análisis de datos en Psicología I* (Cuarta). Madrid: Ediciones Pirámide S.A.
- Estadística para todos. (n.d.). Retrieved December 4, 2018, from http://www.estadistica-paratodos.es/historia/histo_esta.html
- Estadística y algo mas: Análisis de Regresión y Correlación. (n.d.). Retrieved October 16, 2019, from <http://estadisticayalgotomas.blogspot.com/2013/04/analisis-de-regresion-y-correlacion.html>
- Frases Célebres aplicadas a la Percepción. (n.d.). Retrieved December 3, 2018, from <https://es.slideshare.net/pbermudez10/frases-clebres-aplicadas-a-la-percepcin>
- frases sobres observacion - Buscar con Google. (n.d.). Retrieved December 3, 2018, from https://www.google.com.ec/search?tbm=isch&sa=1&ei=swuQW5SZMIjYzwLHuUw&q=frases+sobres+observacion&oq=frases+sobres+observacion&gs_l=img.3..35i39k1.1163667.1164240.0.1165561.2.2.0.0.0.130.249.0j2.2.0....0...1c.1.64.img.0.1.130....0.RyOk7QlGuLc#imgsrc=SI
- González Betanzos, F., Escoto Ponce de León, M., & Chávez López, J. (2017). *Estadística aplicada en Psicología y Ciencias de la salud* (Primera). México: Editorial El Manual Moderno.
- Guardia, J., Freixa, M., Pero, M., & Turbany, J. (2008). *Análisis de datos en Psicología*. (F. García, Ed.) (Segunda). Madrid: Delta Publicaciones.
- Noronha, E. (n.d.). Origen - Catedu. Retrieved December 5, 2018, from <https://studylib.es/doc/317522/origen---catedu>
- Pagano, R. (2011). *Estadística para las ciencias del comportamiento*. (O. Ramírez & C. Islas, Eds.) (Novena). México: CENGAGE Learning.
- Pérez Juste, R. (2012). *Estadística aplicada a las Ciencias Sociales*. (UNED, Ed.) (Primera). Madrid.
- Pérez Juste, R., García Llamas, J., Gil Pascual, J., & Galán González, A. (2009). *Estadística aplicada a la educación*. (A. Cañizal & M. Varela, Eds.). Madrid: Pearson Educación, S.A.
- Spiegel, M. (1970). *Estadística*. México: McGraw-Hill.

ANEXO 1



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0,1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0,2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0,3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0,4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0,5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0,6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0,7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0,8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0,9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1,0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1,1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1,2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1,3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1,4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1,5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1,6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1,7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1,8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1,9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2,0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2,1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2,2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2,3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2,4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2,5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2,6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2,7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2,8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2,9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3,0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3,1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3,2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3,3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3,4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3,5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998

ANEXO 2



gl	0,2	0,1	0,05	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,695	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
dist. Norm	1,282	1,645	1,960	2,326	2,576	3,291

ÍNDICE

INTRODUCCIÓN.....	7
Cómo se desarrolla el libro	7
Los números, la Estadística y la Psicología	9
Qué sí y qué no contiene este libro.....	11
La observación como elemento fundamental del análisis	13
¿Para qué sirve la Estadística en Ciencias Sociales y particularmente en Psicología o Educación?	15
¿En qué momento debe iniciarse el estudio de la Estadística en Psicología?	16
Herramientas a utilizar para el análisis de datos	16
Capítulo 1:	
ANTECEDENTES	18
La Estadística como ciencia	18
Breve historia de la Estadística	18
Definiciones de Estadística	20
División de la Estadística	21
Conceptos iniciales	22
Población y muestra	22
Población	23
Muestra	23
Muestreo	24
Problemas del muestreo	27
Escala de medida.....	28
Tipos de variables y clasificación	29
Tipos de variable	29
Capítulo 2:	
ESTADÍSTICA DESCRIPTIVA.....	31
Estadística descriptiva.....	32

Medidas de tendencia central.....	32
Media (promedio).....	33
Mediana.....	33
Moda	35
Relación entre las tres Medidas de Tendencia Central.....	36
Ejercicios de aplicación del capítulo.....	39
Cálculo de las Medidas de Tendencia Central en datos agrupados.....	46
Media	47
Mediana.....	47
Moda	48
Gráfico de resultados	51
Ejercicios propuestos para el capítulo.....	59
Solución ejercicios impares	63
Capítulo 3:	
MEDIDAS DE FORMA	69
Distribución de frecuencias	69
Asimetrías (sesgos)	70
Asimetría Negativa.....	70
Asimetría Positiva	71
Curtosis	74
Capítulo 4:	
MEDIDAS DE DISPERSIÓN O VARIABILIDAD.....	78
Ejercicios de aplicación del capítulo.....	90
Ejercicios propuestos para el capítulo.....	99
Solución ejercicios impares	105
Capítulo 5:	
MEDIDAS DE POSICIÓN	114
Las medidas de Posición	115
Rango percentil (Excel)	126
Ejercicios propuestos para el capítulo.....	128
Solución ejercicios impares	132

Capítulo 6:

CUATRO CONCEPTOS ADICIONALES	136
Primer tema. Histograma	136
Segundo tema. Estadística Descriptiva abreviada.....	143
Tercer tema. Media ponderada.....	145
Cuarto tema. Media acotada.....	148

Capítulo 7:

DISTRIBUCIÓN NORMAL	152
Características de la Distribución Normal	154
Ejercicios de aplicación del capítulo.....	159
Ejercicios propuestos para el capítulo.....	163
Solución ejercicios impares	165

Capítulo 8:

ESTADÍSTICA INFERENCIAL	172
Evaluación Previa de Proyectos (EPP)	173
Cómo trabaja este sistema	174

Capítulo 9:

RELACIÓN ENTRE VARIABLES	179
Regresión Simple	179
Regresión Lineal.....	180
Regresión Lineal Simple (dos variables)	180
Variables intervinientes	180
Tipos de relación entre las variables.....	181
Fuerza de relación entre las variables	188
Coeficiente de Correlación	189
Coeficiente de Determinación	191
Errores de Estimación (Se) y de Inclinación (Sb)	195
Cálculo del error de Estimación (Se)	196
Cálculo del error de Inclinación (Sb)	196
Elaboración de hipótesis.....	199
Valor de “B” y Niveles de Confiabilidad	200
Contraste entre el valor estándar y el específico	202
Cálculo de valores de estimación	204
¿Para qué sirve y cómo se utiliza la ecuación de regresión?	204
Ejercicios de aplicación del capítulo.....	205

Ejercicios propuestos para el capítulo.....	215
Solución ejercicios impares	218
Capítulo 10:	
REGRESIÓN LINEAL MULTIVARIABLE	
(tres o más variables)	223
El arte culinario como preámbulo ilustrativo a la regresión multivariable.....	223
Proceso del análisis multivariable	224
Proceso para encontrar el modelo y estadígrafos	225
Correlación de cada variable con la variable dependiente	230
Análisis del error de los coeficientes	232
Ajuste del modelo	233
Análisis del coeficiente de determinación r^2	234
Análisis del error estándar de estimación Se	235
Predicción con la ecuación multivariable	235
Ejercicios de aplicación del capítulo.....	237
Ejercicios propuestos para el capítulo.....	246
Solución ejercicios impares	251
BIBLIOGRAFÍA	260
ANEXO 1.....	261
ANEXO 2.....	262



Este libro se terminó de imprimir en el mes de mayo de 2021, bajo el sistema de evaluación de pares académicos (uno interno y otro externo a la PUCE) y mediante la modalidad de «doble ciego», que garantiza la confidencialidad de autores y de árbitros.



La intención de este libro es presentar un formato más amigable en el estudio de la Estadística si se compara con otros textos; además, he tratado de utilizar un lenguaje formal, pero con la intención de ser lo más sencillo de entender y evitando un nivel de desarrollo matemático muy elevado, ya que considero no necesario incursionar en este ámbito para lograr el objetivo de la presente obra.

Cualquier tratado sobre Estadística suele provocar dos tipos de reacción, la primera, y creo yo la más común, es de alejamiento y hasta de cierta aversión dado que es una materia tradicionalmente considerada difícil y, según muchas personas alejadas de la Matemática, poco útil para sus intereses.

La segunda, aunque menos común, es a la que pertenecemos algunos seguidores de esta ciencia, quienes la acogemos con cariño, entusiasmo y necesidad de incursionar en ella de tal manera que hasta nos atrevemos a escribir y proponer, en función de la visión que cada uno tenga, de tal manera que tratamos de dar un aporte para su aplicación en muchos y variados ámbitos del quehacer profesional y estudiantil.

¿Qué propongo con este libro? Pues no otra cosa que el estudio de la Estadística a nivel de grado con aplicación a las ciencias Psicológicas y de Educación; con una visión muy práctica en el estudio de los distintos temas a estudiar a nivel universitario, sin complicarlos con una Matemática avanzada pero tampoco dejando de tener un buen nivel de profundidad en el tratamiento de cada uno de los temas planteados aquí.

El autor



Pontificia Universidad
Católica del Ecuador

Publicaciones Centro de
PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR