

Estadística Descriptiva

eLibro





eLibro



Autor:

Andrés Gálvez-Sobral Aguilar
Maestría en Medición,
Evaluación e Investigación.

“Estadística Descriptiva”

Curso financiado y adaptado para eLibro por:
Asociación Rafael Ayau (ARA)
ISBN: 978-9929-687-50-9

© ARA, 2017
Guatemala, Centroamérica.
informacion@uomac.net

Este eLibro puede ser reproducido citando a la fuente.



Índice del Curso



Biografía del Autor	v
Introducción del Curso	vii
Objetivos del Curso	viii
UNIDAD I, Introducción a la estadística	10
I. La estadística en la historia	12
II. Áreas de la estadística	23
III. Los datos en estadística	34
IV. Evaluación general de procesos estadísticos	45
UNIDAD II, Comprensión y presentación de datos	57
V. Organización de datos numéricos	59
VI. Tipos de variables y tipo de escalas de medida	67
VII. Distribución de frecuencias	77
VIII. Representar tablas de datos	83
UNIDAD III, Interpretación de tablas y gráficas	90
IX. Representación de tablas de frecuencias	91

X. Representación de variables cualitativas	99
XI. Representación de variables cuantitativas	109
UNIDAD IV, Medidas	116
XII. Medidas de tendencia central	120
XIII. Medidas de posición	139
XIV. Medidas de variabilidad	162
UNIDAD V, La curva normal	175
XV. La curva normal como probabilidades	176
XVI. Usos de la curva normal	185
UNIDAD VI, Relaciones entre variables, la correlación	204
XVII. El diagrama de dispersión	208
XVIII. Coeficiente de correlación lineal	220
XIX. Resolución de problemas usando correlación	229
XX. Principios de la regresión lineal	238

<u>Evaluación Corta de Aprendizaje (ECA)</u>	249
<u>Bibliografía</u>	250
<u>Webliografía</u>	251
<u>Anexo</u>	253

Biografía del Autor



Andrés Gálvez-Sobral Aguilar

Nació en Ciudad de Guatemala en 1979, luego de estudiar Psicología Clínica en la Universidad Rafael Landívar estudió una maestría en Medición, Evaluación e Investigación en la Universidad del Valle de Guatemala. Trabajó en el Sistema Nacional de Evaluación e Investigación Educativa –SINEIE- y la Dirección General de Evaluación e Investigación Educativa -DIGEDUCA/MINEUDUC.

Actualmente dirige el Centro de Investigaciones Educativas de la Universidad del Valle de Guatemala en donde realiza investigación educativa en proyectos de diversas aplicaciones. Ha publicado investigaciones en revistas nacionales e internacionales en temas como Bullying, Implementación Curricular y Capacitación Docente.

Algunas de sus publicaciones:

- Características docentes que influyen en el rendimiento académico. Comparación en dos muestras y exploración multinivel.
- Development of a Dating Violence Assessment Tool for Late Adolescence Across Three Countries: The Violence in Adolescents' Dating Relationships Inventory,
- Variables relacionadas con las conductas agresivas y su efecto en el rendimiento académico. Evaluación Nacional a Graduandos 2012
- La evaluación nacional como evidencia de la implementación curricular en Guatemala. ¿Son útiles las pruebas nacionales para retroalimentar al Currículo Nacional Base?

Introducción del Curso



La estadística es una herramienta de alto impacto en la toma de decisiones, a pesar de las críticas que se hacen a la cuantificación y generalización numérica que hace la estadística por “convertir a las personas en números”, sus aportes siguen siendo sin lugar a dudas una fuente de información segura, confiable y comprobable que la mantiene cada día más vigente en el ambiente educativo. La Estadística Descriptiva enfatiza en la organización de datos numéricos, su capacidad de analizarlos e interpretarlos y la correcta manifestación gráfica de las mismas. Asimismo se enfoca en la habilidad de lectura de gráficas para la descripción de fenómenos sociales y educativos.

En la primera unidad se desarrolla una introducción a la estadística, su importancia y sus usos. La segunda unidad se enfoca en la adecuada comprensión de los datos: tipos de escalas, tipos de variables y tablas de datos. Además se desarrolla la capacidad de presentar la información en gráficas de acuerdo a sus características. La tercera unidad se dedica a la interpretación, la información a partir de tablas y gráficas. La cuarta unidad se destina al estudio de las medidas de tendencia central y de posición y variabilidad. La quinta unidad se dedica al estudio de la curva normal y las operaciones relacionadas. La sexta unidad se utiliza para estudiar el correcto uso de la correlación entre dos o más variables.

Objetivos del Curso



General

Desarrollar las destrezas necesarias para valorar el uso e interpretación de datos numéricos para la toma de decisiones, especialmente en el ámbito educativo utilizando la computadora como herramienta esencial para el análisis y representación estadística y tomando en cuenta su correcta aplicación en la resolución de problemas, la investigación y la administración.

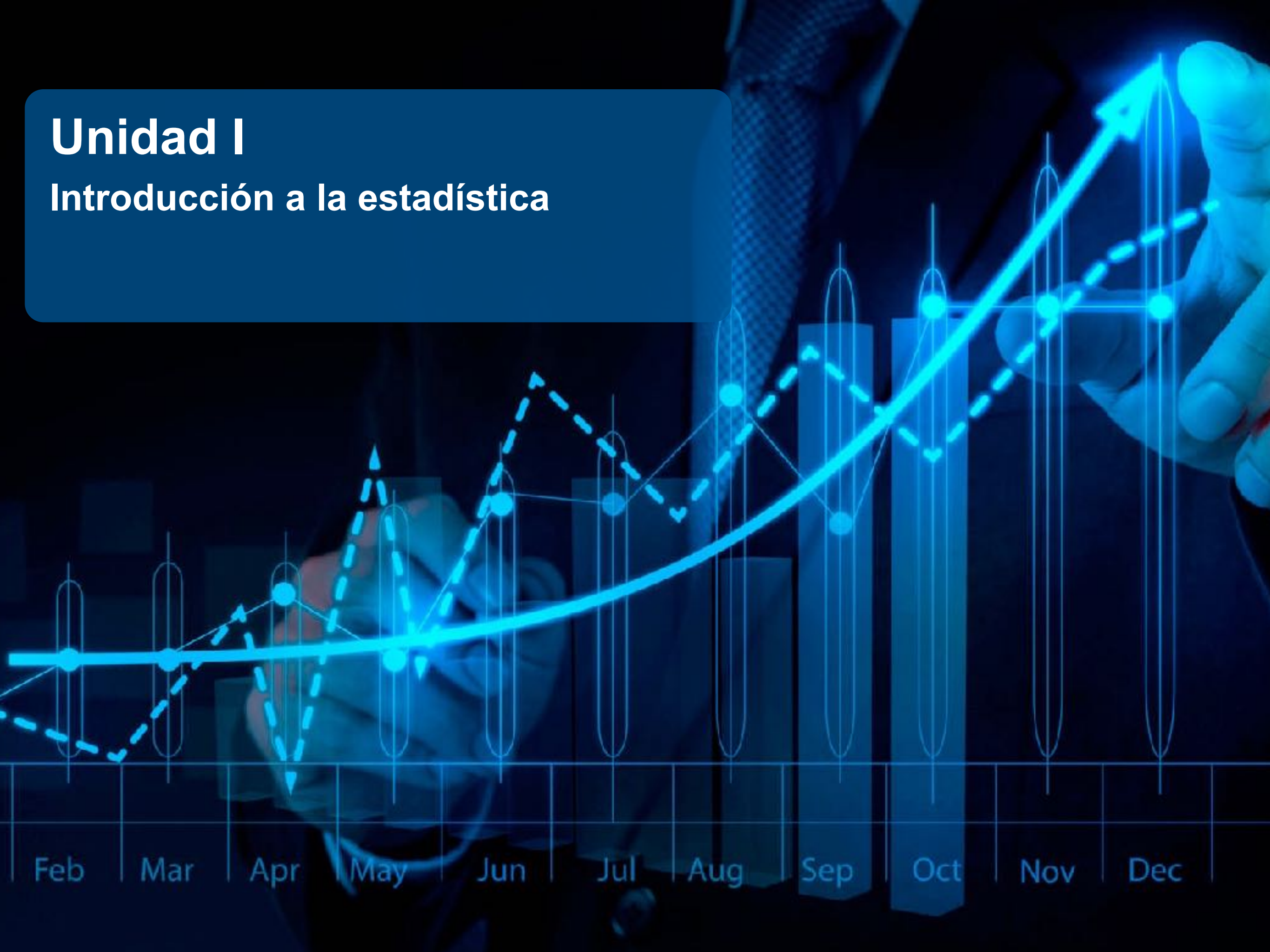
Específicos

- Comprender la importancia de la estadística en distintos ámbitos educativos y sus aspectos generales.
- Desarrollar la capacidad de identificar y seleccionar adecuadamente el tipo de variable y el tipo de gráfica para representarla.

- Utilizar las medidas de tendencia central, posición y variabilidad para toma de decisiones en ámbitos educativos.
- Lograr la correcta interpretación de gráficas y tablas de datos y el vocabulario adecuado para presentar la información proveniente de estas fuentes.
- Comprender el uso y aplicaciones de la curva normal especialmente en el ámbito educativo.
- Expresar la relación entre variables a través de la correlación.

Unidad I

Introducción a la estadística



En la primera unidad se desarrolla una introducción a la estadística, su importancia y sus usos.



Partiremos del origen de la estadística como un área de las matemáticas que poco a poco se fue especializando hasta llegar al punto de poder pensarse en ella como una ciencia por sí misma. Por consiguiente, su relación con las matemáticas es muy estrecha.

Podemos entonces decir que la estadística es la ciencia que se encarga de la recolección, organización y análisis de datos con el fin de tomar decisiones.

Capítulo I

La estadística en la historia



La estadística en la historia

Hace ya algunos años, estaba sentado en un aula universitaria como cualquier novato primerizo junto a otros que estaban igual que yo. Era mi primer día de clases universitarias y el curso era de Estadística 1. Recuerdo muy bien el ingreso del profesor hasta su escritorio, sacó un marcador y preguntó: ¿qué es la estadística?

Después de un silencio de varios segundos nos dijo: “La estadística es una herramienta para mentir científicamente”.

En este capítulo definiremos qué es la estadística e identificaremos los usos que tiene en el campo administrativo, en el ámbito del aula y en la vida cotidiana.

Para iniciar vamos a retomar la idea de la estadística como herramienta para “mentir científicamente” y pensemos un momento en un político que dice: de acuerdo a mis datos, el 75% de la gente dice que está a favor de mi

propuesta! Al escuchar un dato así con esa contundencia cuantitativa se hace más fácil creerle, porque las personas pueden mentir, pero los números no lo hacen.



Estadística

Sin embargo, ¿qué tan a menudo nos preguntamos, lo que significa que el 75% de la gente diga que está a favor? La realidad es que 75 de cada 100 dijeron estar a favor, pero también significa que 3 de cada 4 lo dijeron! Así pues, si este político fue a una reunión de allegados suyos y preguntó a 10 personas si estaban de acuerdo con su propuesta y 7 le

dijeron que sí, 2 que no y 1 que más o menos, entonces puede decir que el 75% está de acuerdo. En este caso, los datos son ciertos, pero no sabemos si representan la realidad.

Otro ejemplo: La economía del país X es muy buena, porque el ingreso promedio per cápita es de US\$500.00 al mes. Nuevamente el dato parece contundente, es fácil pensar que sería muy bueno vivir allí porque al conseguir trabajo tengo buena probabilidad de ganar US\$500.00. Sin embargo como aprenderemos más adelante, el promedio es una medida de tendencia sensible a los datos extremos, así pues en este caso puede ser que existan muchas personas que ganen menos de US\$500 y existan pocas personas que ganen bastante más que esa cifra y el resultado es un dato que no refleja la realidad de dicho país.

El término “estadística” tiene su origen en el vocablo “estado” ya que los gobernantes eran responsables de mantener registros de sus pobladores, en algún



momento de la historia se le conoció como “aritmética política” y era la ciencia del estado. Es hasta 1792 que Sir John Sinclair publica en Inglaterra el primer volumen de una serie titulada “Statistical Account of Scotland” y se empezó a considerar a dicha ciencia como la colección y clasificación de datos.



Sir John Sinclair

Antes de Sir John Sinclair es difícil determinar el origen real de las estadísticas, porque se utilizaban desde civilizaciones antiguas y se encuentra evidencia en pinturas rupestres, en murales y escritos antiguos. Alrededor del año 3000 a.C. en Babilonia se utilizaban tablillas de arcilla para recopilar datos; los Egipcios realizaban censos, registraban movimientos poblacionales y clasificaban sus rentas. El Antiguo Testamento menciona que Moisés recibió la orden de realizar un censo de las tribus y sus linajes. En China se hacían conteos poblacionales alrededor de 2000 a.C., y los Romanos eran también fehacientes organizadores de los datos de sus pobladores.

a. Importancia para la ciencias sociales

La estadística es importante para todas las ciencias, en ocasiones es más difícil entender sus usos en ciencias humanas y comportamientos sociales pues existe una tendencia en contra de “cuantificar” la experiencia humana, bajo la premisa “la experiencia de vida no se puede poner en números”. Así encontramos pues que no

es fácil ponerle un número a la felicidad que sentimos, o calificar en una escala de 1 a 10 qué tanto amor sentimos por nuestros hijos, incluso resulta complicado tratar de inventar una medida numérica que represente un “te extraño mucho”.

A pesar de ello, la estadística juega un papel muy importante en las ciencias. En sus inicios científicos se considera a un mercader de Londres, John Graunt, como el autor del primer trabajo estadístico serio sobre una población, pues recolectó datos de nacimientos y defunciones desde 1604 hasta 1661. En la misma época Pascal y Huygens realizan trabajos similares. Unos años más tarde Edmund Halley presenta la tabla de mortalidad que se convierte en la base de los estudios contemporáneos.



Es hasta el siglo XIX que se generaliza el método hacia las ciencias sociales, Galton y Pearson son considerados los padres de la estadística moderna pues a ellos se debe la evolución de la estadística deductiva a la estadística inductiva.

3

*John Graunt*

El foco de este curso es La Estadística Descriptiva y se puede decir que se utiliza como herramienta para describir, resumir o reducir las propiedades de un conglomerado de datos con el objeto que sean manejables.



En Ciencias Sociales, también se puede usar La Estadística Inferencial que se utiliza para estimar las propiedades de una población, a partir del estudio de las propiedades de una muestra.

b. ¿Por qué necesita conocer estadística?

Conocer estadística es importante en todos los campos de la vida. La capacidad de recolectar datos permite realizar categorías, comparar grupos, cuantificar las cualidades y entonces predecir sucesos. El buen uso de tablas de recolección permite a las personas representar principalmente fenómenos de una forma numérica.

Por ejemplo, podemos decir con palabras muy bonitas que la asistencia de los niños ha mejorado gracias a un programa de motivación que implementamos y podemos decir que la asistencia inicial era de 50% y en un mes después subió a 53% y que en julio estaba por 65% y llegamos a noviembre con un 95% de la asistencia. Todo lo que acabo de escribir está muy bien y muestra resultados, pero sería mucho más fácil si puedo mostrar eso mismo, con más detalle y más fácil de entender; la estadística descriptiva nos ayuda a ello:



Ejemplo de un registro ficticio de asistencia

En este caso, se presenta una gráfica que muestra de forma más simple y comprensible que lo descrito en las líneas anteriores.

Otro uso importante es el descubrimiento de patrones en los sujetos de quienes recolectamos datos que permiten tomar decisiones importantes en beneficio de las personas. Un ejemplo puede ser el descubrimiento del impacto del agua potable en la talla de la población. Resulta que el ejército de los Estados Unidos solía medir a todos

sus aspirantes a soldados y luego de muchos años de recolectar la información, alguien notó que había poblaciones que en algún momento mostraban tallas más grandes, es decir, crecían más. Al indagar sobre el fenómeno, se logró descubrir que los crecimientos de las poblaciones se daban en los lugares en donde introducían el agua entubada y por lo tanto la gente bebía agua limpia, se enfermaba menos y entonces crecía más.

En educación, la recolección de datos es inherente al ejercicio y la práctica en el aula. El maestro necesita una herramienta que le permita registrar el progreso de sus estudiantes y por lo tanto genera cuadros de notas y va llevando una documentación de los punteos obtenidos en diversas categorías y al final genera un promedio para otorgar una nota. Este ejercicio es el más común y conocido y cuando algún docente decide no seguir este patrón, los padres se molestan porque no hay forma de que el maestro demuestre que su hijo sí sabe!



Ahora bien, en otros ámbitos educativos sigue existiendo resistencia al uso sistemático de la recolección de datos, veamos unos casos:

Pruebas estandarizadas

Las evaluaciones estandarizadas cumplen, principalmente, una función de rendición de cuentas. Es una forma relativamente fácil de evaluar a mucha gente en un mismo momento y permite hacer comparaciones entre grupos. A muchos educadores no les agrada dicho ejercicio pues consideran que el resultado en una prueba no refleja el trabajo de todo el año en un aula, también argumentan que el uso de estos instrumentos provoca que los maestros enseñen para la prueba y no para la vida. Aunque ambos puntos son válidos, las pruebas estandarizadas son datos numéricos que reflejan una realidad en un momento dado y si las entendemos como tales, nos damos cuenta que los datos que producen son muy útiles para tomar decisiones, por ejemplo podemos saber si una cohorte necesita refuerzo en álgebra, si hay una cohorte



que está siendo muy bien preparada para ir a la universidad, podemos identificar alumnos sobresalientes o docentes que logran mejores resultados que sus compañeros.

Decisiones Académicas



Pruebas estandarizadas

¿Cómo decide un grupo educativo qué libro van a utilizar para enseñar a leer? Le apuesto que nunca ha escuchado usted a un director diciendo: vamos a usar este libro un mes en un salón y vamos a usar el otro libro un mes en el mismo salón y luego recolectaremos datos con una

evaluación y compararemos para saber qué libro es mejor!



Nota:

En la estadística como en cualquier ciencia, recolectar datos es una responsabilidad que se asume seriamente. No es ético capturar datos sin informar a las personas involucradas el uso que se hará de la información. Así mismo, no debe recolectarse información sin una razón bien justificada.

Y ahora le pregunto ¿considera que sería importante hacer este tipo de ejercicios? Podría servir para conocer qué marca de tóner compramos a la impresora, para saber qué tipo de pan rinde más en la refacción, qué áreas de la escuela consumen menos energía eléctrica, etc.

Además de los casos mencionados, se pueden recolectar los intereses, la conducta, las aptitudes, los valores, la asistencia, el consumo, el crecimiento, el desarrollo, la conducta alimenticia, los registros en actividades extracurriculares, las relaciones sociales, entre otras.

c. La toma de decisiones basada en datos

Los líderes, gerentes, directores, administradores y encargados de grupos saben que tomar decisiones es cosa difícil. En el mundo ideal cuando debemos tomar una decisión que afecta a un grupo, todas las personas del grupo estarán a favor de dicha decisión, pero esto es una utopía, la realidad nos muestra que siempre existirán personas a favor y personas en contra de la decisión tomada.

Así pues, las personas que toman decisiones han utilizado la herramienta de la estadística como un método para justificar las mismas y así estar seguros que hay datos y cálculos matemáticos que orientan nuestro juicio hacia la mejor decisión.



Como podemos ver, el uso de los datos estadísticos y los cálculos matemáticos que conlleva nos facilitan la capacidad de tomar una decisión con menos probabilidad de cometer errores.

Supongamos que en la escuela planificamos un día familiar para el próximo sábado. Tendremos actividades al aire libre para que padres e hijos jueguen, se diviertan y aprendan juntos. El sistema meteorológico ha dicho que hay un 60% de probabilidad de que ese día llueva. ¿Qué hacemos con el día familiar? Ciertamente el dato estadístico de probabilidad te ayuda a justificar la decisión que tomes.

Los niños de la comunidad se están enfermando del estómago, varios están enfermos y parece que se contagian unos a otros. El sistema de salud consigue un donativo de medicamentos recientemente vencidos que curan la enfermedad en 75% de los casos, aunque se sabe que por estar vencidos 4% de los que la tomen pueden morir. ¿Les damos el medicamento? Ciertamente, 4% significa que de cada 100 niños que la tomen,

4 morirán pero también sabemos que 75 se curarán... los datos, permiten tomar decisiones difíciles. Ejemplo: La tienda de un lugar tiene problemas de liquidez, sus clientes le piden crédito y sus proveedores piden el dinero por adelantado, así que aunque ha vendido no tiene el efectivo para seguir comprando productos.

Una persona le ofrece un préstamo por la cantidad que necesita para seguir funcionando con una tasa de interés mensual del 5% y 10 meses para pagar. Otra persona le ofrece el dinero que necesitan para seguir operando a una tasa del 20% por el total del capital. Una última persona le ofrece el préstamo de lo que necesita a cambio de 5% de los ingresos de la tienda por los próximos 2 años.

El dueño de la tienda acude a usted para pedirle consejo: ¿qué le conviene más?

En este curso, aprenderemos a usar la información para tomar buenas decisiones laborales y personales usando la estadística como herramienta.

6



La Estadística, una herramienta de gestión de datos.



Comprobación del capítulo I: La estadística en la historia

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo I

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



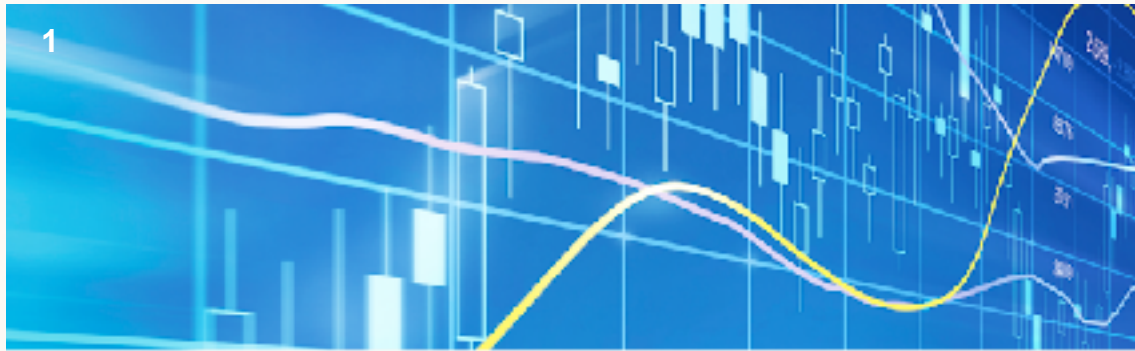
Capítulo II

Áreas de la estadística



Áreas de la estadística

Con el fin de organizar las ideas sobre la estadística como un área de estudio, es importante hacer una división en áreas o componentes que faciliten clasificar el esquema mental para abordar el estudio.



Exploración de datos

Como primer punto se discute sobre el diseño, sin tratarse de un tratado sobre diseño de experimentos o de estudios científicos, se aborda de forma general una propuesta para pensar en problemas cotidianos y del quehacer laboral desde un punto de vista estadístico.

Posteriormente se aborda la descripción, que es el área de la estadística que permite trasladar un fenómeno, una situación o una relación a su expresión numérica y/o gráfica. Esta área facilita la comprensión de la información, tanto para el entendimiento como para la toma de decisiones.

Por último, se aborda la inferencia en estadística. Aunque este tratado se refiere a Estadística Descriptiva, resulta importante abordar la inferencia de una forma muy general que permita conocer los alcances de la estadística.

a. Diseño

A inicios del siglo XX, el prolífico escritor de varios géneros H.G. Wells escribió una frase con mucha visión de futuro y que cobra vigencia en nuestros días. Además la frase tiene relevancia para el tema que estamos tratando.

“El pensamiento estadístico será un día tan necesario como la habilidad de leer y escribir”.



Para cualquier persona en posición de liderazgo, un problema puede pensarse como una relación entre variables, como una acumulación de datos o como una probabilidad matemática, sin embargo lo importante es tener claridad de la razón por la cual se desea pensar la solución del problema a través de los datos.

Suele decirse que el primer paso para la solución del problema es la planificación; si tenemos claros los pasos que debemos tomar para llegar a la meta, será fácil saber si vamos por el camino deseado. El diseño usando estadística descriptiva es en buena medida esa herramienta que ayuda a guiarnos por el camino.



2



Herbert George Wells

Sucede que en la actualidad, vivimos en la era de la información. El acceso a internet permite tener al alcance, conocimiento del mundo entero, la capacidad de los equipos de cómputo para almacenar información de forma eficiente,

clasificarla y accederla permite a los usuarios acceder a la información a través de búsquedas específicas.

Además existe en el mundo entero una cultura de solicitar y entregar datos personales, puede verse claramente en las redes sociales, las compras por internet, pero también al registrarse a un evento, al inscribirse a un programa, cuando llamamos para ordenar comida rápida, entre otros. En todos estos casos brindamos información que queda almacenada para el uso posterior y los publicistas y expertos en mercadotecnia la utilizan para diseñar campañas publicitarias específicas.

En la lógica descrita en el párrafo anterior, cabe entonces hacer la reflexión: Con todos estos datos disponibles ya no se trata de preguntarse ¿dónde obtengo la información? Ahora se trata de responder a la pregunta ¿cómo selecciono y utilizo la información disponible para resolver el problema que enfrento?

En el contexto educativo, solamente basta con ingresar a las páginas de las unidades de evaluación nacional; en Guatemala y en la

mayoría de países con sistemas de evaluación sólidos y serios es mandatorio por ley que las bases de datos de los procesos de evaluación realizados estén disponibles para todo público.



Selección de información

Con dicha información se encuentra un registro de cada estudiante evaluado y toda la información recopilada. En dichos procesos los alumnos responden cuestionarios que registran información del tipo de vivienda, del idioma hablado en casa, del acceso a servicios del hogar, de los estudios de los padres, cantidad de hermanos y hermanas, auto identificación étnica, necesidades educativas especiales y muchas categorías más de datos.



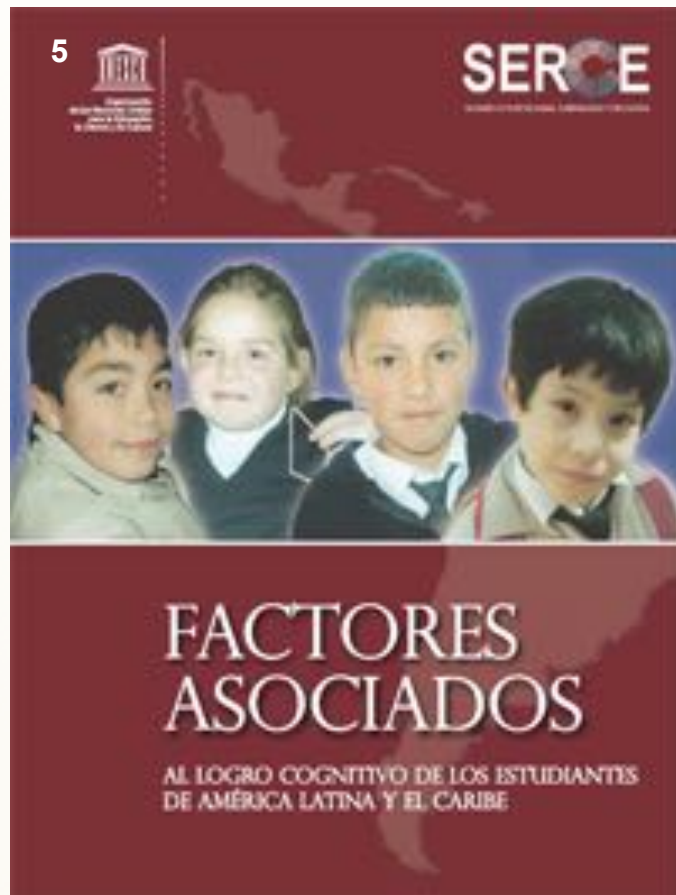
Base de Datos de Evaluaciones Nacionales

Por supuesto, las bases de datos públicas deben respetar la confidencialidad de los evaluados y por lo tanto no tienen nombres, sin embargo, el director de cada establecimiento tiene la potestad para pedir la información de su establecimiento y revisar la información para conocer mejor a sus estudiantes.

Además de las preguntas obvias de los resultados académicos hay preguntas que podría hacerse y ser de utilidad para la institución.

- ¿Cuál es el nivel de estudios de los padres de mis alumnos? Con esta información puede organizarse talleres de padres de acuerdo a su nivel de comprensión o se puede organizar talleres de capacitación para mejorar su nivel.
- ¿Cuántos de mis estudiantes desayunaron antes de venir a clases? Puede servir de justificación para solicitar más apoyo en programas sociales.
- ¿Cómo es el tipo de vivienda de los alumnos que vienen a esta institución? Permite hacer una caracterización del nivel de vida de los alumnos para planificar la educación de acuerdo al contexto.
- ¿Cuántos de mis alumnos de primero primaria asistieron a preprimaria? Brinda datos para justificar la organización de un programa de aprestamiento.

Con el uso de estadística inferencial (un curso siguiente) se puede además responder preguntas más complejas:



Estudio de Unesco

- ¿La asistencia a preprimaria afecta el rendimiento de los alumnos?
- ¿Hay relación entre el idioma materno de mis estudiantes y su resultado en lectura?
- ¿Qué tanto se relaciona la habilidad de lectura de mis estudiantes con sus resultados en la resolución de problemas matemáticos?

Luego del análisis realizado podemos entonces darnos cuenta que los líderes, gerentes, directores y coordinadores de la actualidad no se enfrentan a la pregunta de ¿dónde están los datos? Sino a la pregunta ¿cómo uso la información disponible para tomar mejores decisiones?



Toma de decisiones

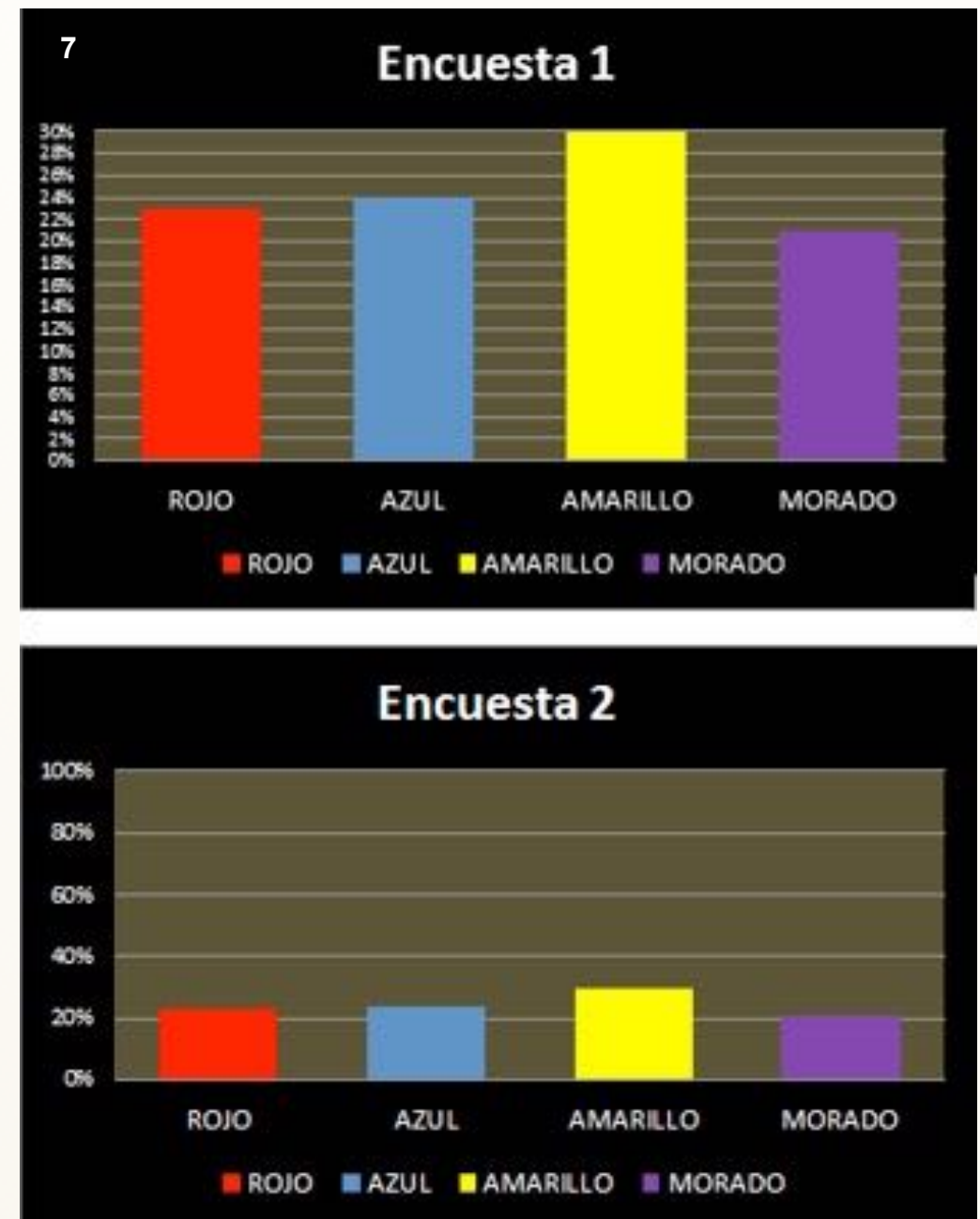
Así pues, se puede resumir que la estadística sirve como apoyo para planificar métodos que guían la toma de decisiones en cuatro aspectos:

1. Presentar y describir la información de forma adecuada.
2. Obtener conclusiones de grandes cantidades de datos.
3. Conocer cómo mejorar procesos.
4. Realizar pronósticos.

b. Descripción

La estadística puede utilizarse perfectamente para describir fenómenos sociales usando los números. Estamos constantemente bombardeados con este tipo de información y por ello es importante reconocerla y saber interpretarla.

Un ejemplo muy claro son las encuestas de intención de voto: en ellas un grupo de expertos selecciona una muestra representativa de personas y las buscan por todo el país. Se entrevista aproximadamente a 1,200 sujetos y se les pregunta ¿por quién piensa votar en las elecciones? Los resultados se registran, se crean bases de datos y se generan tablas de resultados, el producto final es una gráfica como las dos que veremos.



Comparación de encuestas

¿Puede usted identificar qué diferencias hay entre la encuesta 1 y la 2?

A primera vista, en la encuesta 1 el amarillo tiene una diferencia marcada respecto a los demás, sin embargo los datos son exactamente los mismos, la única diferencia que se ha manipulado es el eje y de la gráfica; la escala de comparación y eso hace que la diferencia se vea más grande aunque las cifras sean exactamente las mismas. En ambas figuras el amarillo equivale a 30%, el azul a 24%, el rojo a 23% y el morado a 21%.

Definición de estadística descriptiva

Si hacemos una búsqueda en libros o en internet, encontramos varias definiciones, aquí se han escogido tres, que resumen la misma idea para definir la estadística descriptiva.

1) Wikipedia nos dice: “La estadística descriptiva es la rama de las Matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por

ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto”.

2) “La página ditutor lo plantea así: La Estadística descriptiva registra los datos en tablas y los representa en gráficos. Calcula los parámetros estadísticos (medidas de centralización y de dispersión), que describen el conjunto estudiado”.

3) “La página Aula Fácil: La estadística descriptiva es una ciencia que analiza series de datos y trata de extraer conclusiones sobre el comportamiento de estas variables”.

Después de leerlas y para los fines que nos interesa, diremos que la estadística descriptiva es la rama de la estadística que se enfoca en registrar datos para analizarlos con el fin de poder describirlos.



c. Estadística inferencial

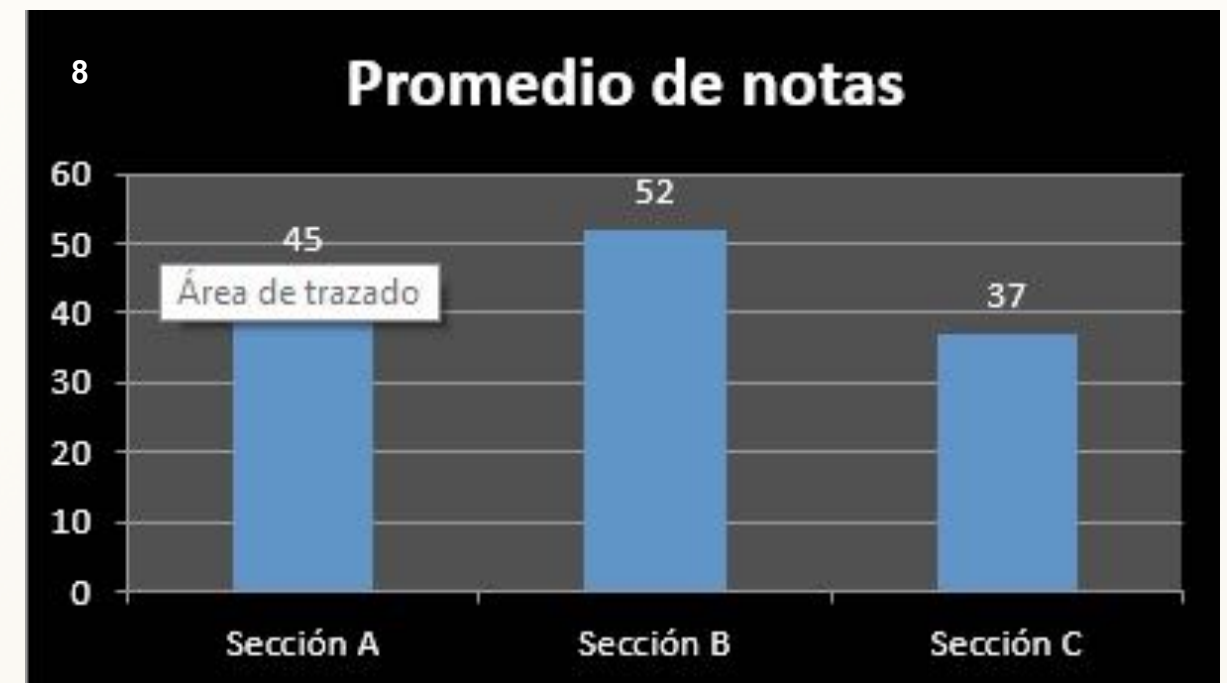
Al igual que la definición de estadística descriptiva, podemos encontrar definiciones de la estadística inferencial que nos brindan claridad para entender la diferencia entre una y otra. Aunque este documento no se dedica a la estadística inferencial, es importante saber de qué se trata, sus alcances y limitaciones.

Entre las definiciones podemos mencionar las siguientes de fácil acceso:

“La estadística inferencial es una parte de la estadística que comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población estadística, a partir de una pequeña parte de la misma”.

“La Estadística inferencial o Inferencia estadística estudia procedimientos para sacar conclusiones generales para toda la población a partir del estudio de una muestra, y el grado de fiabilidad o significación de los resultados obtenidos”.

El concepto es más complicado que el de estadística descriptiva, pues la intención de la estadística inferencial es obtener conclusiones científicamente más rigurosas que lo que se puede observar en la estadística descriptiva. Por ejemplo, veamos la siguiente gráfica:



Promedio de notas por sección

La ilustración anterior muestra unos resultados, los que supondremos son de una evaluación de matemáticas en donde se han escogido 10 niños al azar de cada sección, para conocer cómo están avanzando en ésta área.

Al generar los resultados la estadística descriptiva nos dirá que la sección B tiene un promedio de notas mejor que las otras dos secciones.

Sin embargo, la realidad es que ese dato proviene de una muestra de estudiantes y por lo tanto con esa gráfica es muy arriesgado concluir que en efecto la sección B es mejor.

Correríamos un riesgo si decidiéramos felicitar a ese maestro y usarlo de ejemplo para que los demás lo imitaran, porque puede ser que el resultado se deba más bien a la selección de los 10 estudiantes y así, si hubiéramos escogido a 10 alumnos diferentes, el resultado podría ser diferente.

Aquí es donde la estadística inferencial toma su turno y con sus procedimientos puede responder otras preguntas como por ejemplo:

- ¿Esa muestra representa a toda la sección? Si la respuesta es sí, entonces la diferencia de resultados no se debe directamente a la escogencia del grupo.
- ¿La diferencia entre los promedios de las secciones es estadísticamente significativa? Si la respuesta es sí, entonces podemos decir que la sección B es mejor.

La estadística inferencial tiene varios elementos importantes.

- Muestreo, se refiere a la forma adecuada de obtener una muestra.
- La estimación de parámetros, permite generalizar los resultados a partir de una muestra hacia una población.
- Pruebas de hipótesis, que permiten brindan rigurosidad para decidir si dos muestras son estadísticamente diferentes.



Comprobación del capítulo II: Áreas de la estadística

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo II

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo III

Los datos en estadística

The image shows a person's hands holding a tablet that displays a web-based survey. The survey is titled "Survey" and is part of a "Customer Satisfaction Survey". It asks the user to "Please click on each line to indicate how much you rate level of service". The survey includes five rows of questions, each with five radio button options: Excellent, Good, Average, Fair, and Very Poor. The questions are:

- a. Location
- b. Location
- c. Facilities
- d. Staff
- e. Value for money

Below the questions is a text input field labeled "If you have any comments, please write them here". At the bottom of the form is a blue "Submit" button. The tablet is held by a person wearing blue jeans, and the background is a blurred indoor setting.

	Excellent	Good	Average	Fair	Very Poor
a. Location	☐	☐	☐	☐	☐
b. Location	☐	☐	☐	☐	☐
c. Facilities	☐	☐	☐	☐	☐
d. Staff	☐	☐	☐	☐	☐
e. Value for money	☐	☐	☐	☐	☐

If you have any comments, please write them here:

Los datos en estadística

Hasta ahora hemos discutido en varias oportunidades sobre el fácil acceso a los datos en la era en la que vivimos. También hemos tratado la importancia de usar los datos para la toma de decisiones. Se ha discutido sobre la confidencialidad de archivar y usar los datos de personas, especialmente cuando son poblaciones vulnerables. También se ha hablado de la recolección y organización de los datos para luego analizarlos y presentarlos.

Así pues, en este capítulo discutiremos un poco sobre qué son los datos; para ello vamos a iniciar hablando de otro término: Variable

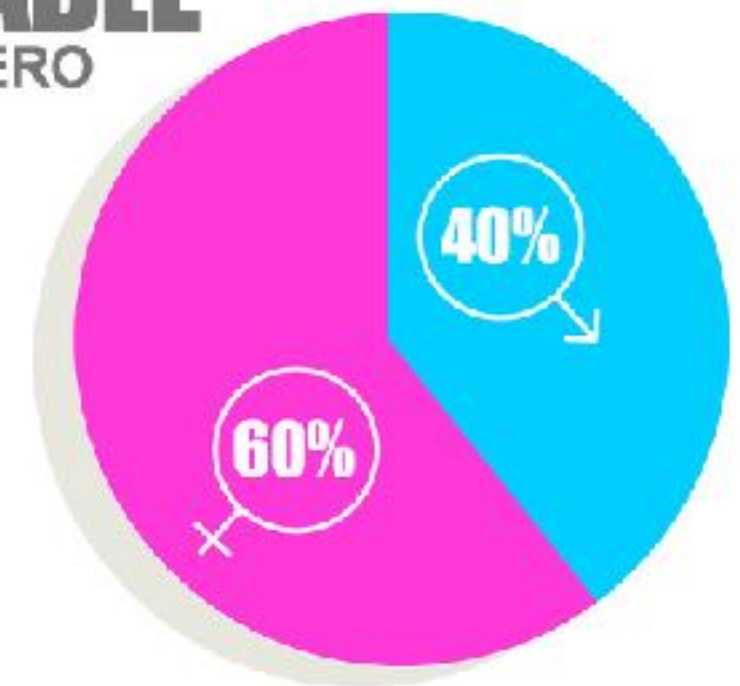
Una **VARIABLE** es un atributo o una característica que puede adquirir más de una modalidad, por ejemplo: Género, es un atributo del ser humano que puede adquirir dos valores: masculino o femenino y al usar género como una variable podemos usarla para describir una población.



Ejemplo: En esta escuela un 60% de los estudiantes es de género femenino y 40% de género masculino.

1

VARIABLE
GÉNERO



Variable

También hay variables que adquieren valores de jerarquía como “cantidad de horas viendo TV al día” que puede adquirir valores desde 0 hasta 24.

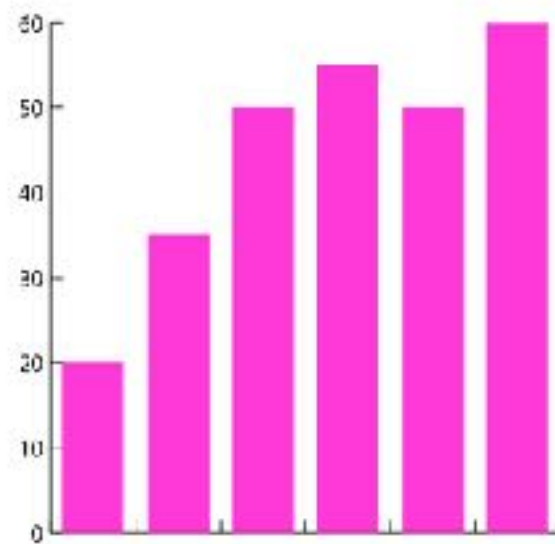
Aparte de las variables, también diferenciamos a las **CONSTANTES** y desde la definición anterior puede deducirse que es un atributo o

característica que puede adquirir únicamente una modalidad. Por ejemplo el equivalente en centímetros de una pulgada es siempre 2.54 cm/pulgada.

2

CONSTANTE

GÉNERO FEMENINO



Constante

a. ¿Qué son los datos y por qué son importantes?

Entonces llegamos a la pregunta importante ¿qué son los datos? Los datos se entienden entonces como la representación numérica de las modalidades que



puede adquirir una variable.

Su importancia radica en que representan la modalidad de la variable en el momento del estudio, así pues la edad de una persona en el momento de la evaluación puede ser por ejemplo, 53 años y ese dato quedará registrado.

Cuando tenemos muchos datos archivados podemos empezar a descubrir tendencias y podríamos suponer que estamos tratando de determinar la edad de contratación más común en docentes universitarios, así que vamos a nuestro archivo y encontramos que una persona había reportado 53, pero que la mayoría de los registros que tenemos reportan entre 40 y 45 años.

Tener la comprensión de lo que el dato significa es una virtud que todos debemos desarrollar. Veamos un ejemplo:

Una universidad quiere contratar a alguien para que organice y utilice los datos que tienen en registro, así que diseñan una prueba práctica. Al final del proceso de selección, únicamente hay dos candidatos que se someten a la prueba que

consiste en revisar los expedientes de 100 catedráticos, encontrar allí la ficha de solicitud de trabajo y usar el dato de la edad para hacer una recomendación en base a lo encontrado, a la universidad.

3



Archivos

Persona 1: revisa los datos, hace una tabla con las edades y calcula el promedio. Reporta que la edad promedio de los docentes en la universidad es 35 años así que sugiere que la universidad tenga facilidades para padres jóvenes porque la mayoría están en edad de tener hijos con necesidades de guardería.

Persona 2: también revisa los datos y hace una tabla con las edades y también calcula el promedio y concuerda con el dato de 35 años. Sin embargo, su propuesta es diferente. Propone que para atraer nuevos docentes es importante dirigir los esfuerzos a un grupo objetivo de entre 25-35 años.

¿Quién de los dos ha hecho la recomendación más apropiada en base a los datos colectados?

La respuesta es la persona 2; este sujeto utilizó los datos de forma apropiada y por lo que representaron en el momento de su recolección. Los archivos que les brindaron eran la ficha de solicitud de trabajo, por lo que el promedio de 35 representa la edad de contratación de los docentes y no la edad actual de los mismos.

b. ¿De dónde provienen los datos?

Como se ha mencionado en varias ocasiones, los datos abundan en la actualidad, así que al preguntarse dónde están la respuesta es “en todas partes”, simplemente hay que saber buscarlos para poder utilizarlos. Recordemos, que la buena recolección es la primera etapa del buen uso que les demos.

Para organizar la fuente de nuestros datos vamos a pensar en dos tipos de orígenes para los datos:

PRIMERO: Los datos que provienen de una recopilación (fuente primaria) o compilación de datos (fuente secundaria). En este caso se habla de datos que ya están publicados por el gobierno, la industria, los periódicos, las universidades, entre otros.

Este tipo de información suele estar disponible para el público en general y basta con hacer búsquedas en internet para encontrarlas. En Guatemala puede accederse a la página del

Ministerio de Educación, en la Dirección General de Planificación y se encontrarán los datos educativos por año, grado, género, identificación étnica, por departamento y en ocasiones por municipio.

Se puede además revisar para cada establecimiento las tasas de aprobación, la matrícula, la repitencia, la cantidad de niños inscritos, la cantidad de maestros, la cantidad de escritorios para cada centro, en fin, hay mucha información disponible que ya fue recopilada para su uso.



Datos

SEGUNDO: Otra fuente de datos son los diseños de experimentos, entendiéndolo como un estudio o ejercicio que controla las variables que pueden afectar un resultado y que dicho resultado se convierte en datos.



Por ejemplo, un investigador desea saber si la refacción escolar tiene un efecto sobre el rendimiento de lectura de los niños. Para ello hará un experimento y buscará dos escuelas con condiciones similares, donde una brinde refacción a los niños y la otra no. Las condiciones para que sean similares las deberá definir, como: el tamaño (cantidad de alumnos y docentes), el tipo de población (etnicidad), el nivel socioeconómico, entre otras variables que se sabe que pueden afectar los resultados.

Luego de identificar a las escuelas, el investigador necesitará evaluar a los niños con la misma prueba y probablemente en más de una ocasión y así al final de un período, uno o dos años, podrá decir si existe alguna diferencia en el rendimiento de lectura en los niños que reciben refacción

escolar, respecto a los que no la reciben. El investigador además tendrá una tabla de datos con el nombre de cada niño, sus condiciones “variables” y los resultados en las pruebas que aplicó. Todos estos datos y sus resultados serán publicados y cualquiera podrá acceder a ellos como una fuente secundaria.



Datos

TERCERO: Los datos que provienen de una encuesta. La principal diferencia de los datos que vienen de una encuesta, con los datos que vienen de un experimento, es que en la encuesta no hay posibilidad de controlar las variables, simplemente obtenemos los datos que nos brindan al responder nuestras preguntas. Todas las respuestas pueden codificarse para generar una tabla de datos que se puede analizar.



Hay muchos ejemplos de encuestas, como las encuestas de vivienda, de juventud, de creencias, de intereses, de opinión, de preferencias sobre productos, de trabajo, entre otras.



Encuesta

CUARTO: La cuarta y última fuente de datos es a través de la observación. En muchas ocasiones es imposible que un investigador realice un experimento, por ejemplo, en el caso de la refacción escolar, no sería ético que el investigador decida no darle comida a un grupo de niños con tal de investigar. Otra razón por la cual se necesita observar es



porque las personas pueden no responder bien a las encuestas, tal vez no saben leer, tal vez no entienden el idioma o tal vez sea ofensivo o intrusivo preguntar ciertos temas. Por lo tanto se realizan observaciones y el investigador crea códigos y contabiliza lo que desea observar.



EOURM San Miguel Milpas Altas, Sacatepéquez, Guatemala.

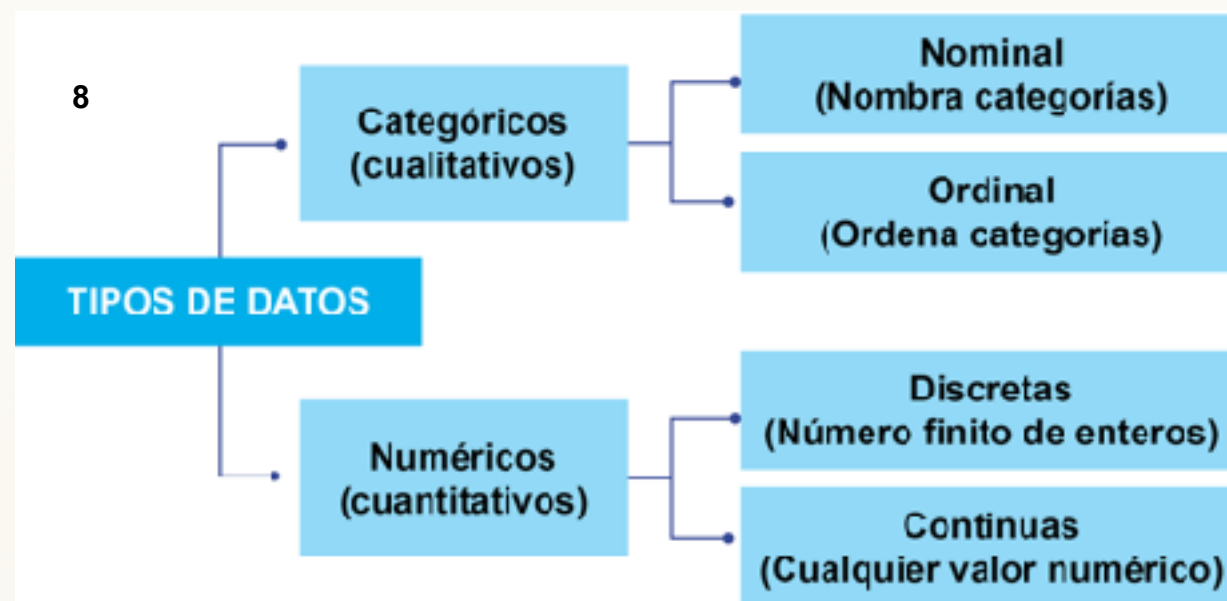
Algunos ejemplos:

- En el recreo de la escuela, se observa y cuantifican las veces que sucede un hecho de agresión entre estudiantes.
- En un aula, se observa y cuantifica las veces que intervienen los niños en clase y las veces

que intervienen las niñas en clase (para saber si el docente promueve la equidad de género).

- En las reuniones mensuales de padres de familia, se observa y cuantifica la cantidad de veces que los padres proponen cambios que los involucran a ellos.

c. ¿Qué tipos de datos existen?



Tipos de datos

En la ilustración se presenta una clasificación de los tipos de datos. En realidad la agrupación de los datos es conocida como “variable” que en matemáticas se le representa con un símbolo o representación y puede adquirir más de un valor dadas las circunstancias.



En nuestro caso, estas variables son las que vamos a clasificar para responder a la pregunta ¿qué tipos de datos existen?

Variables Categóricas: son todas aquellas cuyos valores son categorías que se excluyen entre sí: por ejemplo respuestas de sí o no ante una pregunta, pueden únicamente adquirir uno de los dos valores. Estado civil, color de ojos, sexo son otros ejemplos de datos que se convierten en variables categóricas.

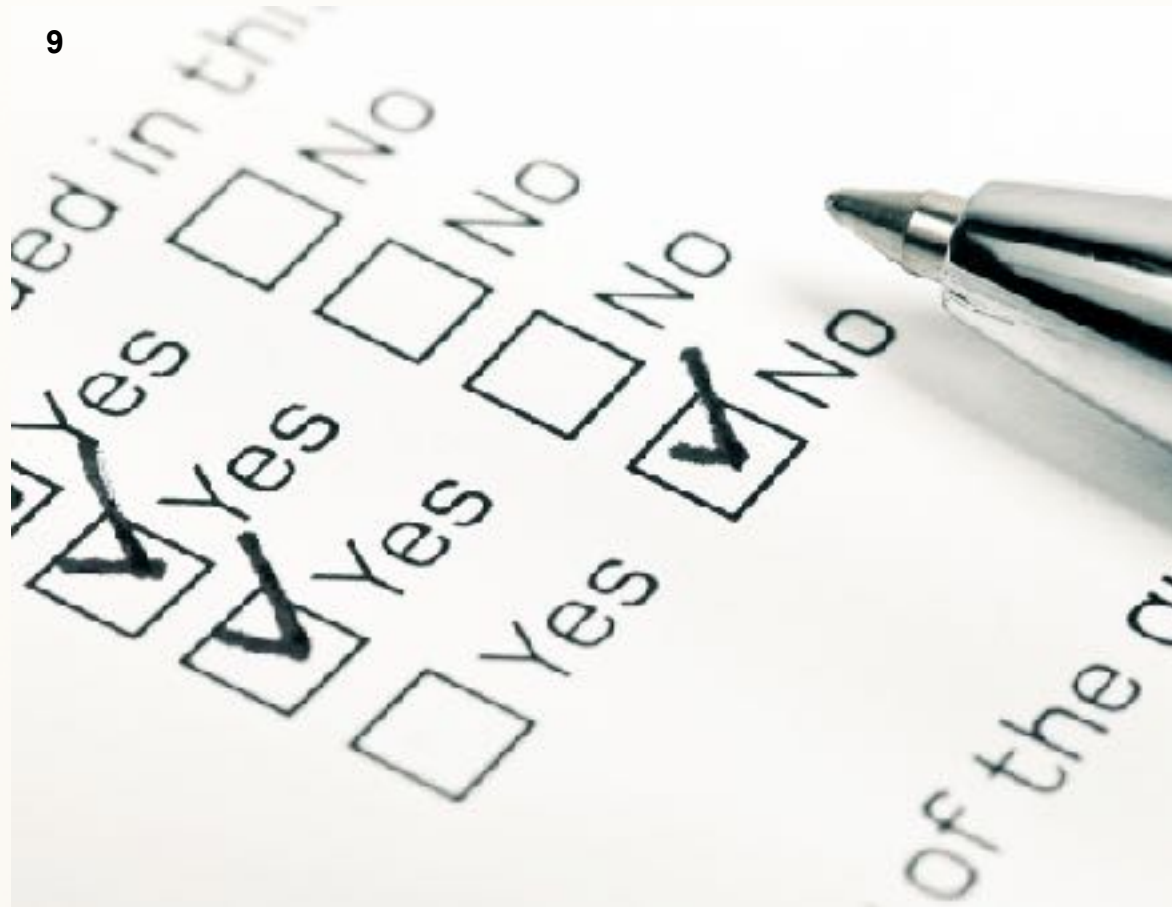


Además, las variables categóricas pueden ser de dos tipos Nominales y Ordinales.

Nominales: son aquellas en donde los valores potenciales provienen únicamente de una lista de alternativas:



9



Variables categóricas

¿Cuál es su deporte favorito?

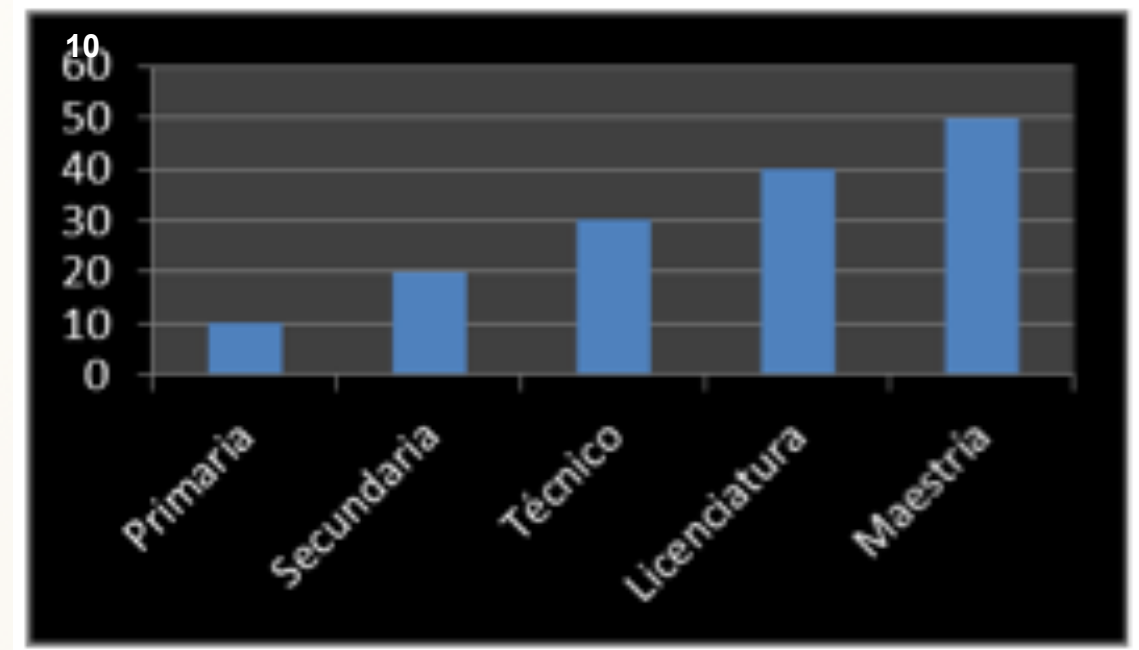
Fútbol

Baseball

Basquetbol

En este caso, ninguna de las opciones es más o es menos que las demás, simplemente representan una categoría diferente.

Ordinales: son aquellas en las cuales los valores potenciales obedecen un orden natural.



Variables categóricas ordinales

¿Cuál es el último grado académico que obtuvo?

Primaria completa

Secundaria completa

Técnico universitario

Licenciatura

Maestría

En este caso, cada valor posible para la variable representa un nivel superior porque son más años de estudio para la persona y mejor preparación académica.

Variables Numéricas: son todas aquellas cuyos valores representan un número o tienen implícito un número. Por ejemplo: edad, calificación obtenida, estatura, peso, salario etc.

Estas variables numéricas pueden ser Discretas y Continuas.

Discretas: son aquellas variables en donde los valores numéricos son finitos y exactos.

¿Cuántos hijos tiene?

1, 2, 3, 4, 5

Estos números responden a la cantidad de hijos que alguien puede tener. Es un número exacto.

Continuas: son aquellas variables que obedecen a un valor intermedio entre un intervalo.

¿Cuánto pesas?

120.3



Comprobación del capítulo III: Los datos en estadística

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo III

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo IV

Evaluación general de procesos estadísticos



Evaluación general de procesos estadísticos

Para finalizar la primera unidad haremos una revisión rápida de algunos elementos importantes en la evaluación de procesos estadísticos.

Esta situación es importante tomarla en cuenta porque muchas veces los directores, coordinadores, gerentes, supervisores y administradores tienen en sus manos la decisión de solicitar o autorizar procesos de investigación o de evaluación y se espera que la persona tenga suficiente conocimiento para determinar que la información sea verdadera.

Para evaluar el proceso la primera pregunta que debemos hacer es: ¿para qué se va a realizar el estudio?

Para que la respuesta sea satisfactoria debe resolver un problema o apoyarnos a entenderlo para resolverlo. Es frecuente la



realización de estudios solamente porque al investigador le llama la atención conocer el resultado y esto equivale a gastar tiempo y dinero en algo que no servirá a nadie más que a la curiosidad del investigador.

Así pues usted puede preguntarse cosas como:

¿Necesita mi institución el resultado que ofrece este estudio?

¿Vale la pena usar el tiempo y recursos para este estudio?

¿Confío en las capacidades de la persona o el equipo investigador?

a. Métodos de muestreo

Una muestra se puede definir como una parte de la población que deseamos estudiar. Los recursos económicos, el tiempo y la logística hace que muchas veces sea imposible obtener información de todas las personas que quisiéramos. Los investigadores seleccionan una muestra, la cual es suficiente para representar a un conjunto mayor; con base a esa muestra extrapolan los resultados



para hacer sus predicciones. El mejor ejemplo son las encuestas de intención de voto en procesos electorales, se entrevista a unas cuantas personas para que puedan determinar qué partido político ganará las próximas elecciones. Normalmente las muestras bien hechas producen buenas predicciones.

Para clasificar las muestras, hablaremos de muestras probabilísticas y muestras no probabilísticas.



Metodos de muestreo

Muestras Probabilísticas

En el caso de las muestras probabilísticas se requiere seleccionar a los sujetos de acuerdo a la probabilidad de su aparición en la población. El primer paso para hacer una muestra de este tipo, es obtener datos de toda la población que deseamos estudiar. Si estamos hablando de los alumnos de un colegio, necesitamos el listado de todos los estudiantes inscritos; si hablamos de todos los niños del país en 4to grado primaria, se requiere también el listado completo. Esta información se usa para crear un “marco muestral” que respete las características de la población y entre más exacto sea el marco, mejor será nuestro estudio.



Por ejemplo: En el caso de los alumnos de 4to grado primaria, necesitamos saber qué proporción existe entre niños y niñas, porque si en 4to grado de todo el país hay 2 niñas por cada niño, cuando elaboremos una muestra ésta deberá tener 2 niñas por cada niño también.

Entre más información tengamos mejor será nuestra muestra y por lo tanto nuestros resultados representarán de mejor forma la realidad.

La forma más simple de muestreo probabilístico es la “Muestra aleatoria simple” en donde cada sujeto de estudio tiene la misma probabilidad de ser seleccionado. También se le puede entender como un proceso de selección de individuos de manera que cada sujeto tiene probabilidad positiva e independiente de ser seleccionado. Equivale a poner el nombre de cada miembro de la población en una bolsa, moverla y con los ojos cerrados sacar uno a uno los papeles con los nombres de quienes forman la muestra.



Una forma más compleja es la estratificación de la muestra. Una muestra estratificada requiere dividir la población en grupos de acuerdo a características seleccionadas, como en el ejemplo de 4to grado primaria, dividir entre niños y niñas, también puede ser entre urbano y rural o entre niños con sobre edad y niños en la edad esperada.



Luego de dividir la población, se realiza una selección aleatoria simple en cada estrato. Las muestras estratificadas son más rigurosas que las aleatorias simples.

Una tercera forma de muestreo es la muestra por conglomerados. Consiste en dividir la población de estudio en grupos llamados conglomerados asegurándose que cada uno sea representativo de la población completa. Luego se seleccionan de forma aleatoria simple los conglomerados que van a estudiarse. Este tipo de muestreo permite hacer estudios a largo plazo en donde el marco muestral de conglomerados se selecciona una vez y en las siguientes recolecciones de datos solamente se cambia los conglomerados.



Muestras No Probabilísticas

Se utilizan cuando el investigador no puede seleccionar de forma aleatoria a los participantes así que los elementos que participarán son escogidos a juicio de una persona o equipo responsable del estudio.

Intentan ser representativos por los criterios del investigador, aunque no tienen el peso estadístico para garantizarlo.

Se puede definir como el tipo de muestreo en donde la captura de datos no brindan a todos los sujetos de la población iguales posibilidades de ser elegidos para participar.

Existen los siguientes métodos:

Método por cuotas: en donde se divide la población en grupos y se llena con sujetos de una forma no probabilística.

Muestreo Discrecional: es cuando los sujetos son seleccionados por el juicio y criterios del investigador. Por conveniencia, es un tipo de muestreo que el investigador selecciona a los sujetos por cercanía, por accesibilidad u otro factor que le conviene.

Muestreo tipo bola de nieve o referidos: es el método en donde se encuentra un sujeto con las características deseadas y se le pide referencia de personas similares para agregarlas al estudio.

Muestreo Accidental o Causal: es cuando se selecciona un lugar y un momento, se hace el estudio con el personal que accidentalmente o de forma fortuita se encuentre allí.

b. Evaluando la certeza de una encuesta

Para iniciar vamos a empezar diciendo que la encuesta es probablemente el principal instrumento de recolección de datos cuantitativos (numéricos) en ciencias sociales y como tal su objetivo es explicar la realidad del fenómeno en estudio y para ello requiere la construcción de una matriz o tabla de datos que puedan analizarse.



Muestreo

Se debe tener muy en cuenta que por ser una técnica importante, su utilización se ha vuelto muy común, y resulta muy tentador para cualquier persona que quiere encontrar respuestas a fenómenos sociales, que decida realizar un encuesta.

Aunque en principio suene como una buena idea, la encuesta no es una herramienta que puede construirse "así porque sí". Tiene ciertos parámetros muy importantes que se debe respetar.

En este apartado revisaremos los puntos que debemos tener en cuenta para revisar si una encuesta cumple los requisitos mínimos para considerarla como válida.

Por lo regular, las encuestas se diseñan con preguntas del tipo "selección múltiple" para generar variables "nominales" u "ordinales".

El primer tema que debemos considerar al vernos frente a una encuesta es su propósito.

Cuando nos vemos frente a instrumentos que tienen como interés convencernos de algo empezamos a notar "alertas". Por ejemplo las encuestas de partidos políticos en épocas de elecciones o las encuestas de preferencia de un producto sobre otro. No quiere decir que todas sean incorrectas o negativas, simplemente que empezamos a considerar si los datos representan la realidad o si se están usando los datos para "mentir científicamente".

Veamos un ejemplo de pregunta en una encuesta

¿Su maestro devuelve las tareas de matemática con retroalimentación?

- a. Nunca*
- b. Casi nunca*
- c. Algunas veces*
- d. Casi siempre*
- e. Siempre*



Ejemplo de encuesta

Luego de determinar el propósito, el siguiente punto que debemos notar a la hora de evaluar si la encuesta es o no es útil para nuestros propósitos es la cantidad de opciones que se presentan en las respuestas. En el caso del ejemplo anterior, la pregunta está diseñada con 5 opciones de respuesta (a,b,c,d,e).



En ciencias sociales se sabe que existe una tendencia natural de las personas por buscar siempre "el medio" la "zona de confort", así que cuando la encuesta tiene un número impar de opciones es muy probable que mucha gente responda la opción de en medio.

Esto sucede porque el encuestado puede pensar que sus respuestas podrían perjudicar a alguien, en este caso, a su maestro; también suele suceder porque no queremos dar mucha información de nosotros mismos, así que si la pregunta fuera ¿Qué tan frecuentemente te enojas con tu familia? Muchas personas preferirían responder "algunas veces".



Así pues, si lo que necesitamos de nuestra investigación requiere una postura de las personas, lo recomendable es que exista un número par de opciones de respuesta como el siguiente ejemplo. En nuestro caso, esto produce que las opciones A y B puedan considerarse negativas y las C y D positivas.

¿Su maestro devuelve las tareas de matemática con retroalimentación?

- a. Nunca*
- b. Casi nunca*
- c. Casi siempre*
- d. Siempre*

¿Tus compañeros de clase te golpean?

- a. Siempre*
- b. Casi siempre*
- c. Algunas veces*
- d. Casi nunca*
- e. Nunca*

Debemos tomar en cuenta, que no siempre queremos colocar a las personas en situación de decidir entre extremos, así que cuando tenga una encuesta frente a usted, pregúntese: ¿en este estudio, necesito que las personas se decidan entre un extremo y el otro o que respondan libremente?



Otro punto de mucha importancia es determinar a quién se estará encuestando; es decir cuál es la población y cómo se ha escogido la muestra.



En el caso de las encuestas, la muestra debe ser seleccionada con cualquiera de los métodos **PROBABILÍSTICOS** que vimos en el capítulo anterior. Si la muestra no se escoge de forma probabilista, los resultados no podrán representar la realidad y por lo tanto la encuesta no sería válida.

Cuando veamos a un investigador hacer preguntas a personas y determinemos que su muestra no es probabilística, NO significa que su método no sea válido; recordemos que puede estar haciendo una investigación cualitativa y en lugar de encuestar está entrevistando a los sujetos de su investigación.

Por último y con datos que van más allá de la estadística descriptiva, existe el llamado "error de medición" que se refiere a la precisión con la cual estamos midiendo, por ejemplo si nos dan una báscula y nos piden registrar nuestro peso, será fácil anotar el valor que obtenemos; sin embargo, si nos dan 2 básculas tendremos que calcular un promedio para dar el dato y si nos dan 10 probablemente nos cueste más dar el dato. Esto mismo pasa con las encuestas, y la ciencia de la psicometría se encarga de diseñar las preguntas y efectuar el análisis estadístico para determinar si el instrumento tiene al menos dos valores en rangos aceptables: Consistencia interna y validez.

En nuestro caso, bastará con preguntar si la encuesta fue sometida a análisis psicométricos básicos, los cuales revisaremos a continuación.



c. Uso de Programas de Hojas de Cálculo (Microsoft Excel™)



Muestras estadísticas

Para los fines de este curso, es importante que conozcamos las herramientas que pueden apoyarnos y facilitar nuestro trabajo a la hora de tener la necesidad de manejar datos.

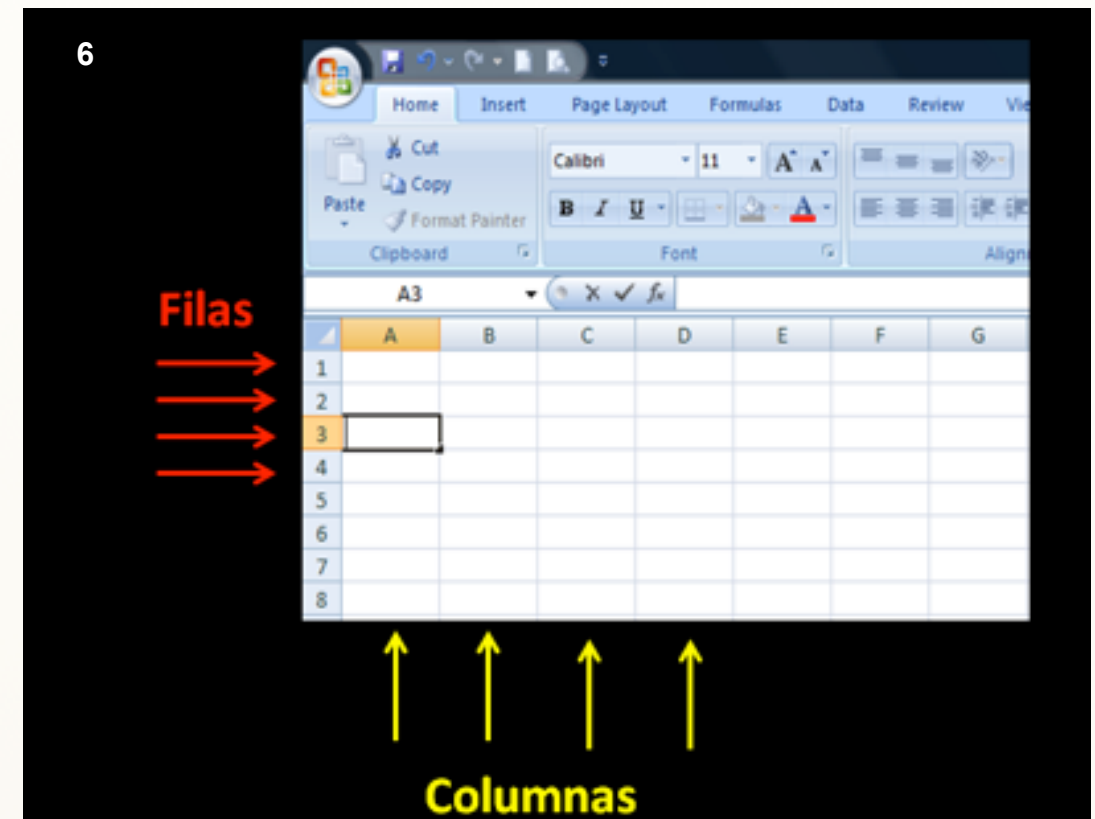
Un programa de hojas de cálculo como el de Microsoft Excel es por excelencia la primer fuente informática para manejar tablas o matrices de datos, principalmente por la accesibilidad a él y su fácil uso. A lo largo de las próximas unidades, se ejemplificarán los métodos para utilizar Excel como un aliado en nuestras investigaciones.

El primer aspecto que necesitaremos conocer de Excel, es la lógica detrás del programa. Surge de la necesidad de registrar datos, administrarlos y la forma más sencilla de hacerlo es a través de una matriz de filas y columnas.



Microsoft Excel

Por lo general se acuerda ordenar los datos de tal forma que las filas representen a los sujetos o a los casos y las columnas a las variables.



Hojas de cálculo

En el programa, las filas están numeradas desde 1 hasta n y las columnas se representan con letras del abecedario.

Como mencionamos, el objetivo es usar cada hoja de trabajo de Excel para fabricar nuestras matrices de datos con la información que nos compete.

Por ejemplo, en la imagen siguiente se muestra una parte del registro de datos generales y notas de un grupo de niños de 5to primaria.

7

	A	B	C	D	E	F	G	H	I
1	Nombre	Edad	Grado	Promedio 1 bimestre	Promedio 2 bimestre	Promedio 3 bimestre	Promedio 4 bimestre	Promedio 5 Bimestre	Acumulado
2	Juan Alberto Rojas	11	5to	67	67	88	65	77	72.5
3	Cristina del Cid	12	5to	71	54	58	58	53	58.8
4	Juan Carlos Ovalle	12	5to	87	61	72	74	85	75.8
5	Marielos Serrano	12	5to	61	71	68	62	54	63.2
6	Diego Alcázar	11	5to	74	94	73	77	81	79.8
7	Matilde López	12	5to	70	88	54	63	77	70.4
8	Rebeca Ramirez	12	5to	74	82	54	64	99	74.6
9	Carlos Solís	12	5to	98	72	96	50	95	82.2

Registro de datos



Comprobación del capítulo IV: Evaluación general de procesos estadísticos

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo IV

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Unidad II

Comprensión y presentación de datos



La segunda unidad se enfoca en la adecuada comprensión de los datos: tipos de escalas, tipos de variables y tablas de datos. Se desarrolla la capacidad de presentar la información en gráficas de acuerdo a sus características.



Métodos de muestreo

The screenshot displays the Microsoft Excel interface. The ribbon at the top includes tabs for FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, and VIEW. The FORMULAS tab is active, showing options like Define Name, Use in Formula, Create from Selection, and Formula Auditing. A blue overlay box in the center contains the text "Capítulo V Organización de datos numéricos". Below this, a table is visible with columns A, B, and C. The table contains the following data:

	A	B	C
1			
2	Name	Age	Amount Raised
3	Suzie	25	150
4	Ron	40	67
5	Sharon	35	80
6	Kelly	22	240
7	Sam	24	175
8	John	55	140
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			

The formula bar at the bottom shows the active cell (C9) is empty, and the status bar at the bottom left indicates "READY".

Organización de datos numéricos

Para poder utilizar la información que vamos recopilando con nuestras encuestas o con la observación o cualquier método que escojamos, vamos a necesitar organizar los datos para que tengan sentido.

A través del tiempo se desarrollaron las tablas de datos, una forma simple y lógica de organizar y almacenar la información. Las tablas de registro son muy utilizadas en todos los ámbitos académicos y profesionales por su practicidad y la facilidad de su uso. Poco a poco, las tablas de registro se fueron quedando cortas, en cuanto a la capacidad de análisis que producían y los usuarios demandaban una forma más eficiente de manejar los datos. Por lo tanto se empezaron a desarrollar las bases de datos y formas electrónicas de manejarlas.

Aunque existen diferentes programas en la actualidad para manejo de bases de datos, la mayoría cumplen las mismas funciones, están estructuradas de forma muy similar y se conforman por: tablas de datos, consultas, informes, macros y formularios.



a. Elementos de una tabla de datos

Tablas: en apariencia son lo mismo que una hoja de cálculo, como las que vemos en Microsoft Excel, la información se encuentra organizada en filas o campos y en columnas o registros. A la hora de hablar de bases de datos y registros organizados, lo importante es tener en cuenta que cada categoría de nuestro estudio deberá tener su tabla de datos.

Es decir, si vamos a llevar registro de estudiantes, entonces tendremos una tabla de datos para ellos. Si además llevamos registro de nuestros docentes, tendremos una tabla de datos únicamente para ellos aunque las variables sean las mismas o muy parecidas; si además llevamos registros del producto que vende la librería, esta será también una tabla de datos diferente.

De esta forma mantendremos el orden que permite relacionar variables con sujetos de la misma categoría.



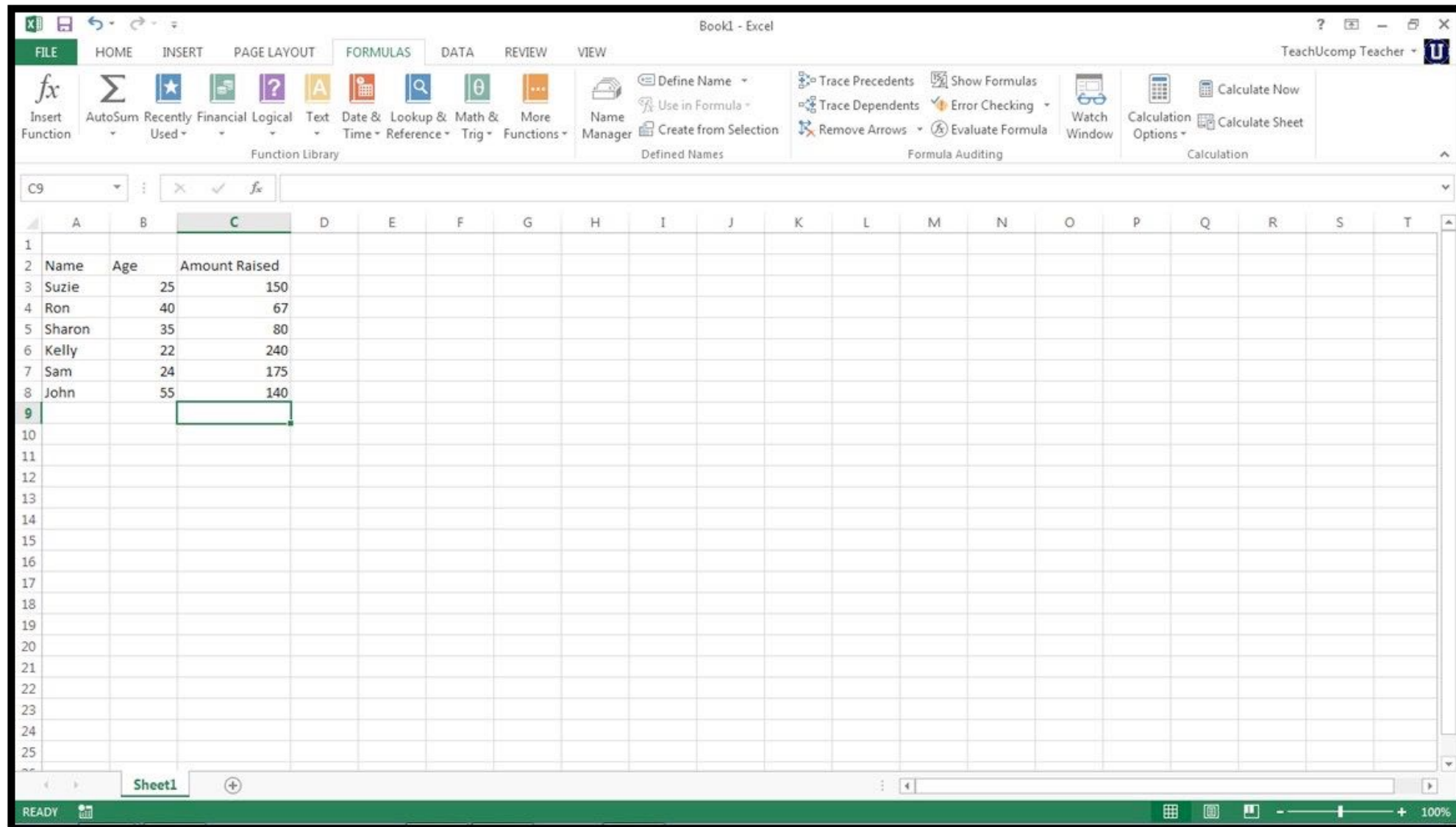
Consultas: en el manejo de tablas de datos como una base de datos, resultan ser el elemento que tiene más relevancia. Su principal función es recuperar información específica archivada. Las consultas se acompañan de la función “filtrar”, por ejemplo si se quiere saber las edades de los docentes que ingresaron a laborar a esta institución entre 1989 y 1999.



Existen consultas que generan procesos de selección, así como el ejemplo anterior, únicamente se obtienen datos para mostrarlos.

Pero también hay consultas que tienen la capacidad de generar acciones, por ejemplo generar tablas de datos nuevas, o añadir variables, incluso borrar datos, por ejemplo si estoy haciendo una recopilación de datos desde hace años y me doy cuenta que tengo una variable en donde registro la cantidad de libros en cada aula, cuando reviso me doy cuenta que existen demasiados valores y muy diferentes, desde 0 hasta 50 y yo quisiera reducir esa cantidad de valores a categorías, digamos: de 0 a 10 = 1, de 11 a 20 = 2, etc. Puedo realizar una consulta para que sustituya dichos valores.

1



The screenshot shows the Microsoft Excel interface with the 'FORMULAS' tab selected. The worksheet contains a table with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1																				
2	Name	Age	Amount Raised																	
3	Suzie	25	150																	
4	Ron	40	67																	
5	Sharon	35	80																	
6	Kelly	22	240																	
7	Sam	24	175																	
8	John	55	140																	
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				
21																				
22																				
23																				
24																				
25																				

Archivo de datos

Informes: un informe sirve normalmente para responder alguna pregunta sobre la información registrada; supongamos por ejemplo que deseamos saber cuántas niñas tienen puntajes arriba de 80 puntos en gimnasia, para ello podemos hacer una consulta y su resultado será un informe.

Macros: son elementos de programación simplificada que ofrecen la mayoría de paquetes de manejo de bases de datos. Permiten al usuario generar secuencias de procesos cortos automatizados.



Formularios: varios programas de manejo de bases de datos permiten el uso de formularios para facilitar la captura de datos y que el usuario ingrese éstos para que se archiven en la tabla de datos que nosotros deseamos. Esta es una herramienta potente que facilita la vida del investigador, pues evita la necesidad de digitar los datos y reduce el error humano.



b. Arreglo ordenado de datos

Se le llama así a una técnica sencilla que permite organizar datos recabados de tal forma que permite observar a simple vista algunas tendencias en forma de concentración de valores, datos extremos o datos comunes.



Vamos a suponer que tenemos la Tabla 1 de datos, que representa las notas que un profesor asigna a sus estudiantes en matemáticas de 3ro. primaria y que hemos recolectado la información de un año únicamente.



¿Qué podríamos concluir con esa tabla a simple vista?

2

70	67	70	90	95	64	64	65	65	75	66	95
66	67	95	68	68	75	71	74	69	69	66	66
80	80	85	70	90	70	90	70	70	41	45	56
70	64	71	68	71	71	75	80	68	75	64	75
71	80	65	80	65	85	90	57	90	66	71	70

Tabla 1. Tabla de datos.

Ahora fijémonos en la Tabla 2, tiene exactamente los mismos datos, pero han sido ordenados y hagamos la misma pregunta ¿Qué podríamos concluir a simple vista?

En primer lugar podemos ver que existe una tendencia a calificar entre 68 y 75 puntos. También podemos pensar que hay una tendencia a calificar a los alumnos “buenos” en múltiplos de 5 (80, 85,

90, 95), mientras que los que tienen mal rendimiento no representan números tan redondos (41, 45, 56, 57).

3

41	45	56	57	64	64	64	65	65	66	66	66
66	67	68	68	68	68	68	68	68	69	69	69
70	70	70	70	70	70	70	70	70	70	70	70
70	71	71	71	71	71	74	74	75	75	75	75
75	80	80	80	80	85	90	90	90	95	95	95

Tabla 2. Arreglo ordenado de datos.

Simplemente por haber organizado los datos, tenemos la posibilidad de generar nuevas preguntas para nuestra investigación, tenemos evidencia para preguntarle al docente qué está haciendo para calificar, también para decirle que el grupo parece rendir de una forma pareja pues la mayoría tienen rendimientos similares, podemos preguntarle si tiene estrategias específicas para los alumnos con punteos más altos, aunque son pocos, en fin, ordenar los datos permite que fácilmente y sin hacer cálculos complicados, podamos obtener información valiosa.

c. Diagrama de tallo y hoja

El objetivo del diagrama de tallo y hoja también es presentar información de una forma organizada, diremos que sirve para resumir y entregar datos de forma más amigable al usuario final.



El diagrama está compuesto por dos elementos: Tallos y Hojas. El tallo será el número principal o números principales y la hoja serán los números secundarios. En nuestro ejemplo de la tabla 2. Formamos el Tallo con el primer dígito que lo representan las decenas y la hoja la llenamos con los dígitos de las unidades.

4

Tallo	Hoja
4	5
5	6 7
6	4 4 5 5 6 6 6 6 7 8 8 8 8 8 8 8 9 9 9
7	0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 4 4 5 5 5 5
8	0 0 0 0 5
9	0 0 0 5 5 5

Ejemplo de diagrama de tallo y hoja

El diagrama se llena colocando una hoja por cada vez que aparece el dato, entonces podemos saber a simple vista dónde hay más acumulación de datos. Los 70's son las notas más comunes, hay 23 registros allí y podemos saber que la más común es la nota 70 con 12 registros.

Tanto el arreglo ordenado de datos, como el tallo y hoja sirven para presentar la información y facilitar la interpretación, pero recordemos siempre que debemos tener una razón para la cual deseamos hacerlo.



Comprobación del capítulo V: Organización de datos numéricos

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo V

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo VI

Tipos de variables y tipo de escalas de medida



Tipos de variables y tipo de escalas de medida

Muchas personas dejan por un lado el hablar de datos estadísticos, ya sea por temor, por desconocimiento o desmotivación, sin embargo es importante entenderlos plenamente. Por ello hemos abordado la capacidad de reconocer los diferentes tipos de datos, luego discutimos la lógica atrás de la captura de datos para convertirlo en algo útil, ahora debemos profundizar el tipo de variable que tenemos, sus alcances y sus limitaciones.

a. Variables Cualitativas

También se les conoce como variables de atributo y lo importante es entender que su naturaleza NO es numérica. Podemos estar hablando de los colores de pintura de determinada marca o una cualificación como “bueno”, “regular” y “malo”.

La importancia de entender la naturaleza de la variable es vital para las investigaciones, pues aunque decimos que NO son números, lo que suelen hacer los investigadores es convertirlas en números para poder analizar sus resultados. Por ejemplo: ¿le gusta el brócoli? Si o No.

El investigador puede convertir todos las respuestas Sí en un “1” y los No en “0” para luego encontrar la proporción de respuestas o para combinarla con otra variable, digamos género: Masculino = 1 y Femenino = 0; entonces con un ordenamiento de datos puedo ver en mi tabla a simple vista los hombres que respondieron Sí. En la próxima tabla puede verse el ejemplo de una base de datos ficticia sobre la pregunta ¿Te gusta el brócoli?

1	Género (0=mujer, 1=hombre)	Gusto por Brócoli (0=No, 1=Si)
Caso 1	0	1
Caso 2	0	1
Caso 3	0	0
Caso 4	0	1
Caso 5	0	0
Caso 6	0	0
Caso 7	0	0
Caso 8	0	1
Caso 9	0	1
Caso 10	0	1
Caso 11	0	1
Caso 12	0	1
Caso 13	0	1
Caso 14	0	0
Caso 15	0	1
Caso 16	0	1
Caso 17	0	0
Caso 18	1	0
Caso 19	1	1
Caso 20	1	1
Caso 21	1	0
Caso 22	1	0
Caso 23	1	0
Caso 24	1	0
Caso 25	1	1
Caso 26	1	0

Variables cualitativas nominales cuantificadas

Para profundizar en el tema de variables cualitativas, discutiremos ahora los dos tipos en los que pueden categorizarse estas variables.

Nominal: Son variables que identifican un grupo de pertenencia, por lo tanto únicamente permiten relaciones de igualdad o desigualdad entre sus elementos, en el caso de nuestro ejemplo anterior, las dos variables son nominales. Género puede ser masculino o femenino y aunque le podemos poner valores de números como “0” y “1” su naturaleza nos dice que JAMÁS podremos decir que una de las categorías es mejor o peor que otra, que una categoría es más o menos que la otra. De igual manera con el gusto por el Brócoli, es SI o es NO y ninguna respuesta tiene jerarquía sobre la otra.



Ordinal: Este tipo de variables cualitativas sirve para clasificar en categorías que indican un orden o que exigen una jerarquía entre sus datos.



Por naturaleza, los elementos en estas categorías guardan relaciones de “mayor qué” y “menor qué”. Para seguir nuestro ejemplo anterior, supongamos que también queremos saber si el nivel de estudios de la gente se relaciona con el consumo del brócoli, así que preguntamos su nivel de escolaridad con opciones “secundaria”, “profesorado”, “licenciatura” y “maestría”. Los datos aquí también pueden cuantificarse, pero tomando en cuenta la jerarquía proveniente de la naturaleza de los datos, porque profesorado tiene una mayor jerarquía que secundaria, por ejemplo. Entonces “Secundaria =0”, “profesorado = 1”, “licenciatura = 2” y “maestría =3).

2	Género (0=mujer, 1=hombre)	Gusto por Brócoli (0=No, 1=Si)	Grado de escolaridad (1=secundaria, 2=profesorado, 3=licenciatura, 4=maestría)
Caso 1	0	1	3
Caso 2	0	1	2
Caso 3	0	0	0
Caso 4	0	1	0
Caso 5	0	0	1
Caso 6	0	0	1
Caso 7	0	0	0
Caso 8	0	1	1
Caso 9	0	1	3
Caso 10	0	1	0
Caso 11	0	1	1

Ejemplo de Variables cualitativas ordinales

b. Variables Cuantitativas

La definición de variable cuantitativa es la más simple, debido a que cuando se habla de variables se tiende a pensar en este tipo: diremos que es la representación de la captura de datos numéricos (cuantitativos). La edad, la cantidad de goles, el registro de notas, el tiempo efectivo de clases, el número de visitas, la asistencia, los pagos realizados, todos ellos son ejemplos de variables cuantitativas.

Al igual que las variables cualitativas, las cuantitativas se pueden clasificar de acuerdo a su naturaleza:

Discreta: Es el tipo de variable cualitativa que adquiere valores dentro de un conjunto de datos finito, es decir que se limita a variar dentro del dato “no observable” y el “dato máximo posible”. En sentido práctico, por lo general estas variables se reconocen porque son números positivos enteros (de 1 en adelante) y al preguntarnos ¿podría tener valor 0? La respuesta es sí. La cantidad de personas en una oficina, el número de vehículos

registrados, la edad de las personas, el número de libros en la biblioteca, el total de visitas registradas durante un año laboral, todos son ejemplos de variables discretas.

En el ejemplo al que le estamos dando seguimiento, vamos ahora a revisar la variable discreta que el investigador agregó. Como ya sabía que tendría género, gusto o no gusto por brócol y nivel de estudios, ahora quiere saber qué tanto brócoli comen las personas que estudiará, él espera encontrar alguna relación importante ya que tiene la sospecha científica (hipótesis) que dice que las mujeres consumen más brócoli que los hombres y que la gente más educada consume más brócoli.

3	Género	Gusto por Brócoli	Grado de escolaridad	Frecuencia de consumo de Brócoli
	{0=mujer, 1=hombre}	{0=No, 1=Si}	1=secundaria, 2= profesorado, 3=licenciatura, 4= maestría	# de veces por semana
Caso 1	0	1	3	1
Caso 2	0	1	2	3
Caso 3	0	0	0	2
Caso 4	0	1	0	1
Caso 5	0	0	1	1
Caso 6	0	0	1	1
Caso 7	0	0	0	1
Caso 8	0	1	1	1
Caso 9	0	1	3	1
Caso 10	0	1	0	3
Caso 11	0	1	1	2

Ejemplo de variable discreta

Continua: Se refiere a las variables que adquieren cualquier valor de un conjunto de datos con características de intervalos y por lo tanto, los valores intermedios pueden registrarse. Por lo general se distinguirán porque permiten decimales y negativos y a la pregunta ¿el dato registrado depende de un instrumento de medición? Si la respuesta es sí, por lo general es una variable continua ya que su resultado podría variar si se toma con un instrumento diferente. Las notas en un examen, el peso y estatura de las personas.



En el ejemplo que estamos construyendo, el investigador también buscará relación entre el peso de las personas y el consumo de brócoli, así que también recopiló el peso de los participantes y lo registró en una variable cuantitativa discreta. Ahora, a practicar entre variables cualitativas y cuantitativas...

4	Género	Gusto por Brócoli	Grado de escolaridad	Frecuencia de consumo de Brócoli	Peso
	(0=mujer, 1=hombre)	(0=No, 1=Si)	(1=secundaria, 2= profesorado, 3=licenciatura, 4= maestría)	# de veces por semana	lbs.
Caso 1	0	1	3	1	123.73
Caso 2	0	1	2	3	145.00
Caso 3	0	0	0	2	143.28
Caso 4	0	1	0	1	133.42
Caso 5	0	0	1	1	121.91
Caso 6	0	0	1	1	131.26
Caso 7	0	0	0	1	159.00
Caso 8	0	1	1	1	140.73
Caso 9	0	1	3	1	120.44
Caso 10	0	1	0	3	177.61
Caso 11	0	1	1	2	126.90

Ejemplo de variable discreta

c. Variables Aleatorias

En muchas ocasiones las variables no pueden conocerse sino hasta después de un suceso o experimento y por lo tanto se les conoce como aleatorias ya que su valor dependerá del resultado del experimento. Así pues, puede pensarse que el valor de dicha variable es desconocido hasta que se determine su resultado. Probablemente el ejemplo más simple sea el de una variable que adquiere el valor que sale al tirar el dado, pero también puede ser el valor de la temperatura en la Ciudad de Guatemala tomada a las 11:32 am todos los días del año en donde deberemos esperar cada día para revisar el termómetro y saber cuál es el valor de la variable.



Las variables aleatorias pueden ser Discretas o Continuas bajo la misma definición que ya hemos elaborado en el tema anterior.

En el ejemplo que estamos elaborando, vamos a agregar la hora de la entrevista como una variable aleatoria que nos servirá de referencia y de control para el estudio que realizamos.

5	Género	Gusto por Brócoli	Grado de escolaridad	Frecuencia de consumo de Brócoli	Peso	Hora de la entrevista
	{0=traje, 1=hombre}	{0=No, 1=Si}	{1=secundaria, 2= profesorado, 3=universitaria, 4=maestría}	# de veces por semana	lbs.	
Caso 1	0	1	3	1	123.73	12:30
Caso 2	0	1	2	3	145	11:00
Caso 3	0	0	0	2	143.28	13:45
Caso 4	0	1	0	1	133.42	8:00
Caso 5	0	0	1	1	121.91	19:00
Caso 6	0	0	1	1	131.26	15:30
Caso 7	0	0	0	1	159	9:30
Caso 8	0	1	1	1	140.73	21:00
Caso 9	0	1	3	1	120.44	10:00
Caso 10	0	1	0	3	177.61	20:30
Caso 11	0	1	1	2	126.9	8:30

Ejemplo de Variable Aleatoria

d. Variables Dependientes e Independientes

La clasificación de variables de acuerdo a la relación entre ellas en una investigación se puede entender desde dos perspectivas opuestas, una causa y un efecto. Puede pensarse también como un sistema de sobrepesos en una balanza que produce que el cambio en un lado de la balanza produzca un cambio en el otro lado también.

La mejor manera de explicarlo y definirlo es iniciando por la variable independiente.

Variable independiente: Se entiende como “independiente” a las variables que tienen la potestad de influir, afectar o incidir a otras variables. Se le llama independiente porque NO depende de otros factores para existir dentro del estudio. En el ejemplo que estamos construyendo, podemos pensar en el “gusto por el brócoli” y la “cantidad de días semanales en los cuales se consume brócoli como variables independientes.



6



Brocoli

Variable dependiente: Es la variable que representa el atributo, propiedad o fenómeno que estamos investigando. Es llamada dependiente porque depende de las demás variables que recopilamos (las independientes). Cuando realizamos un estudio, es recomendable que la variable dependiente esté en el título.



Veamos nuestro ejemplo. ¿Ya identificaron la variable dependiente? Sí! Es el peso de las personas. Imaginemos el título de la investigación: Relación entre el consumo de brócoli y el peso de las personas en Santiago Atitlán.

En el estudio, tengo el interés de determinar si “el gusto por el brócoli” y “la frecuencia de consumo de brócoli” (dos variables independientes) afectan el peso de las personas (variable dependiente).

Otra forma de identificar el tipo de variable es preguntándose ¿cuál de estas variables puede manipular un investigador para alterar el resultado? En nuestro caso, el investigador podría pedir a la gente que coma más brócoli o menos brócoli y luego pesarlos, así que estamos viendo su intención para determinar si el peso DEPENDE del consumo de brócoli.



Comprobación del capítulo VI: Tipos de variables y tipo de escalas de medida

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo VI

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo VII

Distribución de frecuencias



Distribución de frecuencias

La distribución de frecuencias es una herramienta útil en estadística descriptiva, nos permite agrupar los datos en categorías de interés. Esto permite observar los datos de una forma reducida y fácil, lo que ayuda a tomar decisiones con los datos.

a. ¿Qué es la distribución de frecuencias?

Diremos que la distribución de frecuencias es una forma de agrupación de datos en categorías que son mutuamente excluyentes, es decir, que cada dato puede pertenecer única y exclusivamente a una categoría.



b. Utilización de la distribución de frecuencias

Elaboremos un ejemplo: Supongamos que entrevistamos a 500 niñas de una escuela y les preguntamos cuál es su número favorito entre 1 y 10. Tendremos entonces un listado de 500 datos entre 1 10: algo así

1

10	9	3	8	1	4	2	3	2	5
2	10	2	4	1	1	8	9	10	1
5	3	7	8	6	9	9	1	5	5
6	66	8	3	7	1	10	8	9	7
5	5	5	8	8	10	5	5	1	5
7	5	4	4	8	4	7	4	7	8
4	9	4	8	6	9	4	1	6	7
2	10	2	3	7	7	1	6	5	4
9	1	0	3	10	1	5	6	4	9
8	7	1	1	8	5	1	7	1	4
8	9	2	7	1	9	3	1	3	1
4	10	8	2	5	10	1	9	9	8
10	1	0	9	6	4	3	4	3	5
3	7	3	4	1	4	7	3	10	6
2	9	2	8	10	2	1	3	9	4
5	10	7	7	7	3	1	10	5	7
8	8	10	8	1	8	7	5	1	5
2	8	3	9	8	6	3	4	5	6
5	1	7	8	8	1	10	8	7	3
8	1	0	8	6	2	1	3	1	9
4	2	2	1	2	3	6	10	1	2
1	6	1	1	1	1	8	1	6	4
2	6	7	3	3	6	0	7	6	7
6	7	10	3	8	9	6	4	6	9
8	9	0	9	2	8	2	2	1	9
3	9	4	6	4	8	2	3	9	7
3	4	3	1	1	5	1	9	1	10
9	9	2	7	7	8	2	7	3	5
6	9	8	2	8	10	10	9	6	5
6	1	0	10	9	1	8	5	1	8
5	9	4	3	3	2	8	9	8	1
10	6	3	9	6	8	7	1	7	3
9	6	3	6	5	9	3	7	10	8
9	9	7	2	8	9	5	1	7	6
1	1	9	1	5	3	1	9	10	8
8	7	7	4	1	10	8	5	10	5
1	9	8	1	10	9	7	2	8	2
8	7	3	5	7	10	0	1	3	4
6	8	5	3	2	8	6	1	8	10
10	9	0	10	2	5	7	5	6	1
5	6	0	5	10	5	7	3	10	8
8	1	6	8	7	10	1	5	2	3
5	10	0	8	10	7	3	4	8	1
9	9	3	7	2	1	8	10	2	10
5	2	3	1	5	2	5	7	4	9
6	1	0	2	7	3	7	3	3	7
10	2	3	9	6	1	9	3	9	1
7	1	5	5	2	9	4	4	4	3
9	8	1	3	1	6	3	1	1	3
10	1	9	1	5	2	4	10	1	8

Distribución de frecuencias de 500 datos

Podemos darnos cuenta que en este caso es muy complicado tratar de encontrar sentido a los datos, por lo tanto puedo crear categorías que me permitan facilitar su interpretación. Así que podemos crear categorías, por ejemplo categoría 1: los datos menores o iguales que 3; categoría 2: los datos mayores que tres y menores o iguales que 6; y categoría 3 los datos mayores que 6. El resultado de este ejercicio es la FRECUENCIA ABSOLUTA y se representa como f_i . La sumatoria se representa por la letra Σ griega sigma y el resultado de la sumatoria se representa como N que debe ser el total de niñas entrevistadas.

2	Frecuencia Absoluta f_i
Menor o igual que 3	245
Mayor que 3 menor o igual que 6	149
Mayor que 6	106
Σ	500

Frecuencia absoluta



Nota:

Es muy importante verificar que la suma de frecuencias en cada categoría sume el total de casos que tenemos disponibles para asegurarnos que contabilizamos todos los datos y que ninguno está duplicado.

Además de la frecuencia absoluta, se puede calcular la Frecuencia Relativa que hace la comparación entre categorías más apropiadas pues se representa como porcentaje de cada categoría de estudio. Se obtiene al dividir cada f_i entre N y puede representarse como porcentaje al multiplicarlo por 100. Se representa por n_i .



3	Frecuencia Absoluta f_i	Frecuencia Relativa n_i
Menor o igual que 3	245	0.49
Mayor que 3 menor o igual que 6	149	0.298
Mayor que 6	106	0.212
Σ	500	1

Frecuencia absoluta y relativa

Es muy importante que la frecuencia relativa siempre sume 1 si se expresa como proporción y que sume 100 si se expresa como porcentaje.

La siguiente adición a la tabla de frecuencias es la columna de frecuencias acumuladas. Se refiere a la acumulación de datos iguales o inferiores a la categoría deseada. Se obtiene sumando la f_i previa a la categoría que analizamos, así que la primera categoría es igual a su f_i la segunda será el f_i anterior (245) más la f_i propia (149) = 394. La última será (394+106=500).

4	Frecuencia Absoluta f_i	Frecuencia Relativa n_i	Frecuencia Acumulada F_i
Menor o igual que 3	245	0.49	245
Mayor que 3 menor o igual que 6	149	0.298	394
Mayor que 6	106	0.212	500
	500	1	

Frecuencia absoluta, relativa y acumulada

Es importante verificar que la frecuencia acumulada de la última categoría sume N , en este caso 500.



Por último, la tabla de frecuencias se completa con las Frecuencias Relativas Acumuladas, se representarán por N_i . Se aplica la misma lógica que la acumulación de frecuencias, se toman las n_i y se le suma la n_i anterior para evidenciar la acumulación.

5	Frecuencia Absoluta f_i	Frecuencia Relativa n_i	Frecuencia Acumulada F_i	Frecuencia relativa acumulada N_i
Menor o igual que 3	245	0.49	245	0.49
Mayor que 3 menor o igual que 6	149	0.298	394	0.788
Mayor que 6	106	0.212	500	1
	500	1		

Frecuencia Absoluta, Relativa, Acumulada y Relativa Acumulada

Es importante que se verifique que la frecuencia relativa acumulada sume 1 en la última categoría si se usan proporciones o que sume 100 si se usan porcentajes.



Atención

Si desea aprender a hacer las distribuciones de frecuencias apoyándose en Microsoft Excel, en este enlace puede encontrar una forma de hacerlo:

<https://exceltotal.com/distribucion-de-frecuencias-en-excel/>

c. Interpretar la distribución de frecuencias

Utilizaremos la tabla anterior para ejemplificar los usos que se pueden dar a la distribución de frecuencias.

En primer lugar, podemos saber cómo están organizados los datos de un solo golpe de vista, sabemos que el número favorito de la mayoría de las 500 niñas de la escuela está entre 1 y 3. Sabemos que el número menos favorito es mayor que 6.

Podemos decir que casi la mitad de las niñas (49%) prefieren un número menor que 3 (lo sabemos porque si se multiplica el ni de la categoría por 100 obtenemos el porcentaje relativo).

Podemos además decir que un poco más de 3 de cada 4 niñas prefieren números entre 1 y 6 porque en la Frecuencia Acumulada, la preferencia nos dice que el 78% de niñas escogen esas dos categorías.



Comprobación del capítulo VII: Distribución de frecuencias

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo VII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo VIII

Representar tablas de datos



Representar tablas de datos

La representación de datos a través de tablas es, sin lugar a dudas, una tarea vital de todo investigador y analista de datos. Una correcta representación garantiza la interpretación adecuada y genera valor al trabajo científico. Por el otro lado, una mala representación de los datos, suele hacerse por dos razones: 1) Por desconocimiento y 2) por mala fe, con el fin de manipular los resultados.

a. Gráficas de barras

Las gráficas de barras representan una de las formas más básicas de representación de datos y su fácil interpretación han sido factores importantes que para su popularidad aumente en ámbitos académicos, políticos, económicos, sociales, etc.

También se le conoce como diagrama de barras o diagrama de columnas y su uso es para representar frecuencias y frecuencias relativas.



Gráficas de barras

Barras Simples: su uso principal es la representación de variables categóricas cualitativas, ya que cada barra representará una categoría. El trazo de cada barra debe ser independiente de la siguiente para mantener la separación visual.



El ejemplo siguiente muestra una gráfica de registro de asistencia en una semana.

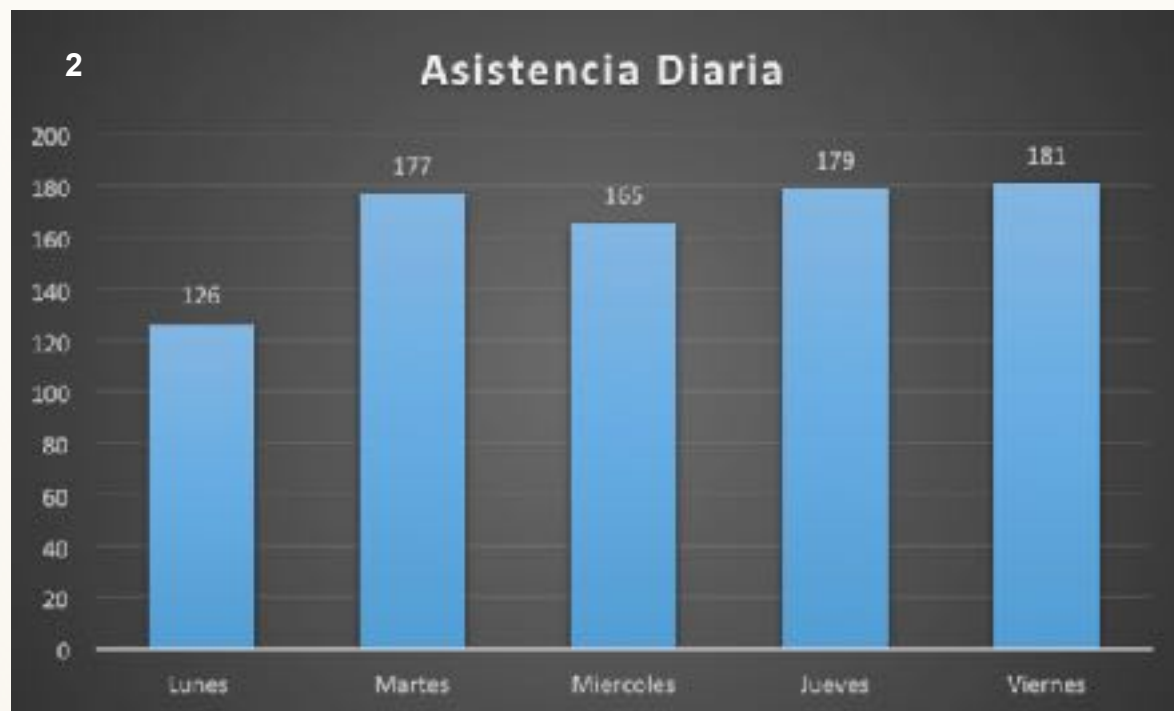


Diagrama de barras simple

Barras Compuestas: Es un tipo de gráfica que permite hacer comparaciones más específicas. En realidad sirven para mostrar las ocasiones en las que una categoría tiene subdivisiones, así que su cualidad principal es que la capacidad de hacer comparaciones a lo interno de las categorías y entre categorías en una misma gráfica. En el ejemplo vemos la asistencia en el primer semestre comparando niños y niñas.

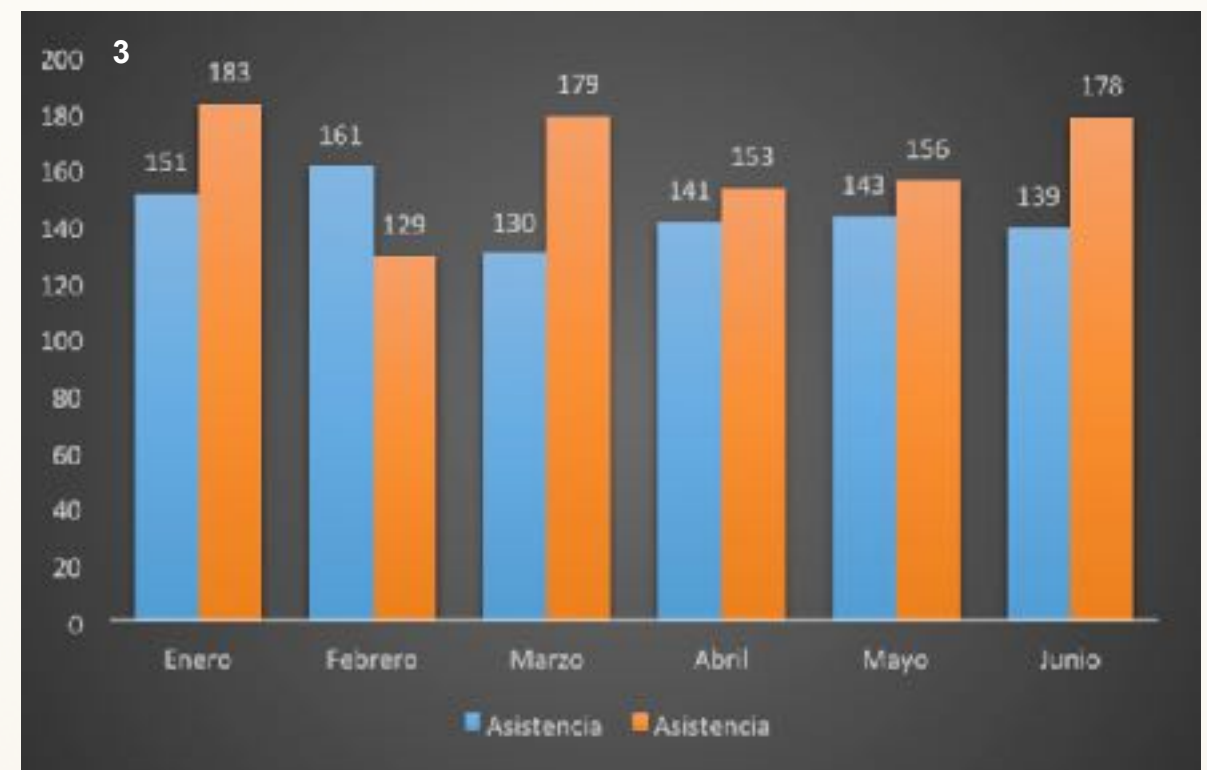


Diagrama de barras compuestas

Horizontales: Por lo general se usan para diferenciar las clases de datos categóricos, se pueden usar para representar frecuencias absolutas, frecuencias relativas, proporciones y porcentajes.



En el caso de las barras horizontales, se recomienda organizarlas siempre de mayor a menor o menor a mayor. En el ejemplo el administrador de un colegio necesita saber cómo se ha comportado la matrícula para mostrarle a la junta directiva de la institución.

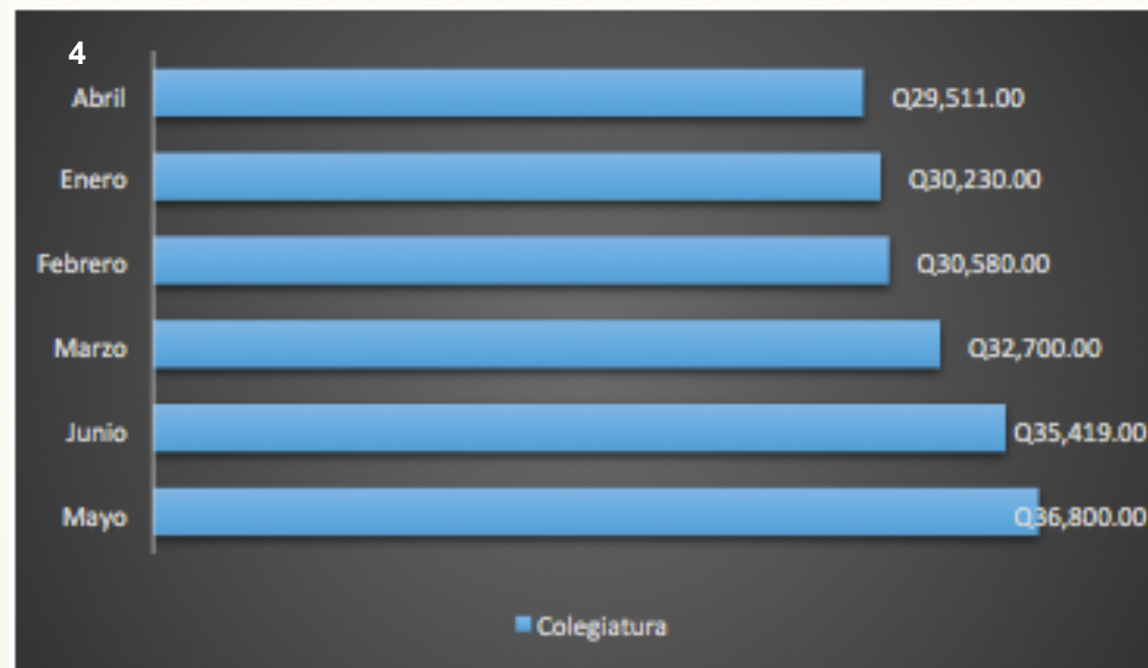


Diagrama de barras horizontales

b. Diagrama de Sectores o Circular

Se refiere a la famosa gráfica de pie o pastel. Su representación gráfica es muy útil porque permite mostrar las partes de un todo en tamaño relativo.



En el ejemplo veremos un posible uso para este tipo de diagrama, en donde se presenta a un grupo de docentes de primer ingreso, la distribución del tiempo de instrucción que se espera en la institución educativa para lograr el perfil de egreso propuesto.

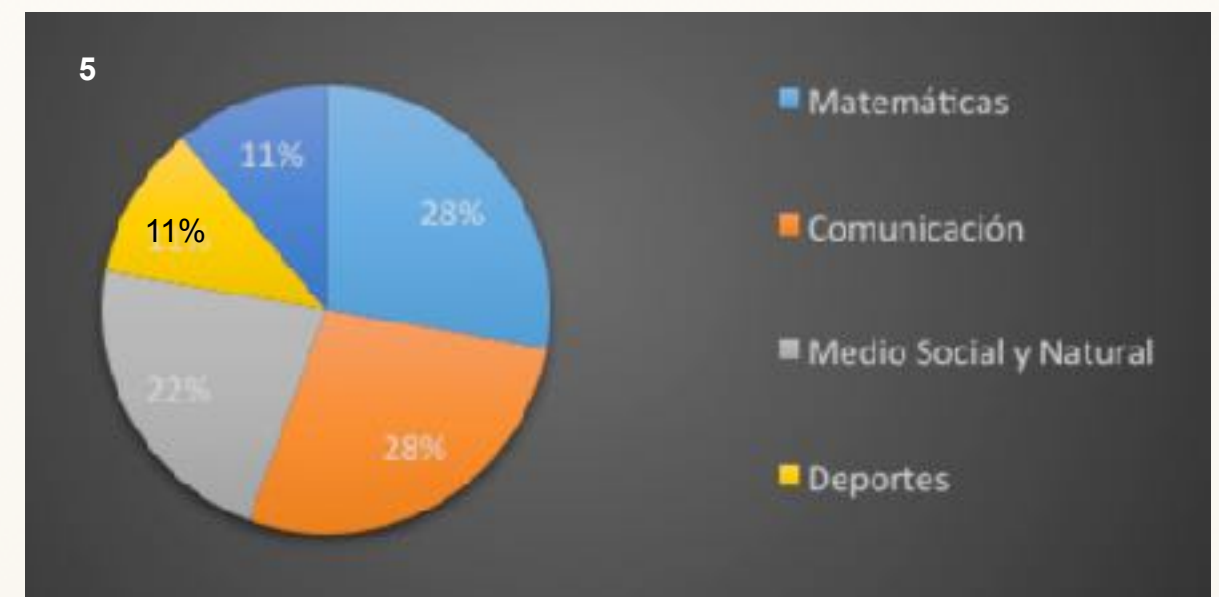


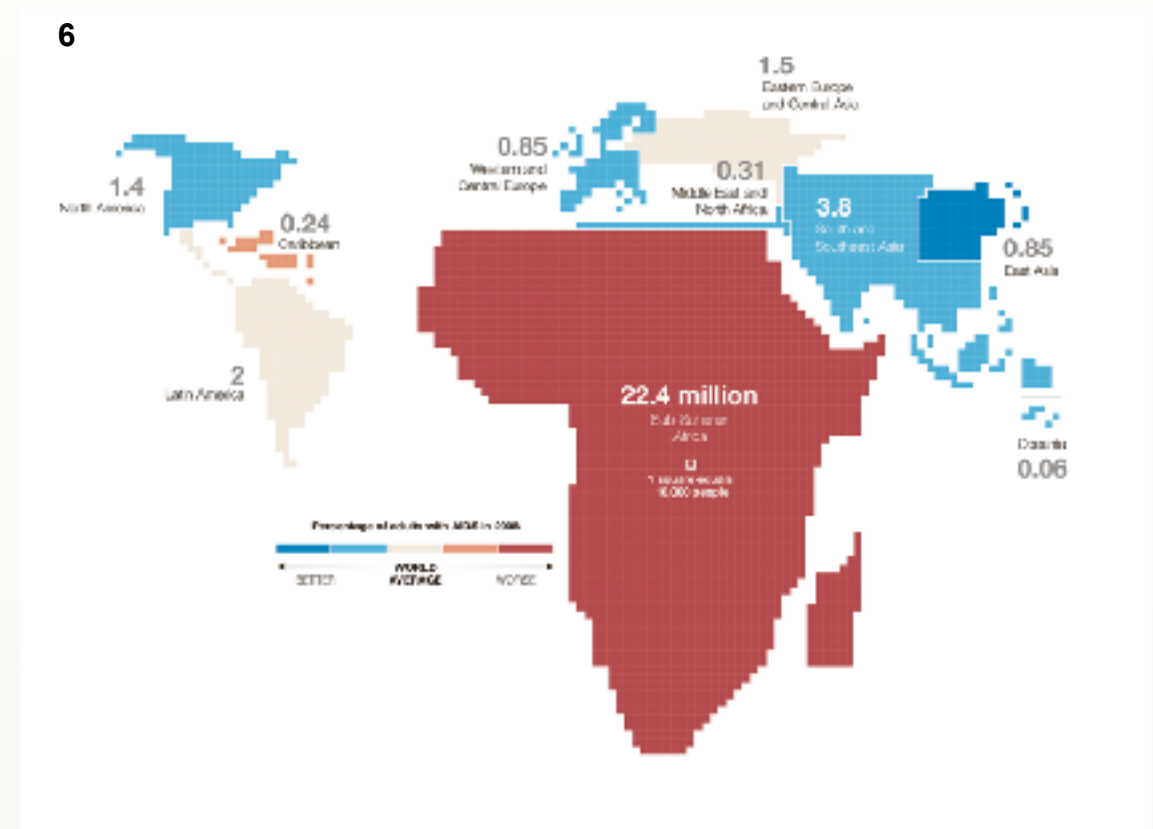
Diagrama de barras circular

Los últimos dos tipos de gráfica que se discutirán en este apartado son menos frecuentes y de más compleja elaboración. Normalmente requieren de programas especializados para generarlos ya que los paquetes más comunes como Microsoft Excel no pueden hacerlos. Vamos a hablar de Cartogramas y Pictogramas para conocer de ellos y poder interpretarlos.

c. Cartogramas

Es una gráfica que parte de un mapa y muestra los datos asociados a la región mediante el tamaño de sus unidades. Suelen usarse para mostrar tamaños de poblaciones en países, tamaño económico de regiones, casos de alguna enfermedad, cantidad de conexiones de internet, en fin, sus usos son múltiples. En resumen, podemos decir que un cartograma es un mapa distorsionado, con información.

Consulte el siguiente Cartograma y agrande, según le interesa un área.



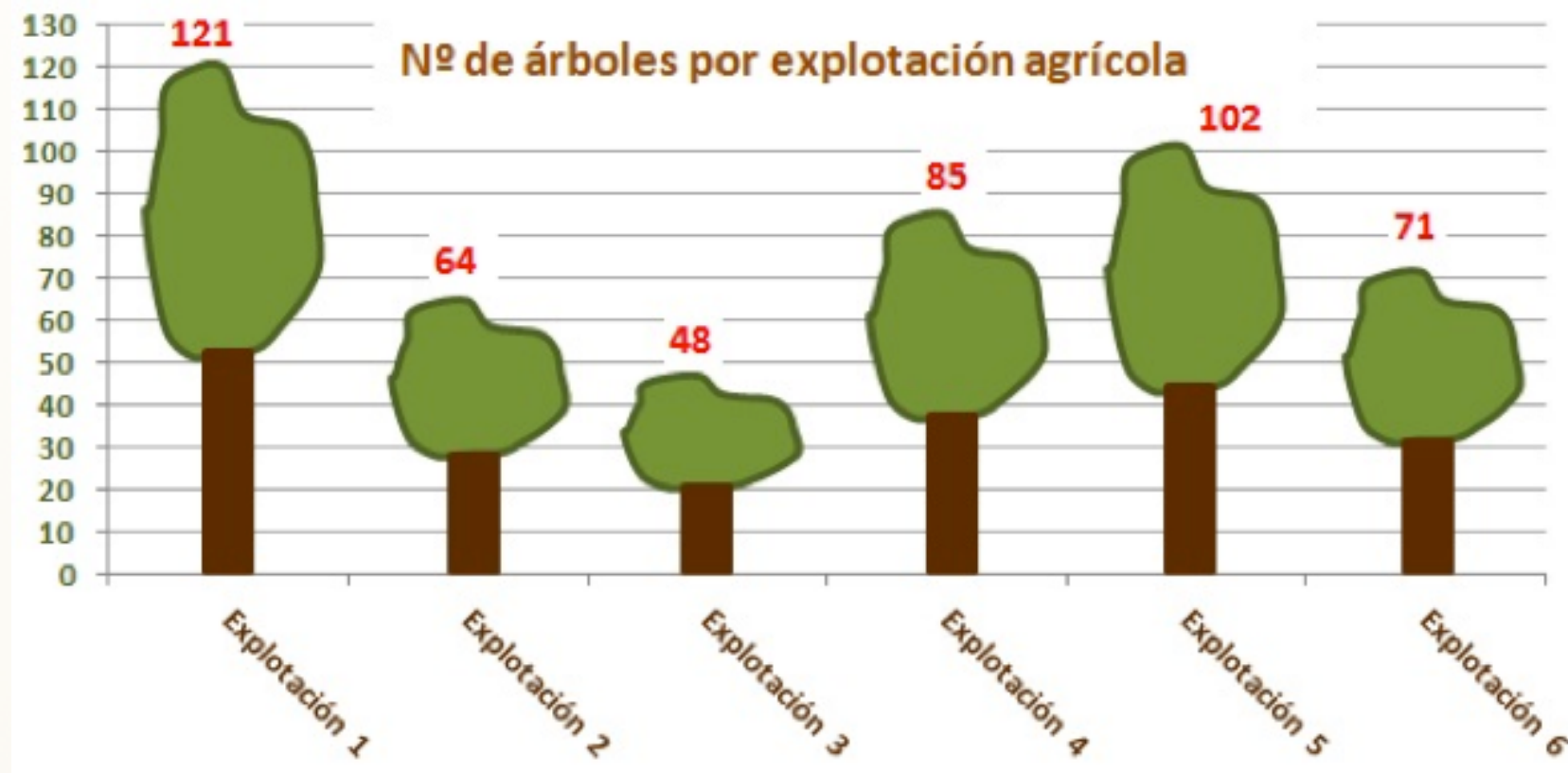
Ejemplo de cartograma

d. Pictograma

Es una forma más gráfica y puede considerarse informal para representar frecuencias absolutas o relativas. Muestra las relaciones mediante dibujos alusivos a la característica estudiada.



7

*Ejemplo de pictograma*

Con la revisión de los pictogramas finalizamos la unidad, en la próxima unidad revisaremos más a profundidad los usos correctos de la representación de la información y la correcta interpretación de los mismos, además de aprender a elaborar cada tipo de representación en programas computacionales.



Comprobación del capítulo VIII: Representar tablas de datos

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo VIII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Unidad III

Interpretación de tablas y gráficas



Capítulo IX

Representación de tablas de frecuencias



Representación de tablas de frecuencias

La unidad se dedica a la elaboración de gráficas en Excel y la correcta interpretación de la información a partir de tablas y las gráficas.



Gráficas estadísticas

En esta unidad profundizaremos y ejercitaremos sobre el proceso de elaboración de gráficas, principalmente en Excel y más importante aún, abordaremos el tema de cómo llegar a su correcta interpretación.

La representación de datos de forma gráfica es una herramienta que nos permite presentar datos de manera más sencilla y permite llegar a conclusiones más fáciles de entender. No es una tarea compleja pero requiere de un proceso de toma de decisiones que permite lograr la representación adecuada de los datos.

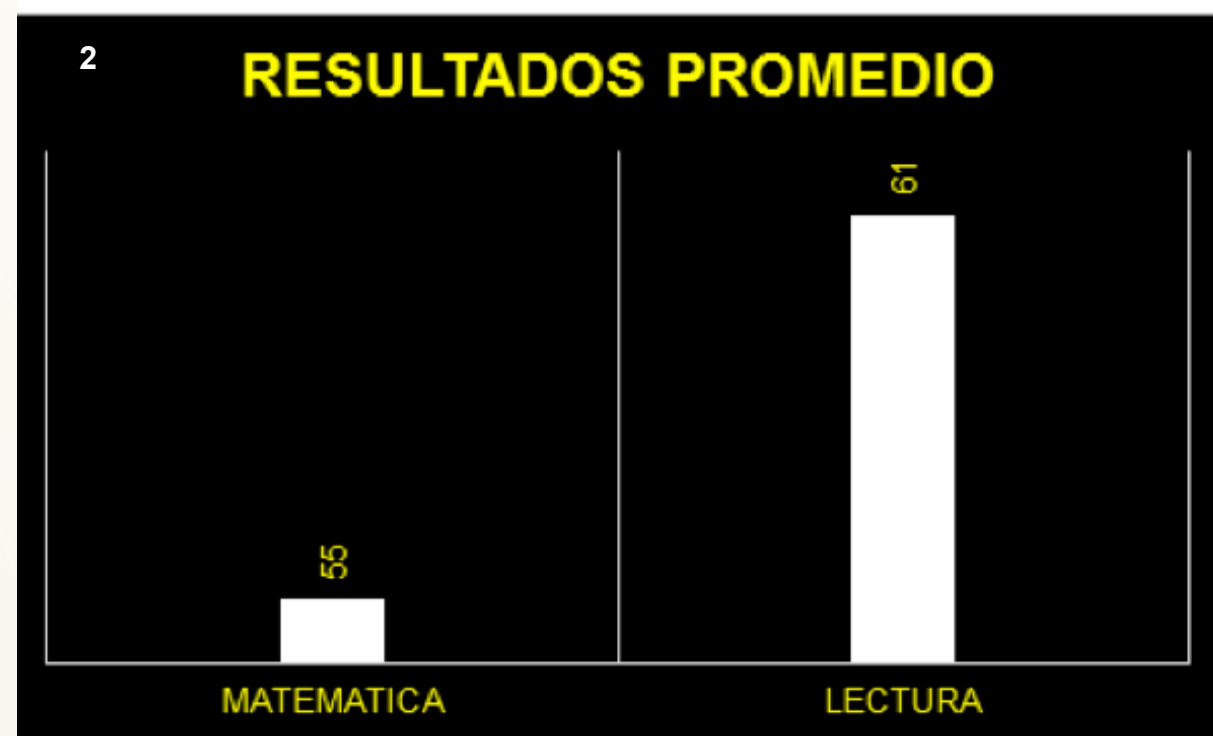


En las ciencias sociales a esta tarea se le considera “*intelectualmente respetable*” y se ha convertido en una disciplina denominada: The visual display of quantitative information, que en español significa: La representación visual de información cuantitativa.

a. Importancia de la correcta representación de los datos

Para mostrar la importancia de mostrar adecuadamente los datos recabados hablaremos de un ejemplo que me contó una colega y nos permitirá evidenciar el tema en discusión:

“Discutíamos con un grupo de docentes sobre los resultados que sus estudiantes habían alcanzado en unas pruebas de matemática y lectura en un ciclo escolar. Los resultados estaban presentados en una gráfica como la siguiente:

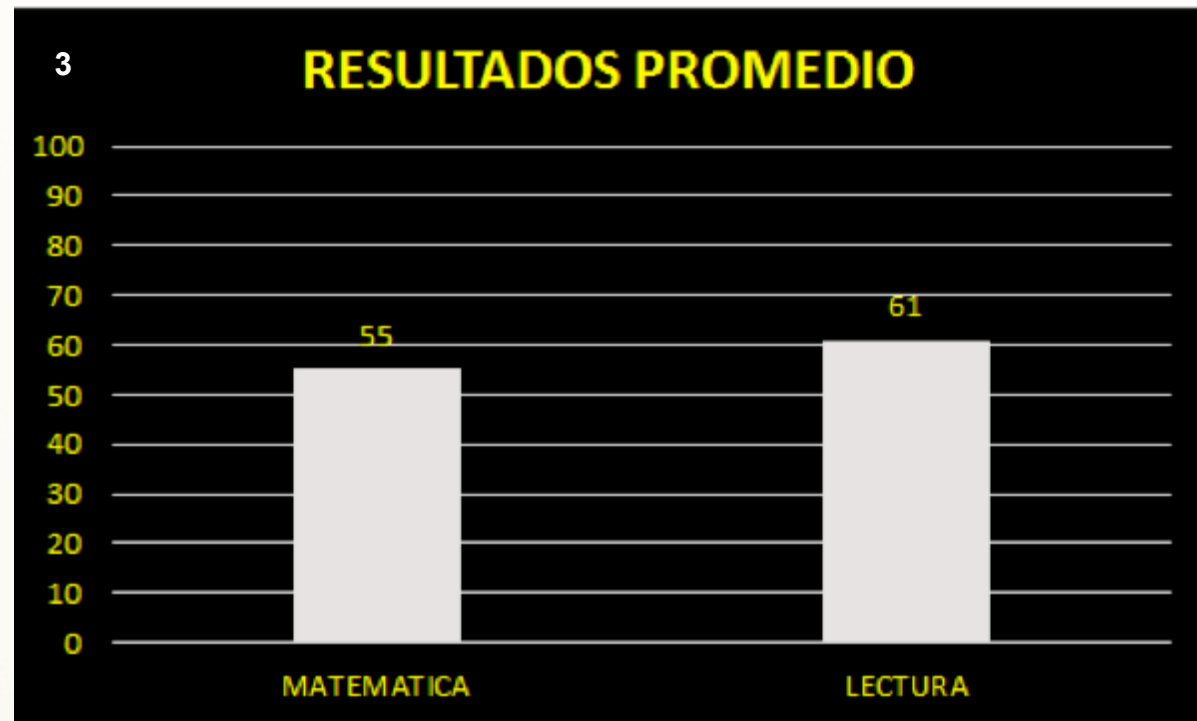


Resultados de estudiantes de primaria. Gráfica de barras 1.

A simple vista, la gráfica mostraba un resultado muy bajo en el área de Matemática, comparado con lo alcanzado en Lectura. Alrededor de esta imagen se generó gran parte de la discusión docente donde cada uno planteaba razones por la cuales los estudiantes obtenían notas bajas en el área de Matemática, soluciones para mejorar dichos resultados y experiencias negativas y positivas que habían tenido durante el año escolar. La discusión pretendía analizar resultados y proponer medidas de mejora. Luego se convirtió en una evaluación al proceso de enseñanza y aprendizaje de la Matemática, sin darse cuenta que la diferencia entre los resultados de las áreas evaluadas, no era tan dramática como la gráfica lo presentaba y que las propuestas de mejora debían de diseñarse y aplicarse en las dos áreas evaluadas”.

Así que una mala gráfica hizo que un grupo de docentes perdiera más de 2 horas en una discusión que no tenía mucho sentido. Así que vale la pena preguntarse.

¿Qué cambios se podrían hacer a la gráfica anterior para lograr una mejor representación de los datos?



Resultados de estudiantes de primaria. Gráfica de barras 2.

Ahora veamos la gráfica presentada de otra forma, podemos notar grandes diferencias entre las dos formas de representar los datos y está de más decirlo, pero esta forma es la correcta.

b. Utilidad de la representación de datos

Para obligarnos a hacer buenas manifestaciones de los datos a través de gráficas, podemos aprovechar las 7 reglas para generar un documento guía para nuestro trabajo.

- Ante todo, revela los datos.
- Induce a pensar sobre la sustancia.
- Anima a comparar unos datos con otros.
- Evita distorsionar lo que dicen los datos: no miente.
- Es denso en datos: mucha cifra, poco espacio, poca tinta.
- Sirve un propósito claro: describir, explorar, tabular datos, etc.
- Existe íntimamente integrado con el texto y las piezas que acompaña.

Vamos a utilizar esta lista de cotejo para cada gráfica que produzcamos de ahora en adelante, podemos estar seguros que si lo hacemos, cuando presentemos datos las personas van a comprender nuestras intenciones.

4

NO.	CONDICION	SI	NO
1	¿Revela los datos?		
2	¿Induce a pensar sobre el tema?		
3	¿Anima a comparar unos datos con otros?		
4	¿Evita distorsionar lo que dicen los datos?		
5	¿Es densa en datos?		
6	¿Sirve un propósito claro?		
7	¿Está íntimamente integrado con el texto y las piezas que acompaña?		

Lista de cotejo, 7 reglas para gráficas

Hay un último punto que es importante mencionar con el uso de las gráficas. Cada vez que vayamos a utilizar una debemos preguntarnos: ¿la gráfica va a simplificar el texto? En otras palabras, es un error muy común que pongamos una gráfica y luego la expliquemos en texto, si hacemos esto estamos redundando, lo correcto es poner una gráfica para evitar escribir el texto y dedicarle nuestros párrafos a la interpretación de los resultados y sus implicaciones.



Representaciones de variables

Como primer punto vamos a discutir sobre los elementos que una gráfica debe llevar adentro de un informe: Título, Número y Fuente.

Título: es la identificación para la gráfica que presentamos, debe indicar exactamente de qué se tratan los datos de la forma más resumida posible.

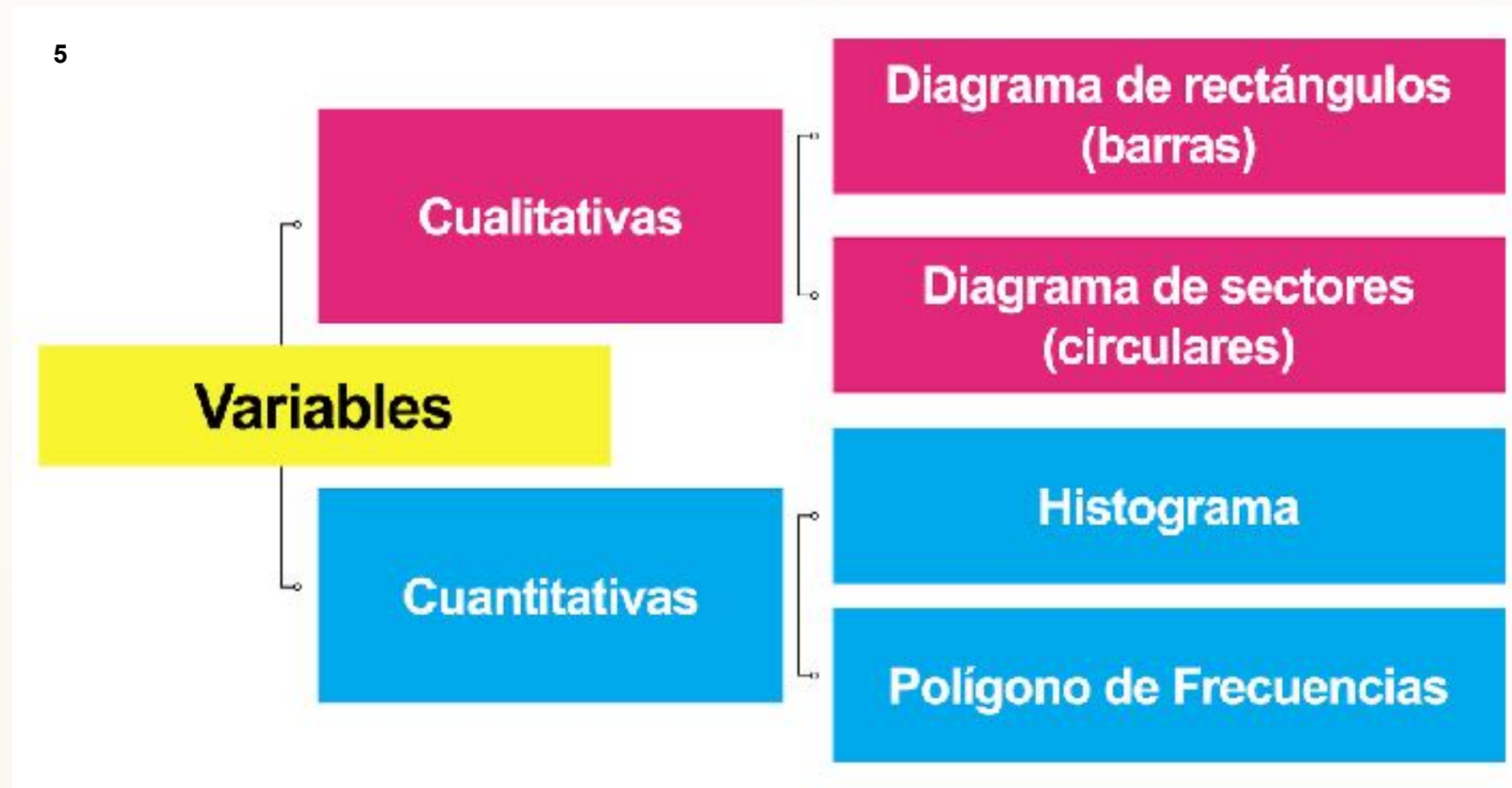


Número: es el correlativo correspondiente a la gráfica que estamos usando. En los programas de edición de texto como Microsoft Word hay opciones automáticas para numerar las gráficas y evitar problemas. Es importante usarlas!

Fuente: es el origen de la gráfica que se propone, si son gráficos creados por tu persona la fuente es “elaboración propia del autor” muchas personas son más detallistas “elaboración propia con base de datos de Ministerio de Educación 2015”, las dos opciones son válidas, depende más del estilo del autor.



En la siguiente ilustración podemos ver un diagrama de orientación para determinar qué gráfica se utiliza. Es importante mencionar que allí veremos la representación más simple, pues hay datos agrupados y datos sin agrupar que hacen que existan barras simples, compuestas y apiladas por ejemplo.



Decisiones sobre tipo de gráficos

Se puede resumir en la tabla que se presenta a continuación, una orientación al uso adecuado de las gráficas:

6	GRÁFICA	USO	TIPO DE VARIABLE
	Gráfica de Barras	Representar Frecuencias	Cualitativa o Discreta
	Gráfica de Barras (Horizontal o Vertical)	Comparar magnitudes y ver la evolución en el tiempo	Cuantitativa y Ordinal
	Gráfica de Barras (Histograma)		Cuantitativa continua
	Gráfico de líneas	Cambios en el tiempo o tendencias temporales	
	Gráfica de sectores		Cualitativa o Discreta
	Pictograma		Cualitativa o Discreta
	Gráfico de dispersión	Relación entre dos variables	

Guía para el uso de gráficos



Comprobación del capítulo IX: Representación de tablas de frecuencias

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo IX

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo X

Representación de variables cualitativas



Representación de variables cualitativas

Ahora que empezaremos a ver cada una de los tipos de gráficos o diagramas empezaremos a notar la importancia del aprendizaje en las unidades anteriores. Vamos a necesitar recordar y comprender el significado de las frecuencias absolutas y relativas, por ejemplo.

1



Diagramas estilizados

a. Diagramas de rectángulo (barras)

Como mencionaba, es importante que recordemos algunos conceptos para no equivocarnos al usar los gráficos. En el caso de las “barras” estoy seguro que todos las hemos visto y las hemos usado más de alguna vez, así que en esta ocasión será importante que pongamos atención al repaso de estos conceptos:

Existen variables cualitativas y cuantitativas.

Las cuantitativas pueden ser continuas (cuando pueden tomar cualquier valor en un intervalo, como peso, notas, etc.)



también pueden ser discretas (solamente toman algunos valores de una distribución, como cantidad de alumnos en el aula, cantidad de hijos de una persona, número de visitas al médico en el último año, etc.).

Las cualitativas pueden ser nominales (cuando no tienen ningún orden



específico, como la respuesta ante la pregunta: usted estudia? Sí o No) y las variables cualitativas pueden ser ordinales

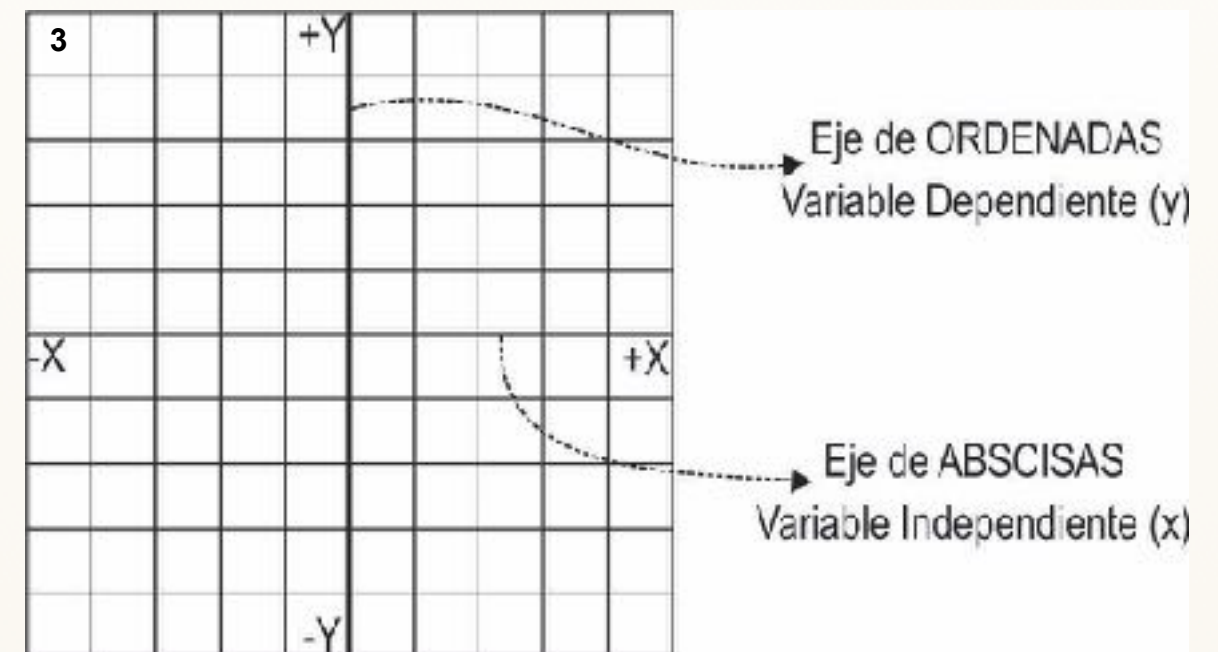
(cuando si existe un orden para las respuestas, por ejemplo: cuál es el último año que usted ha aprobado en la escuela?)

Los diagramas de barras se utilizan para representar variables cualitativas ordinales. Es incorrecto utilizarlas para representar variables cualitativas nominales.



Diagrama de barras

Su construcción requiere que en el eje X se coloquen las categorías, allí se les representa con su nombre y en el eje Y se levantan las barras cuya altura representará la frecuencia de la misma.



Ejes Cartesianos



Ejemplo de diagrama de barras.

Se puede usar un diagrama de barras también para representar variables cuantitativas discretas.

Para elaborar una gráfica de barras en Excel es necesario crear dos columnas de información, en la primera colocamos las categorías y en la segunda las frecuencias relativas o absolutas:



Nota:

Si trabajamos frecuencias relativas, nos debemos asegurar que SIEMPRE sumen 100%.

	A	B	C	D	E
1					
2					
3					
4					
5			Años de Estudio	Frecuencia relativa	
6			Primaria	7%	
7			Básico	14%	
8			Secundaria	19%	
9			Técnico Universitario	48%	
10			Licenciatura	8%	
11			Maestría	3%	
12			Doctorado	1%	

Elaboración de gráfica de barras simple. Paso 1

Con el mouse seleccionamos las columnas y filas con la información que deseamos graficar, hacemos clic en la función INSERTAR y allí nos aparecerá la opción de gráficas. Es importante que tengamos en cuenta que dependiendo de la versión de Excel que tengamos instalada el menú puede variar un poco.

En el menú de gráficos que aparezca debemos seleccionar la gráfica de barras o columnas para que el programa genere nuestro diagrama seleccionado.

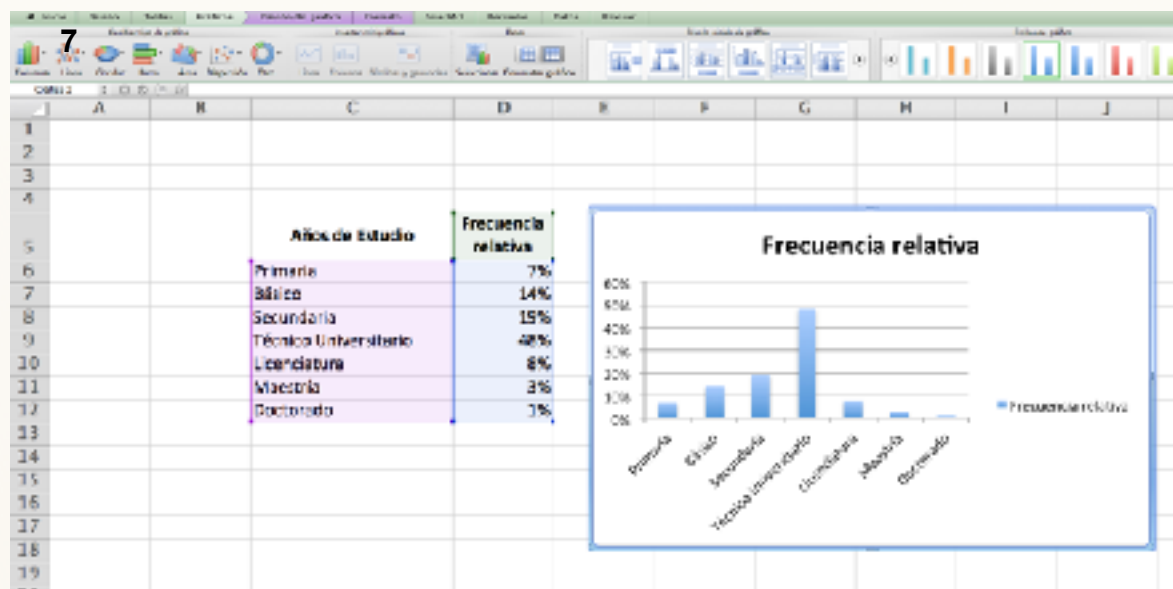
[illegible]

Elaboración de gráfica de barras simple. Paso 2

Una vez generado el gráfico podemos editarlo de acuerdo a nuestros intereses, empezando por el nombre al hacer doble clic sobre el nombre que aparece y sustituyendo el texto por el título que deseamos.

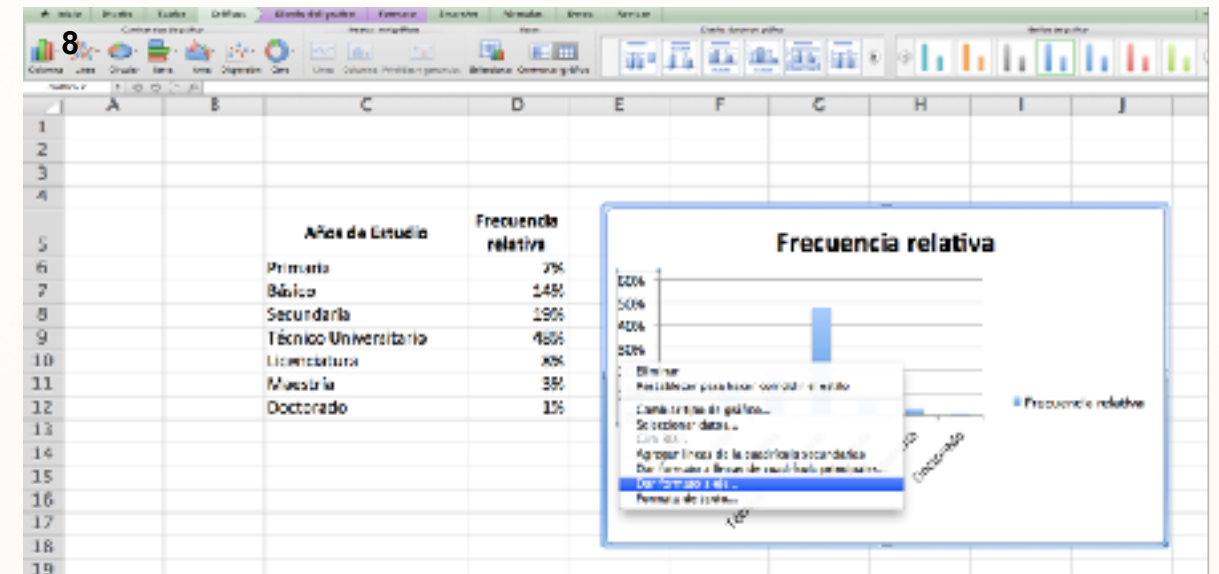
Una modificación que se puede hacer y en ocasiones es necesario realizar, es la modificación del eje. En la introducción al capítulo vimos un ejemplo que modificaba la apariencia de los resultados simplemente porque el eje Y no era el adecuado. Para agregar una situación de esa naturaleza haremos los siguientes pasos:

Primero, marcamos el eje haciendo clic sobre él aparece el eje seleccionado.



Elaboración de gráfica de barras simple. Paso 3

Segundo, hacemos clic derecho sobre el eje seleccionado para que aparezca un menú y allí seleccionamos la opción: Dar formato al eje.



Elaboración de gráficas simple. Paso 4.

En las opciones que aparecen podemos modificar el inicio del eje, el final del eje y el rango de separación entre los puntos del eje. Todas las opciones se pueden volver a generar y en último caso se elimina el gráfico y se vuelve a generar, así que vale la pena probar.

Excel también nos permite cambiar la apariencia de las barras, podemos modificar el fondo, los colores de las barras y el tamaño, todo para que personalicemos nuestro trabajo.

Hasta ahora, los ejemplos que vimos son barras simples, sin embargo también existen las barras compuestas que pueden ser apiladas o agrupadas y cada una cumple una función diferente. Los principios para su creación son los mismos.

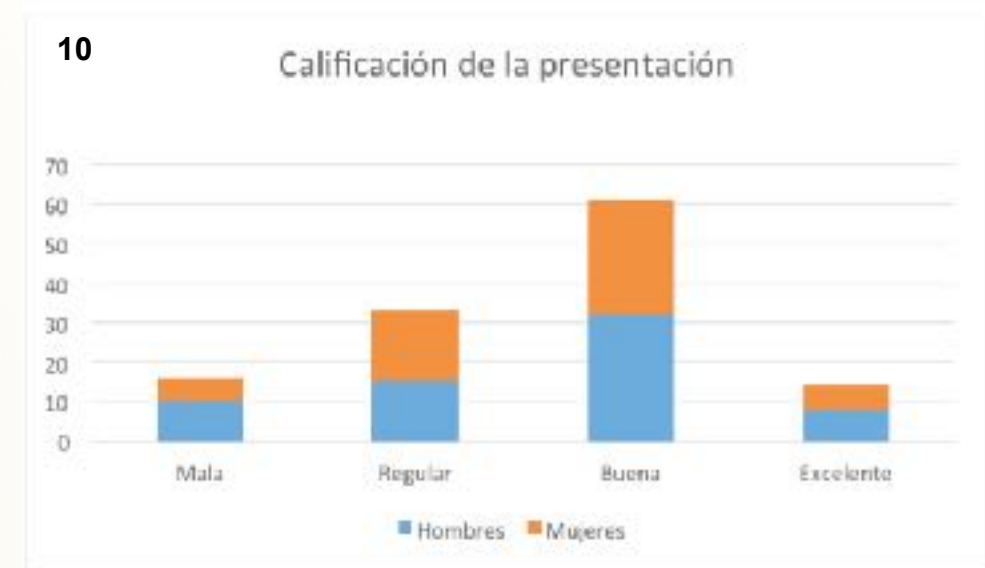
El criterio para usar una gráfica de barras compuesta es la necesidad de graficar una variable representada como matriz de doble ingreso. Supongamos por ejemplo la calificación que un grupo de personas le otorga a un presentador, pero los resultados los deseamos separados por hombres y mujeres.

9

Presentación	Hombres	Mujeres
Mala	10	6
Regular	15	18
Buena	32	29
Excelente	8	6

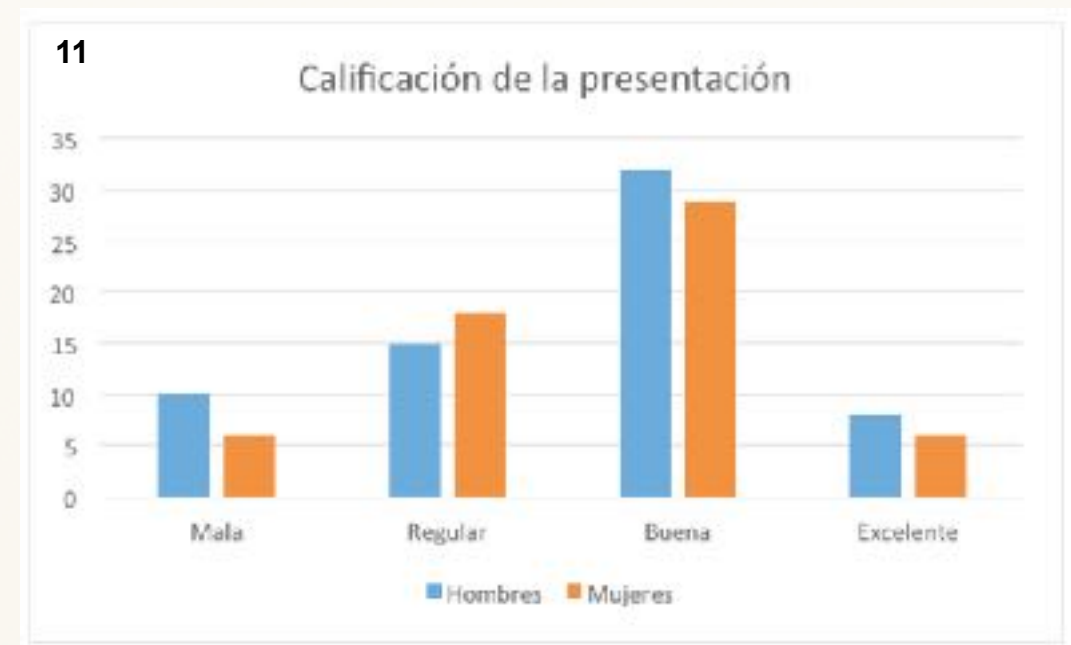
Datos para elaborar una gráfica de barras compuestas

En este caso, necesitaremos las barras compuestas, la altura de la barra indicará la frecuencia de la categoría pero estará dividida en dos grupos cuyas alturas indicarán las frecuencias propias.



Ejemplo de diagrama de barras compuestas compuesta.

Para el mismo ejemplo de doble entrada, se puede generar un gráfico de barras agrupadas, la selección de una u otra opción dependerá del estilo y la selección del autor.



Ejemplo de gráfico de barras agrupadas.

b. Diagramas de sectores (pie)

Es otra opción para representar principalmente datos cualitativos, su principal función es mostrar las divisiones de una población entera de acuerdo a la variable que nos interesa.



El diagrama de sectores consiste en una circunferencia que utiliza los ángulos para representar la proporción de la población en cada categoría o sector, dicho ángulo se obtiene multiplicando 360° , el total del pie, por la frecuencia absoluta de la categoría dividido N.

Es importante recalcar, el ángulo de cada sector del diagrama representa el tamaño proporcional de su frecuencia absoluta.

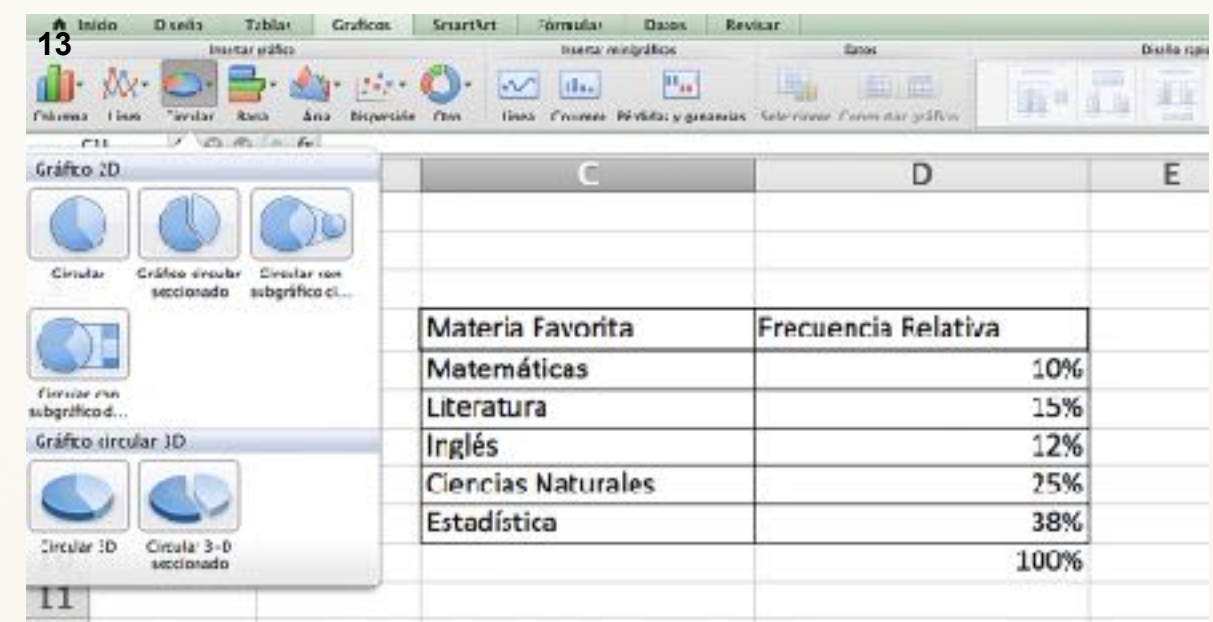
Su elaboración en Excel es bastante simple. En primer lugar seleccionaremos los datos que cumplen con las especificaciones mencionadas. Siempre verificamos que las proporciones sumen 100%.

12

	Materia Favorita	Frecuencia Relativa
5	Matemáticas	10%
6	Literatura	15%
7	Inglés	12%
8	Ciencias Naturales	25%
9	Estadística	38%
		100%

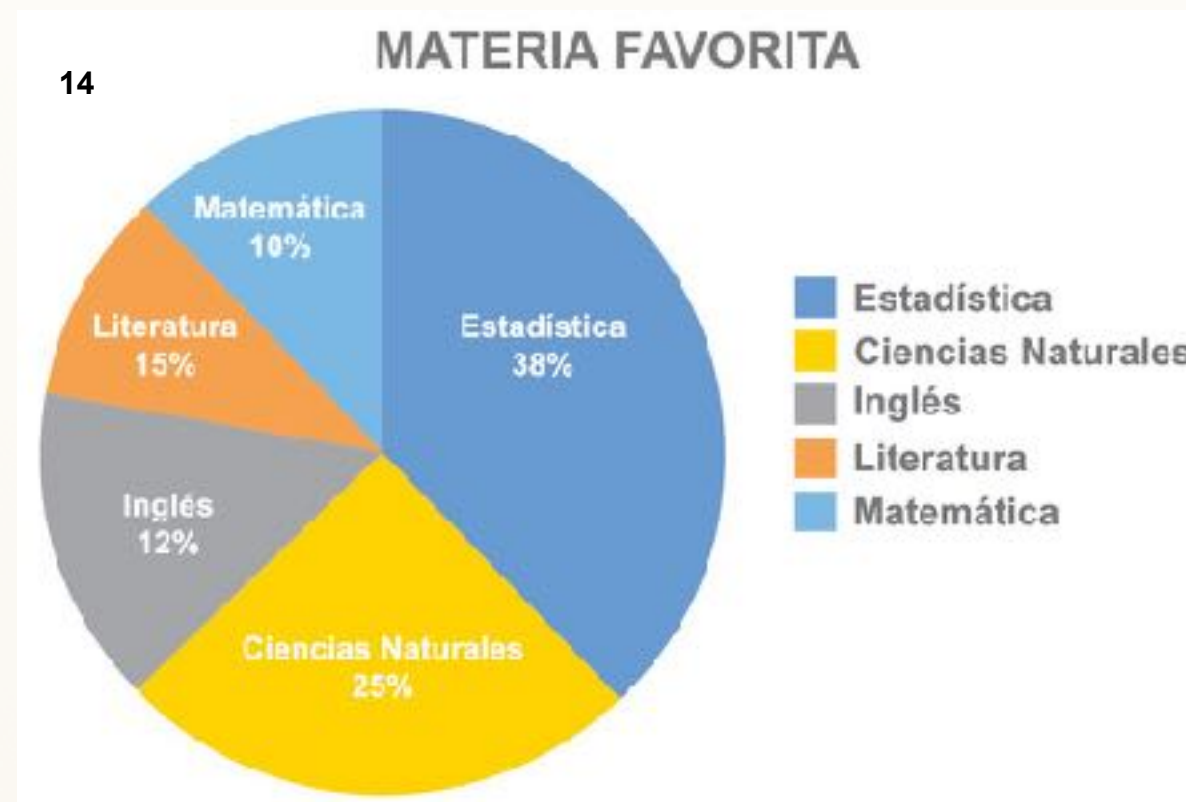
Elaboración de diagrama de sectores. Paso 1.

En el menú de insertar seleccionamos la opción de gráficos circulares (recordemos que dependiendo de la versión de Excel, el menú puede variar levemente).



Elaboración de gráficas simple. Paso 2.

El resultado final es la gráfica que necesitamos y podemos editarla de acuerdo a nuestras preferencias.



Ejemplo de diagrama de sectores



Comprobación del capítulo X: Representación de variables cualitativas

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo X

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XI

Representación de variables cuantitativas



Representación de variables cuantitativas

Para hablar de las variables cuantitativas dividiremos en dos tipos de variables, aquellas que están agrupadas y las que no lo están. Debemos recalcar nuevamente la importancia de saber con qué tipo de variables estamos trabajando en cada momento y qué tipo de variable deseamos representar, de forma que se pueda utilizar la guía de orientación para la selección de gráficas, presentada al inicio de la unidad.

1. Representación de variables cuantitativas no agrupadas

Para la representación de este tipo de variables nos enfocaremos en tres tipos, la primera ya es conocida porque se utiliza también para variables cualitativas ordinales... la gráfica de barras. También exploraremos la opción de las gráficas de escaleras.

Como vamos a ver la selección de una u otra opción depende del tipo de variable y de la idea que deseamos transmitir.

a. Barras: Las barras son el tipo de diagrama más común, lo vemos en reportes, en periódicos, en informes, en televisión y por lo tanto podemos deducir que si son tan utilizadas es por su versatilidad.

En el caso de las variables cuantitativas, son utilizadas para representar variables cuantitativas discretas (pueden tomar únicamente una cantidad limitada de valores). La mecánica para elaborarlas es la misma que se discutió en las variables cualitativas así que veremos un ejemplo ante una pregunta que genera datos cuantitativos discretos: ¿cuántos años de escolaridad tiene usted?



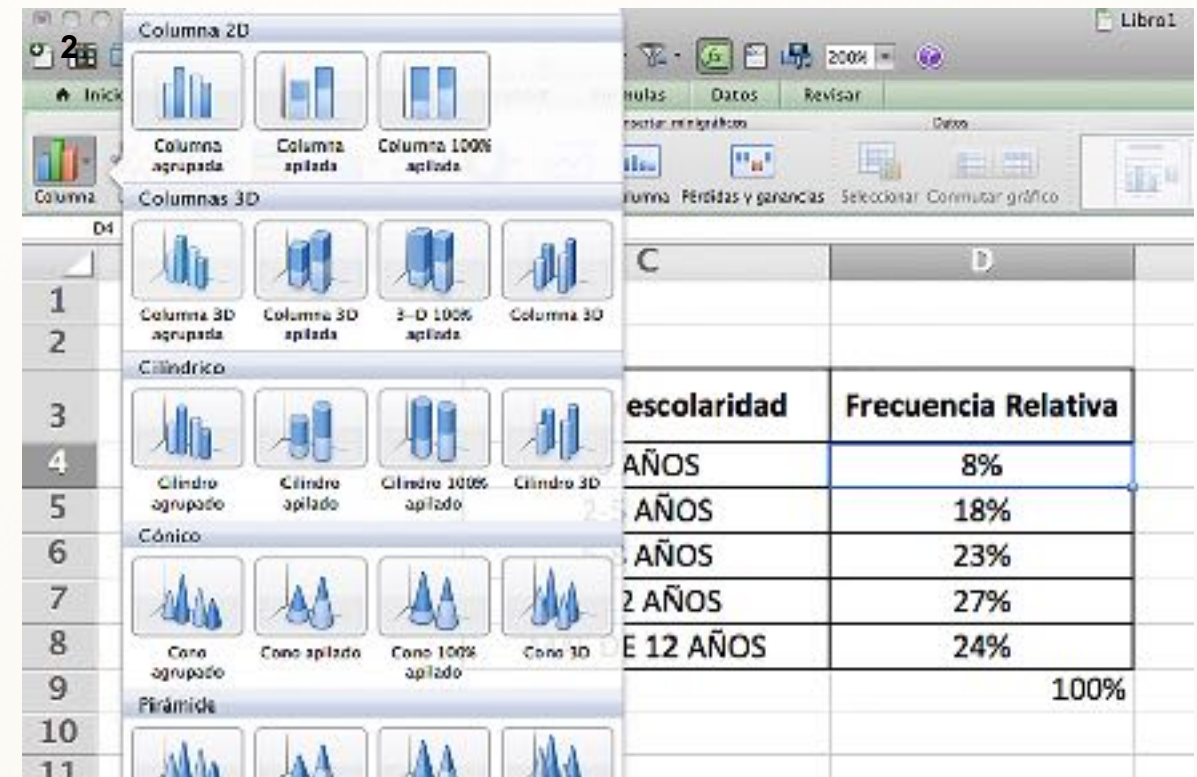
1

Años de escolaridad	Frecuencia Relativa
0 AÑOS	8%
3 AÑOS	18%
5 AÑOS	23%
8 AÑOS	27%
10 o Más	24%

Datos para elaborar una gráfica de barras de datos cuantitativos. Paso 1.

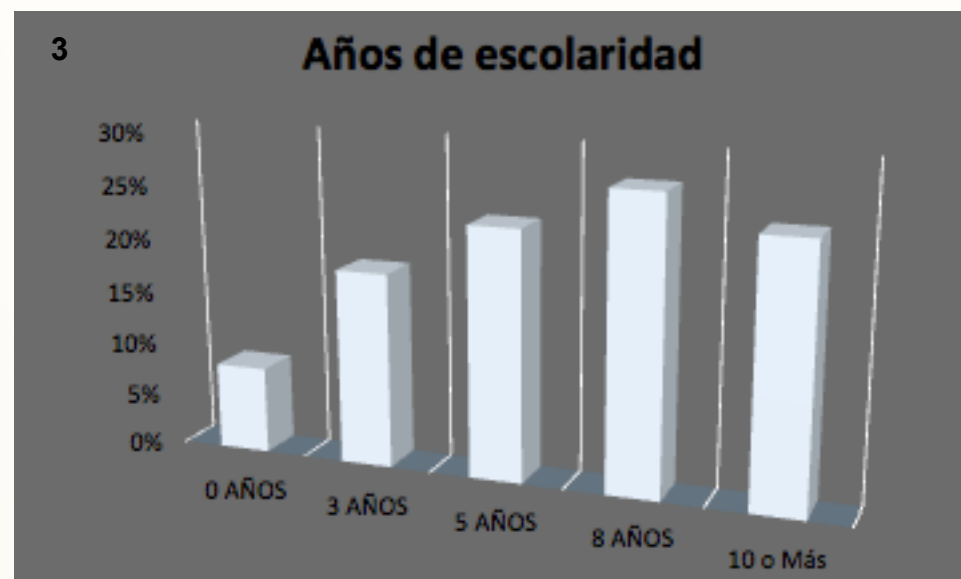
Recordemos que debemos asegurarnos que la respuesta sea 100% si estamos trabajando con datos relativos.

Como primer paso seleccionamos nuestros datos en Excel y buscamos la opción de “INSERTAR” allí buscamos la opción de barras (cada versión del programa muestra leves diferencias en el menú y la ubicación de los íconos).



Elaboración de gráfica de barras de datos cuantitativos. Paso 2.

El diagrama o gráfica final se podrá editar como hemos discutido anteriormente para modificar ejes, nombres, colores y apariencia de acuerdo a nuestras necesidades y buscando la forma más clara de mandar el mensaje con nuestros datos.



Ejemplo de barras para datos cuantitativos.

b. Escaleras: esta versión de las gráficas de barras son utilizadas para representar las frecuencias acumuladas de una distribución.



El principio de la representación gráfica es el de la barra de gráficas pero usando las frecuencias acumuladas y elaborando un paso adicional para representar adecuadamente los datos.

Partimos de la distribución de frecuencias.

4

Años de escolaridad	Frecuencia Relativa	Frecuencia acumulada
0 AÑOS	8%	8%
3 AÑOS	18%	26%
5 AÑOS	23%	49%
8 AÑOS	27%	76%
10 o Más	24%	100%

Datos para elaborar una gráfica de barras de escalera
Paso 1.

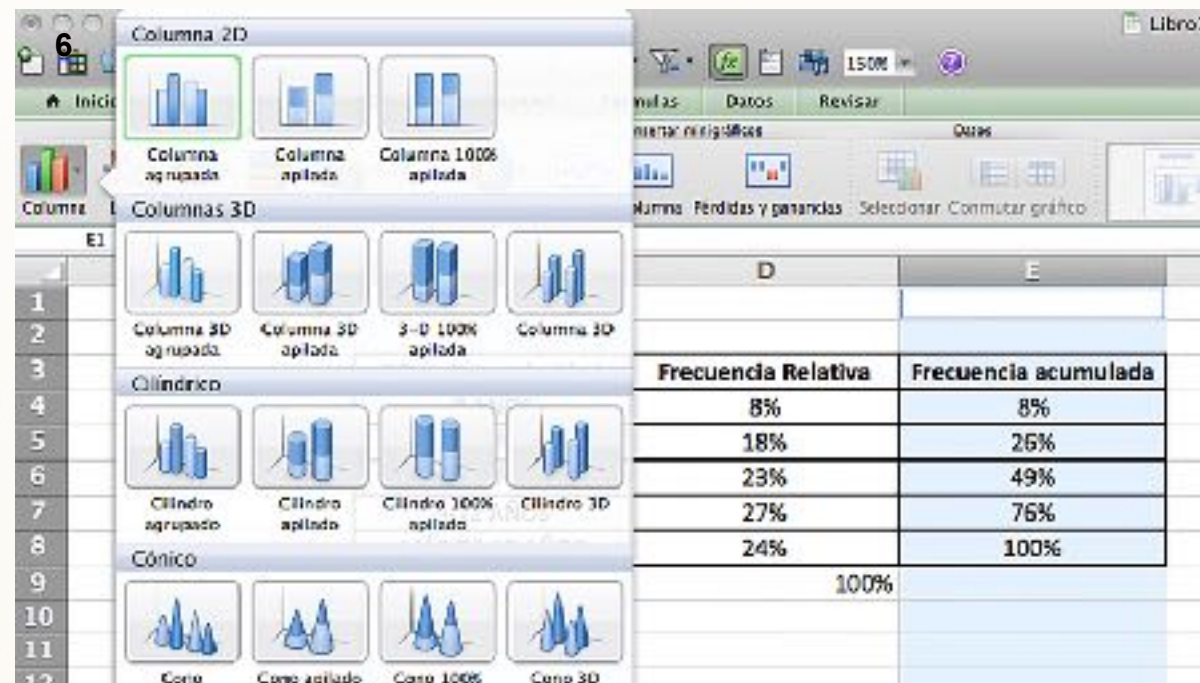
Seleccionaremos las columnas de años de escolaridad y frecuencia acumulada. Un consejo en Excel para hacerlo sin cambiar la tabla. Presionaremos la tecla CTRL y con el mouse seleccionamos la columna que deseamos y sin soltar CTRL seleccionamos la otra columna que deseamos. En Excel se ve así:

5

	A	B	C	D	E
1					
2					
3			Años de escolaridad	Frecuencia Relativa	Frecuencia acumulada
4			0 AÑOS	8%	8%
5			2-5 AÑOS	18%	26%
6			6-8 AÑOS	23%	49%
7			9-12 AÑOS	27%	76%
8			MÁS DE 12 AÑOS	24%	100%
9				100%	
10					

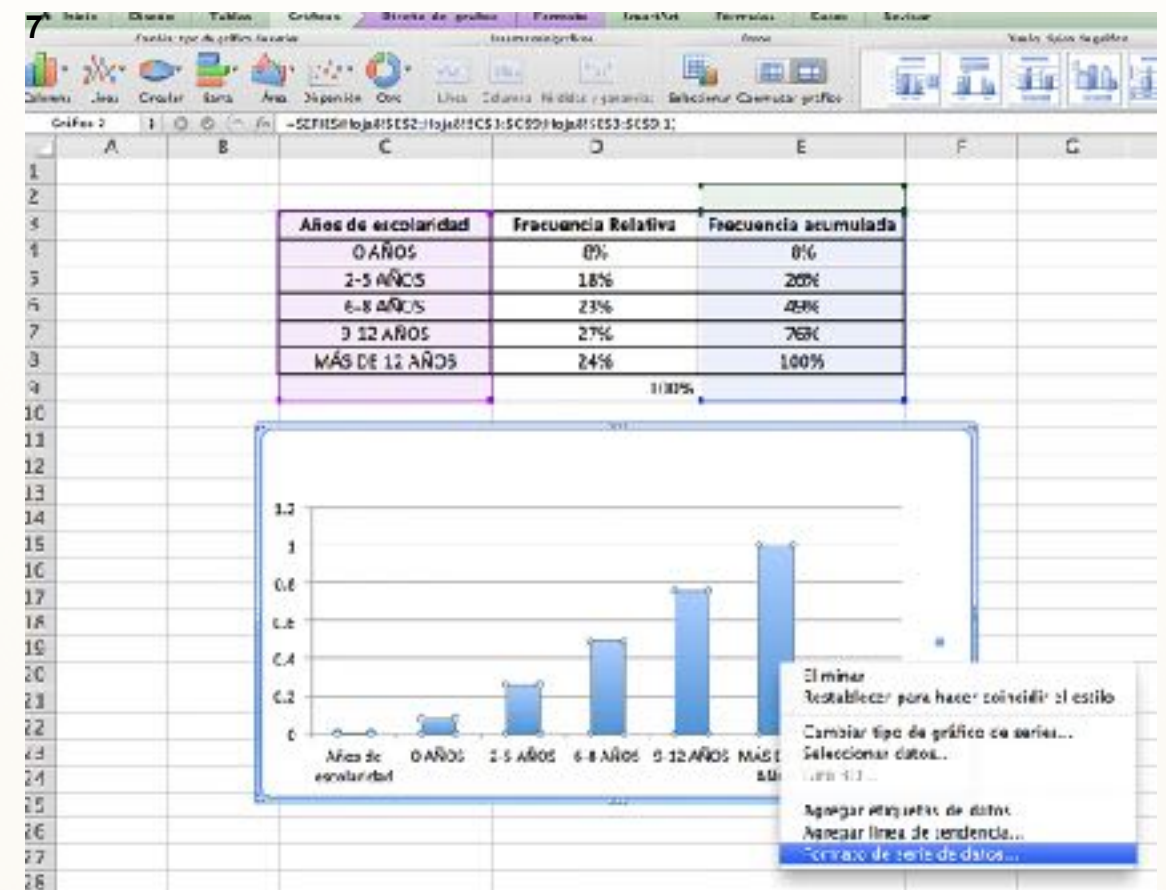
Elaboración de diagrama de escalera. Paso 2.

Seleccionaremos entonces INSERTAR y escogemos la opción de barras (tal y como hemos trabajado hasta ahora).



Elaboración de diagrama de escalera. Paso 3

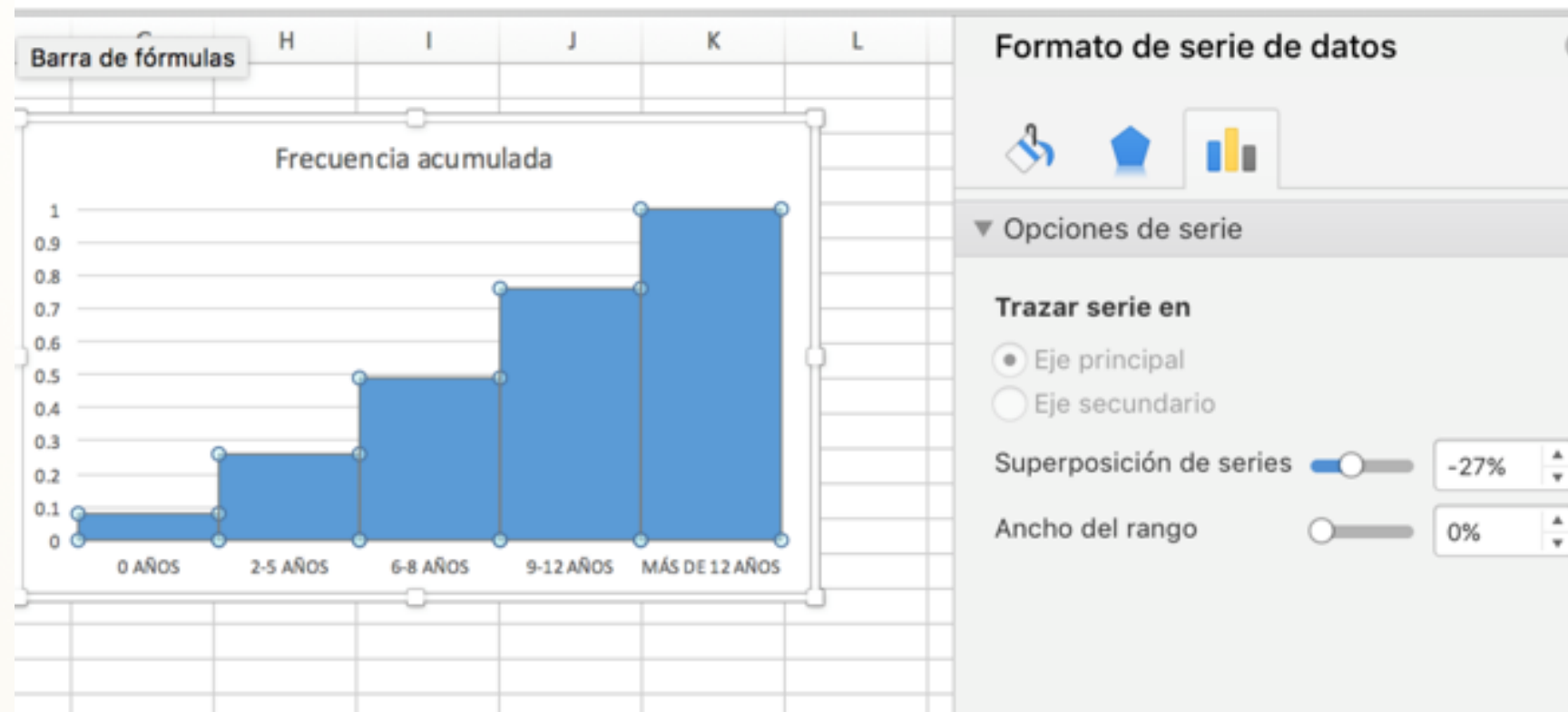
Vamos a seleccionar la gráfica generada (ahora es una gráfica de barras normal) haciendo clic normal con el mouse en alguna de las barras generadas y luego, cuando ya están seleccionadas haremos clic con el botón derecho para obtener un menú. Allí escogeremos la opción “Formato de serie de datos” tal y como podemos apreciar en la imagen.



Elaboración de diagrama de escalera. Paso 4.

Al hacer clic aparece la opción de modificaciones; únicamente debemos ubicar la opción de “ancho de rango” y llevarlo hasta el valor 0. (recordemos que la apariencia del menú y sus opciones varía de acuerdo a la versión de Excel).

8



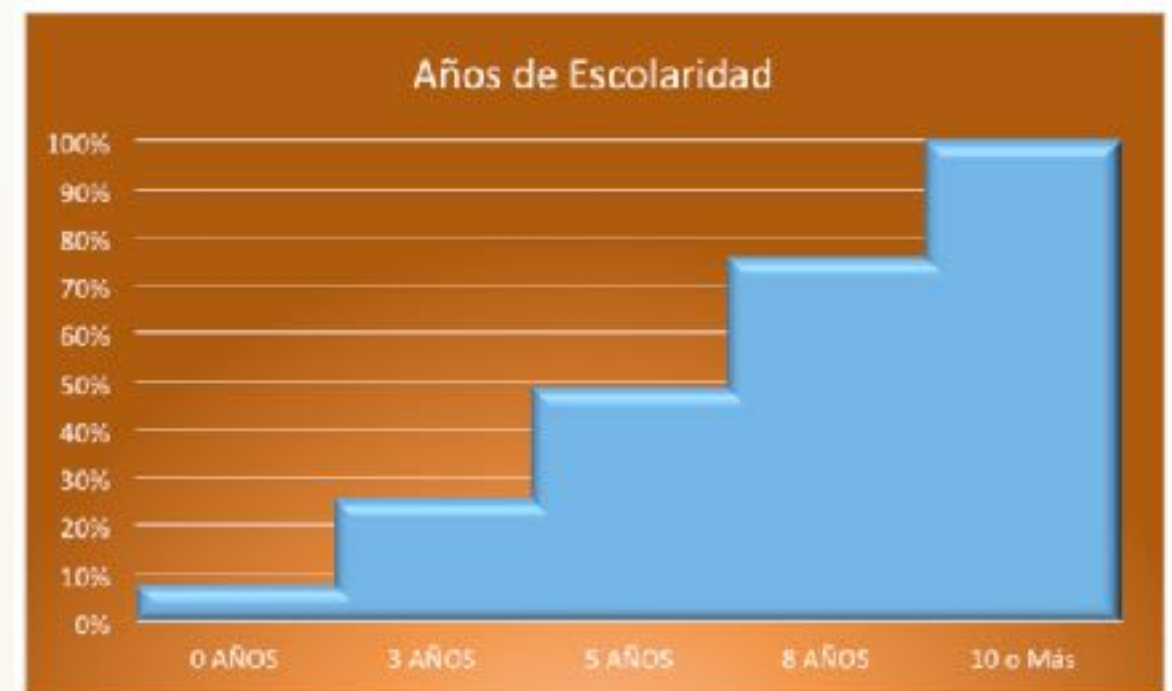
Elaboración de diagrama de escalera. Paso 5.

En este momento tenemos ya nuestro diagrama de escalera o diagrama escalonado.

Podemos modificar su apariencia para mejorar la representación de acuerdo a nuestros intereses.

Para este caso, se modificó el título, luego se seleccionó la imagen y modificó el color del fondo con un degradado, posteriormente al seleccionar las barras con el mouse se modificó el formato dando una apariencia de Bisel para lograr un efecto de escalera más marcado.

9



Ejemplo de diagrama de escalera.



Comprobación del capítulo XI: Representación de variables cuantitativas

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XI

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Unidad IV

Medidas



Esta unidad está dedicada al estudio de las medidas de tendencia central y de posición y variabilidad.

Hasta ahora hemos explorado los datos, sus características y la forma en que pueden representarse. Desarrollamos temas específicos para organizar los datos en tablas y poder hacer algunas interpretaciones. Luego discutimos sobre la forma de representar esos datos gráficamente de acuerdo a sus propiedades. En la tercera unidad profundizamos en la correcta interpretación de las tablas y las gráficas que se pueden producir con datos estadísticos.

En esta unidad hablaremos de medidas que nos permiten representar conjuntos de datos de acuerdo a su ubicación y su variabilidad lo cual nos permitirá tener más herramientas para representar los datos que recabemos en nuestros estudios.

Empezaremos contando una historia, la historia de Manolo. Empezaré por contarles que Manolo es un agrónomo retirado, actualmente vive con su

esposa en una pequeña granja afuera de la ciudad. Él disfruta del campo, de sus siembras y sus experimentos con semillas, pero algo que le apasiona es el cultivo del café. Manolo no se dedica a la producción para la venta, por el contrario produce para su familia y amigos, y se siente orgulloso de su producto de alta calidad. Él mismo prepara la tierra, selecciona la semilla para sembrar, cuida las plantas, cosecha la fruta del café, le quita la pulpa, seca las semillas al sol y por último lo tuesta. En la imagen siguiente podemos ver a Manolo disfrutando un momento de la producción de su café... secándolos al sol.

Ahora bien, quiero que veamos la imagen detalladamente y nos fijemos en los granos de café y vamos a pensar en ellos como datos!! En primer lugar debemos preguntarnos: Estos datos ¿son una muestra o son la población? En este caso, podemos ver que en el suelo hay muchos granos y los que él seleccionó para lanzar al aire son una parte. Así que pensaremos en el grupo que tira al aire como una muestra y a todos los granos que tenía allí como la población.



Historia de Manolo y el café, la dispersión

Al pensar así, diremos que si tomamos datos del grupo que él agarró (muestra), hablaremos de “estadísticos” pero si usamos esa muestra para hablar de datos de todos los granos disponibles hablaremos de “parámetros” (población) y en teoría, en algunos casos se podrían tener datos de TODOS los granos de café que existen, entonces hablaremos del “Universo” Entonces: estadísticos de una muestra y parámetros de una población.

Regresando a los granos de café en el aire (nuestra muestra), podemos observar una agrupación o una tendencia a la agrupación, en el centro se encuentra la mayoría y en los extremos aparecen granos de café, pero en menor cantidad.

En la práctica de la estadística, veremos este fenómeno siempre que tengamos datos, y entre más tenemos más se comportan de forma “normal”, así se le llama a una distribución que tiene agrupados a la mayoría en el centro y pocos en los extremos y una forma para expresar los datos con estos comportamientos son las “medidas de tendencia Central”.

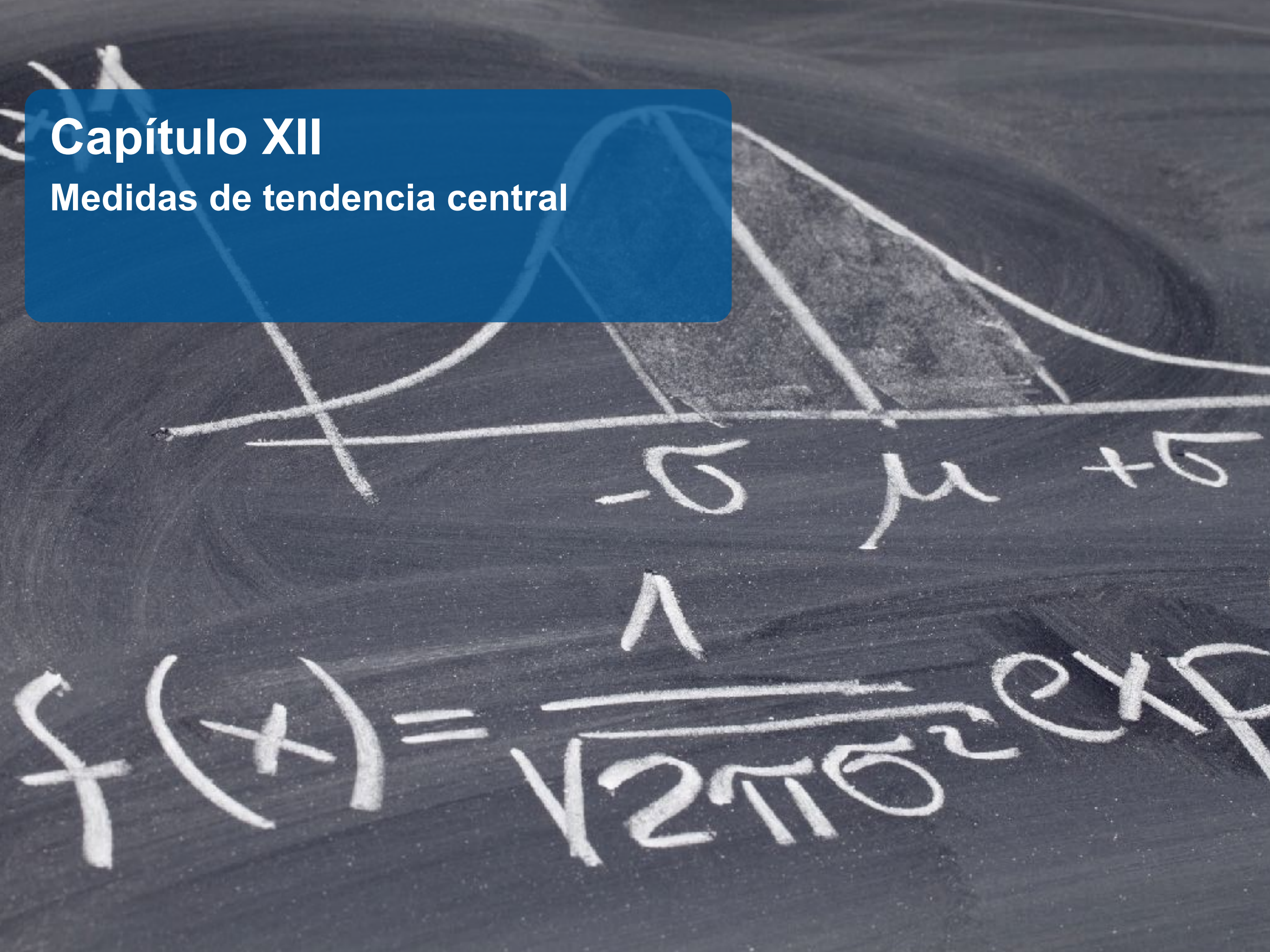
2



Historia de Manolo y el café, la dispersión

Capítulo XII

Medidas de tendencia central



A hand-drawn normal distribution curve is shown on a chalkboard. The curve is bell-shaped and centered at a point labeled μ . Two vertical lines are drawn on either side of the mean, labeled $-\sigma$ and $+\sigma$. The area under the curve between these two lines is shaded. Below the curve, the probability density function is written as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp$$

Medidas de tendencia central

Una revisión rápida en la web genera suficientes resultados para indicar qué se entiende por medidas de tendencia central, la siguiente página lo define así:

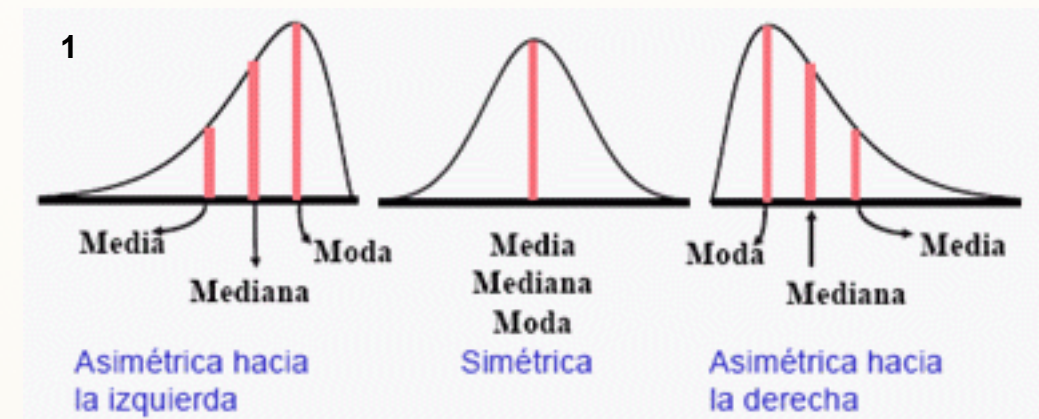
“Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda.”



Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable.

Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí.

De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y dispersión”.



Otra definición disponible es la de Wikipedia:

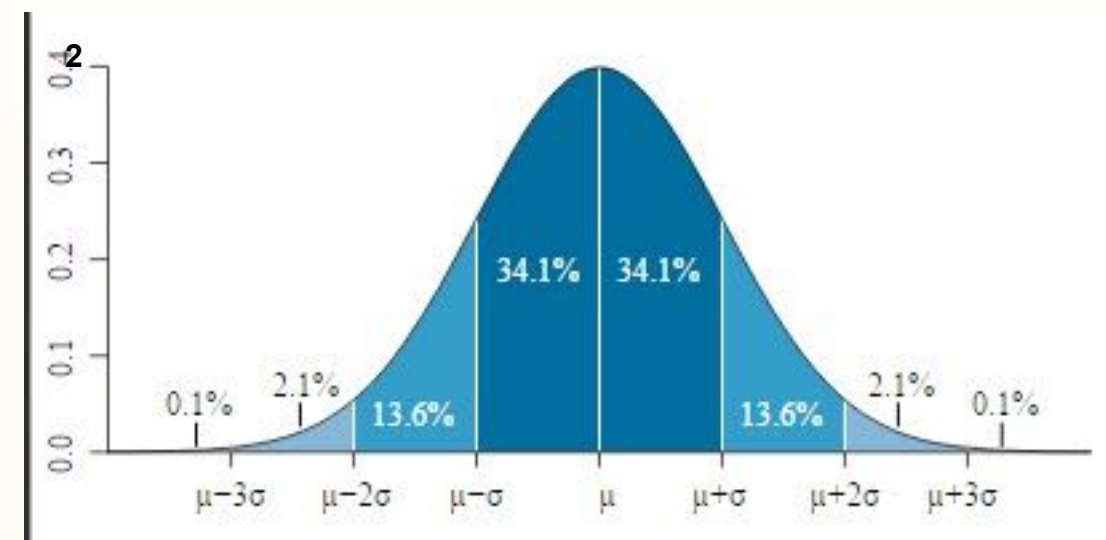
“Al describir grupos de diferentes observaciones, con frecuencia es conveniente resumir la información con un solo número. Este número que, para tal fin, suele situarse hacia el centro de la distribución de datos se denomina medida o parámetro de tendencia central o de centralización.”

Cuando se hace referencia únicamente a la posición de estos parámetros dentro de la distribución, independientemente de que ésta esté más o menos centrada, se habla de estas medidas como medidas de posición.¹ En este caso se incluyen también los cuartiles entre estas medidas.”

Un concepto importante que nos ayudará a entender mejor las medidas de tendencia central es la distribución normal. Así se le conoce a la probabilidad de ordenamiento de datos sociales, naturales y psicológicos ya que por lo regular al organizar los datos se observa este tipo de distribución. Tal y como vimos en el ejemplo de Manolo y como se puede determinar al realizar histogramas hay un grupo grande de datos en el centro y a la medida que nos alejamos del centro encontramos menos datos acumulados.



La distribución normal se representa con la famosa campana de Gauss.



Distribución normal

Al final del capítulo, volveremos a revisar esta imagen para poder entenderla con todas sus medidas, por el momento diremos que la mayoría de recolecciones de datos en ciencias sociales se comportan así y que cerca del promedio o media encontraremos al 68% de la población. En un sentido práctico y si alguien nos dice que evaluó a sus alumnos y el promedio es de 80 ya podemos tener una idea que la mayoría de personas tienen notas alrededor de 80. En cuanto nos adentremos en la unidad podremos ser más exactos para conocer a la población de acuerdo a estas medidas.



a. La Moda (Mo)

Definición y propiedades: La medida de la Moda, es el valor de la variable o dato que más veces se encuentra en nuestra distribución, es el dato cuya frecuencia es la más alta de nuestro conjunto de datos.



En otras palabras, la moda es el valor con mayor frecuencia absoluta.



Nota:

Es importante tomar en cuenta que pueden existir distribuciones de datos con más de una moda.

Cálculo: Cuando se trabajan datos no agrupados en intervalos, la forma de identificarla es revisando la tabla de distribución de frecuencias en la columna de frecuencias absolutas (f_i) e identificaremos la más alta. Esta será nuestra moda.



En distribuciones con intervalos, por ejemplo si tenemos una variable con rangos de edad (rango1= de 1 a 3; rango2 = de 4 a 6; rango3 = de 7 a 9; rango4 = de 10 a 12 y rango5 de 13 a 15) tendremos dos opciones para trabajarlo.

La primera opción es decidir que usaremos un “rango modal” en lugar de una “moda” y en este caso buscamos en la distribución de frecuencias el rango con la frecuencia absoluta (f_i) más alta.

La segunda opción es calcular la moda utilizando los límites de los rangos; el cálculo se complica un poco pero genera datos más confiables, utilizaremos la siguiente fórmula.

3

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

*Moda***Donde:** L_i = Es el límite inferior del intervalo modal f_i = Frecuencia absoluta del intervalo modal f_{i-1} = Frecuencia absoluta inmediatamente inferior al intervalo modal f_{i+1} = Frecuencia absoluta inmediatamente posterior al intervalo modal a_i = Amplitud de la clase o intervalo

Veamos un ejemplo entonces, aquí hay parte de una distribución de frecuencias.

Con la opción 1 que se mencionaba, diremos que el intervalo modal es el de 7 a 9 ya que su f_i es 66, la más alta.

4

Intervalos	f_i
de 1 a 3	10
de 4 a 6	37
de 7 a 9	66
de 10 a 12	42
de 13 a 15	14

Distribución de frecuencias

Con la opción 2 debemos hacer los cálculos según la fórmula antes mencionada.

Paso 1: Identificamos L_i = Es el límite inferior del intervalo modal en otras palabras el límite de abajo del rango que identificamos en la opción anterior. En nuestro caso es 7.

Paso 2: identificamos f_i y f_{i-1} para realizar la resta. En esta caso la frecuencia absoluta (f_i) de nuestro intervalo modal es 66 y la inmediatamente inferior es 37, así que $66 - 37 = 29$ y ya tenemos el numerador de nuestra operación.

Paso 3: identificamos f_{i+1} para operar el denominador y en este caso es 42 (la frecuencia absoluta superior al de nuestro intervalo modal).

Nos ubicamos en el primer paréntesis y tenemos que restar lo mismo que en el paso 2, así que sabemos que el resultado es 29. Ahora debemos tomar f_i y restarle f_i+1 , En este caso será $66-42 = 24$.

Y ahora operamos los resultados del denominador $29 + 24 = 53$.

Paso 4: identificamos a_i , para ello veremos cuántos datos están representados en cada intervalo y vemos que de 1 a 3 hay 3 datos el 1, el 2 y el 3. Lo mismo pasa con los demás intervalos, así que la amplitud del intervalo (a_i) es 3.

Ahora tenemos todos los datos para operar respetando la jerarquía de las operaciones

Primero el numerador que ya sabemos es 29, debemos dividirlo entre el denominador 53

El total entonces de la fracción es $= 0.55$

Ahora debemos multiplicar este resultado por la amplitud, así que $0.55 * 3 = 1.65$

Por último le sumamos a Li . Así que $7+1.65 = 8.65$
 $Mo = 8.65$

¡Esta es nuestra moda!

Uso de Excel: como ya lo hemos dicho, las herramientas se han creado para ayudarnos a facilitar nuestro trabajo. Aunque no deja de ser importante conocer la lógica y la operación de los valores que vamos a usar, también es importante que podamos apoyarnos en los programas que nos ayudan a hacer la vida más fácil.

Tomemos en cuenta también, que cuando tenemos un grupo de 20 datos, 50 datos incluso puede que sea fácil recurrir a una inspección visual para determinar la moda, sin embargo cuando hablamos de 500 datos, 1000 datos o cientos de miles de datos, es muy complicado hacerlo de esta forma.

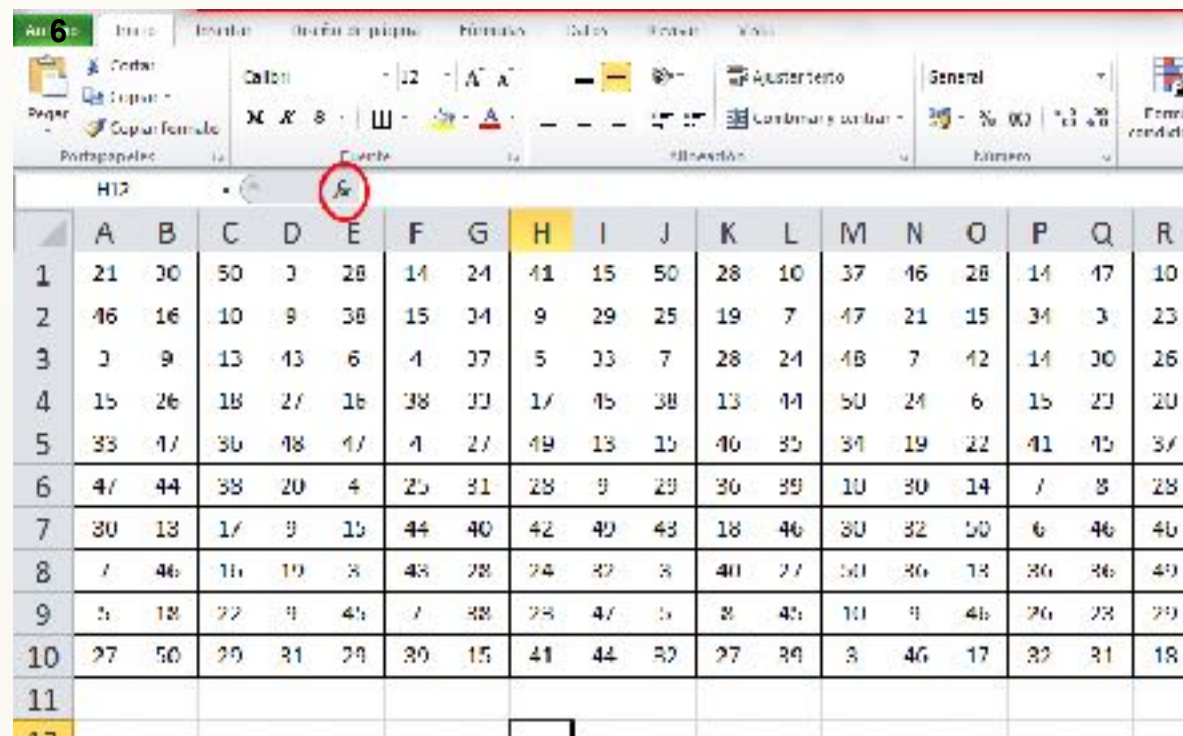
5

21	30	50	3	28	14	24	41	15	50	28	10	37	46	28	14	47	10	7	46	47	12	33	33	31	46	33	46	29	25
46	16	10	9	38	15	34	9	29	25	19	7	47	21	15	34	3	23	34	41	29	22	42	43	47	50	49	28	34	13
3	9	13	43	6	4	37	5	33	7	28	24	48	7	42	14	30	26	20	26	34	32	41	17	4	13	48	48	10	28
15	26	18	27	16	38	33	17	45	38	13	44	50	24	6	15	23	20	30	5	50	21	4	21	47	16	41	14	45	24
33	47	36	48	47	4	27	49	13	15	46	35	34	19	22	41	45	37	48	19	29	34	30	32	38	12	27	36	5	28
47	44	38	20	4	25	31	28	9	29	36	39	10	30	14	7	8	28	5	17	41	38	38	13	32	14	38	27	42	33
30	13	17	9	15	44	40	42	49	43	18	46	30	32	50	6	46	46	37	7	17	21	49	29	8	23	43	9	34	19
7	46	16	19	3	43	28	24	32	3	40	27	50	36	13	36	36	49	30	48	34	25	12	50	39	14	8	18	40	25
5	18	22	9	45	7	38	23	47	5	8	45	10	9	46	25	23	29	23	17	19	33	48	37	27	29	27	50	45	36
27	50	29	31	29	39	15	41	44	32	27	39	3	46	17	32	31	16	35	38	32	18	8	36	26	12	13	50	34	11

Ejemplo de datos, calcular moda

En este ejemplo tenemos 300 datos, vamos a suponer que son las edades de todas las personas que han visitado nuestra institución durante el año y ahora queremos saber cuál es la edad más frecuente que nos visita para adaptar el área de recepción acorde a dicha edad.

Veamos cómo se ven estos datos en Excel.



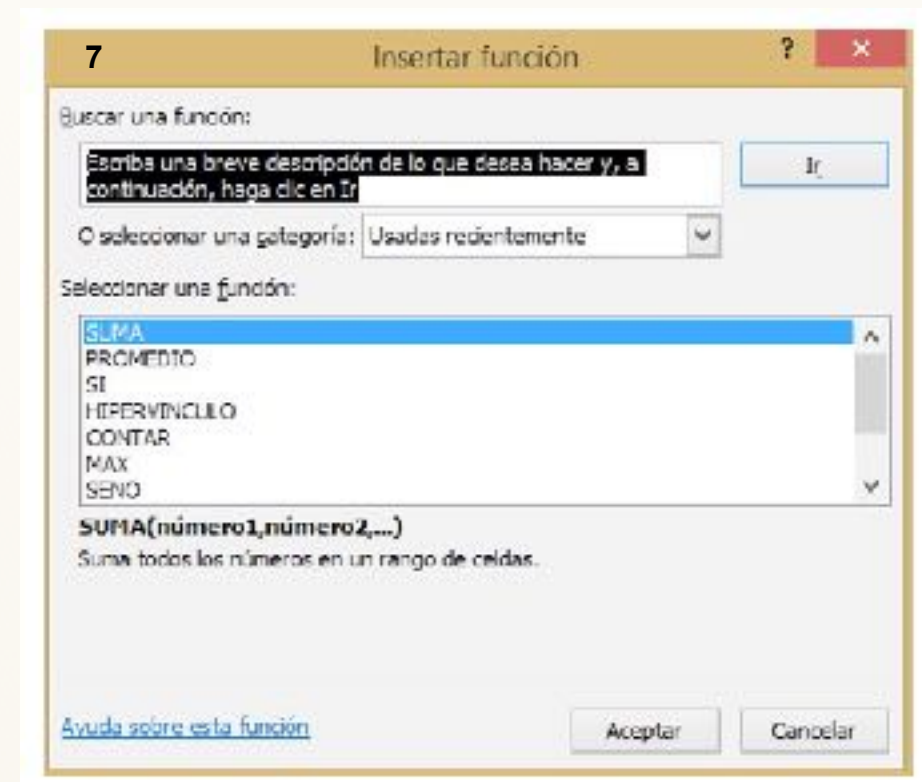
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	21	30	50	3	28	14	24	41	15	50	28	10	37	46	28	14	47	10
2	46	16	10	9	38	15	34	9	29	25	19	7	47	21	15	34	3	23
3	3	9	13	43	6	4	37	5	33	7	28	24	48	7	42	14	30	26
4	15	26	18	27	16	38	33	17	45	38	13	44	50	24	6	15	23	20
5	33	47	30	48	47	4	27	49	13	15	40	35	34	19	22	41	45	37
6	47	44	38	20	4	25	31	28	9	29	30	39	10	30	14	7	8	28
7	30	13	17	9	15	44	40	42	40	43	18	46	30	32	50	6	46	40
8	7	46	11	19	3	43	28	24	32	3	40	27	10	36	13	36	46	49
9	5	18	22	9	45	7	38	23	47	5	3	45	10	9	46	26	23	29
10	27	50	25	31	29	30	15	41	44	32	27	34	3	46	17	32	31	18
11																		

Ejemplo de datos, calcular moda en Excel

Tenemos en la imagen 10 filas (de 1 a 10) y 30 columnas (de A hasta AD).

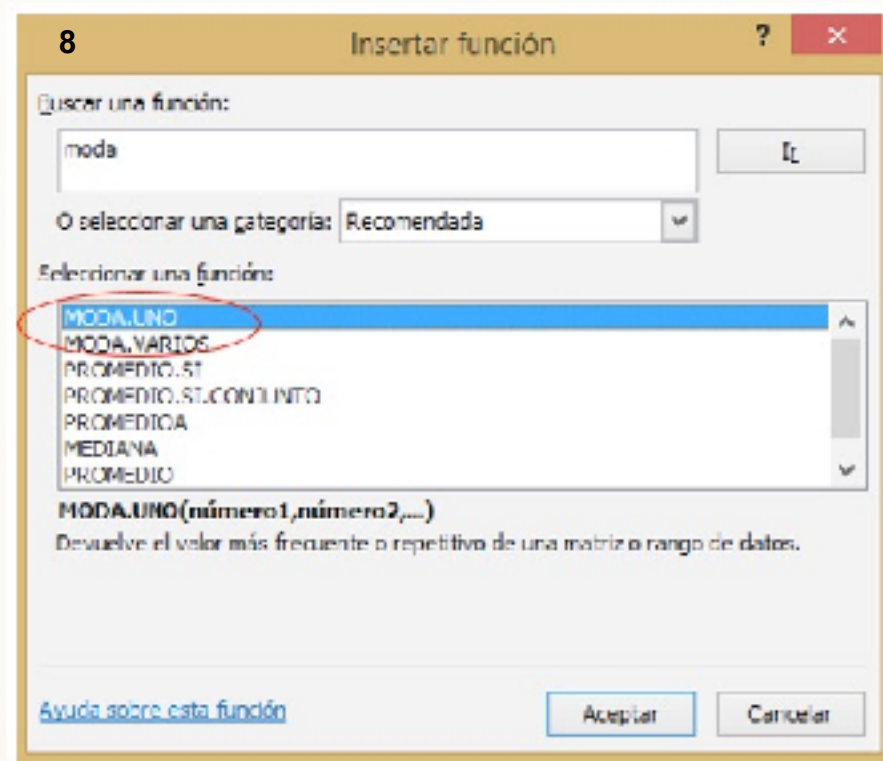
Si observamos arriba de los datos hay un área que se identifica con fx que está señalada en la imagen con el círculo rojo. Esta área es el menú de funciones de Excel que organiza operaciones que los usuarios pueden necesitar.

Así pues, si hacemos click en el área fx obtenemos un nuevo menú que nos permite buscar y seleccionar la función que deseamos, en el cuadro de “Buscar una función” escribiremos la palabra “moda”.



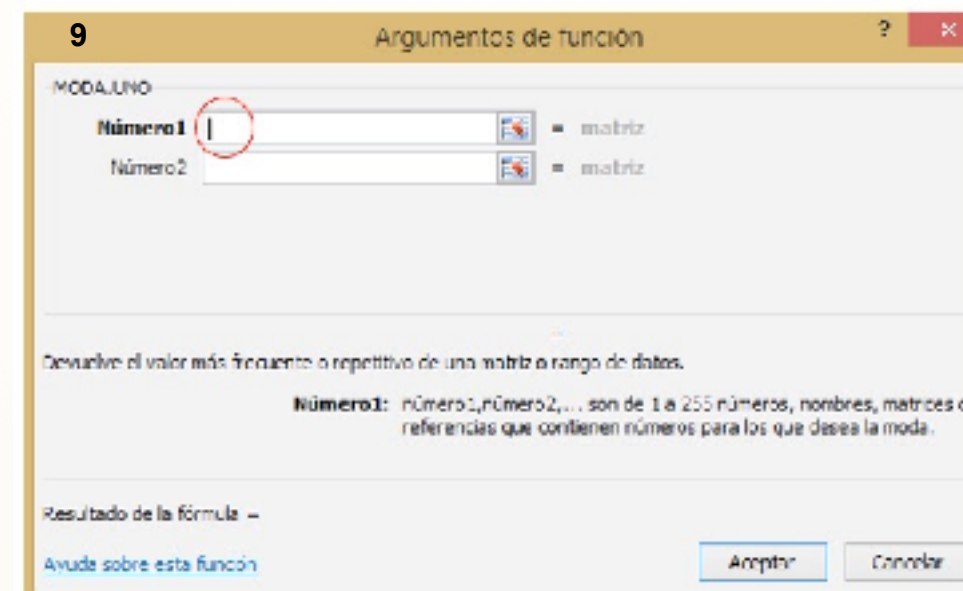
Ejemplo de datos, calcular moda Paso 2

Al presionar “*enter*” Excel nos mostrará las opciones que tiene para el término que usamos; en este caso aparecen 2 opciones “*moda.uno*” y “*moda.varios*”. En principio ambos nos funcionan pero para lo que estamos trabajando la opción *Moda.uno* es la opción que necesitamos.



Ejemplo de datos, calcular moda Paso 3

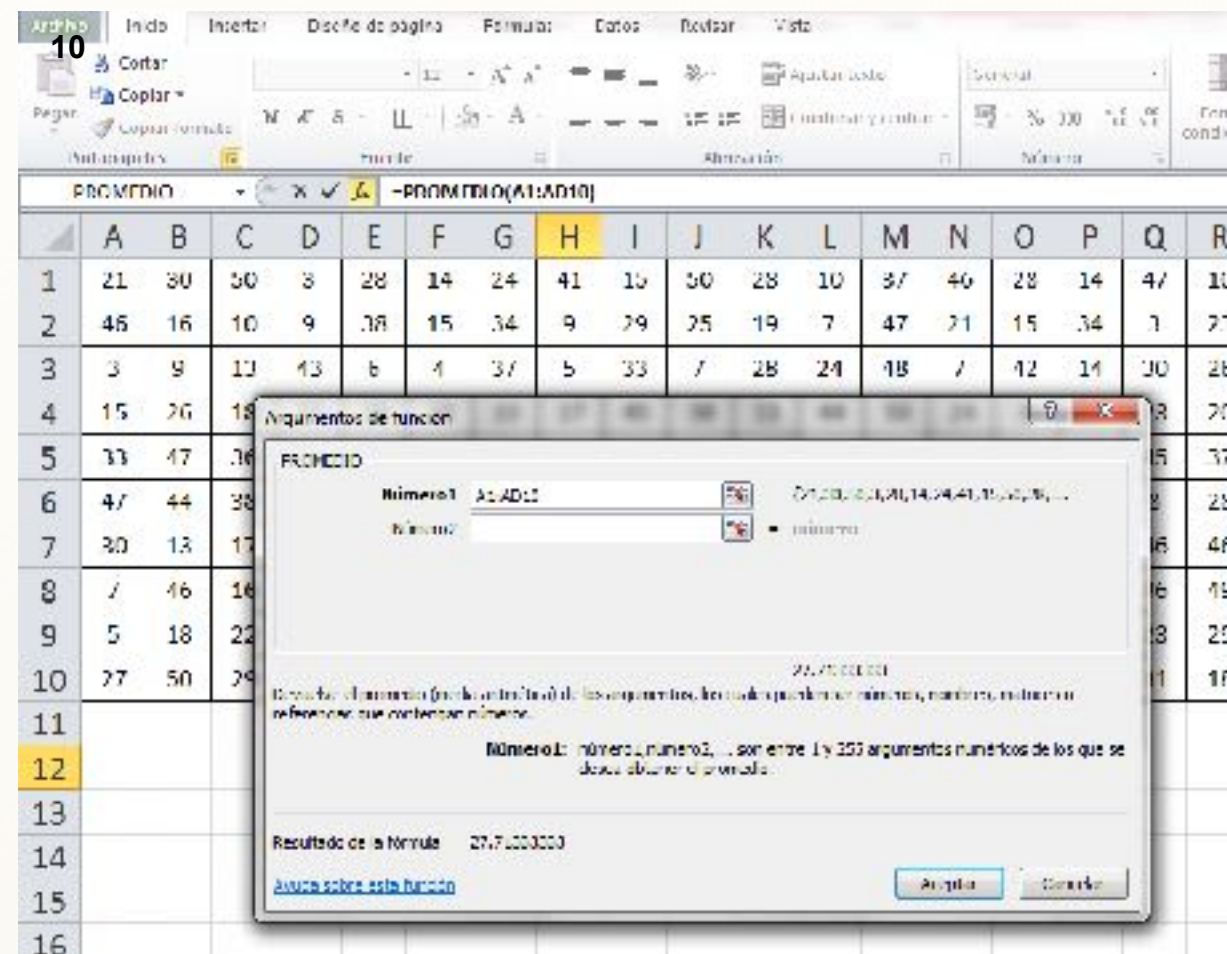
Después de seleccionado y al hacer click en aceptar, obtendremos una pantalla nueva que nos pedirá los argumentos de la función, es decir, nos va a pedir que seleccionemos los datos de los cuales necesitamos obtener la moda.



Ejemplo de datos, calcular moda. Paso 4

Es importante que notemos que nos da la opción de ingresar 2 matrices de datos, esto es muy importante para cuando tenemos varias tablas de datos separadas y deseamos la moda de todas ellas. En nuestro ejemplo solo trabajaremos con una tabla de datos, por lo que solamente llenaremos el área en negrillas “*número 1*”.

El primer paso es colocar el cursor en la casilla identificada por “*número 1*”, tal y como se muestra en la imagen anterior. Luego seleccionaremos con el cursor toda el área de datos, esto hará que Excel reconozca que debe realizar la función para todos los datos que están en el área que seleccionamos.



Ejemplo de datos, calcular moda. Paso 5.

Al hacer click en aceptar, obtendremos la moda de nuestros datos en la celda que teníamos seleccionada antes de abrir el menú de función.

11

=MCDA.UNO(A1:AD10)																														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	21	30	50	3	23	14	24	41	15	50	23	10	37	46	26	14	47	10	7	46	47	12	30	30	31	46	33	46	29	25
2	45	16	10	0	33	15	34	0	20	25	10	7	47	21	15	34	3	20	34	41	20	22	42	40	47	50	40	26	34	10
3	3	9	13	43	0	4	37	0	33	7	23	24	43	7	42	14	30	20	20	20	34	32	41	17	4	13	46	46	10	23
4	15	20	18	27	10	33	33	17	40	38	13	41	50	24	0	15	23	20	30	0	50	21	4	21	47	10	41	14	40	24
5	34	47	40	48	47	4	27	49	13	15	40	30	34	19	20	41	40	37	43	19	24	44	30	37	34	12	27	40	5	24
6	47	44	48	20	4	20	31	28	4	24	30	34	10	30	14	7	8	28	0	17	41	38	34	14	37	14	48	27	42	34
7	30	13	17	9	15	44	40	42	49	43	18	45	30	32	50	6	45	45	37	7	17	21	49	24	8	23	43	9	34	19
8	7	46	16	19	3	43	28	24	32	3	40	27	50	36	13	36	35	49	30	48	34	25	12	50	29	14	8	18	40	25
9	5	18	22	0	45	7	18	23	47	5	8	45	10	0	46	26	20	20	23	17	10	30	43	17	27	20	27	50	45	15
10	27	50	20	31	20	30	15	41	44	32	27	30	3	46	17	32	31	13	35	38	32	16	6	35	25	12	13	50	34	11
11																														
12								46																						
13																														
14																														
15																														

*Ejemplo de datos, calcular moda. Paso 6.
En este caso, la moda es 46.*

b. Media Aritmética (promedio)

Definición y propiedades: Probablemente es la medida de tendencia central que todos los lectores ya conocen, se trata del famoso “promedio” o la “media”. La forma de representar a la media cuando se trata de una muestra (los granos de café que tomó Manolo en nuestro ejemplo anterior) es una equis con una barra encima ($\bar{X}$) y cuando estamos hablando de la población se expresa con la letra griega mú (μ). En nuestro caso, por lo general estaremos hablando de muestras.

12

$$\bar{X} = \frac{\sum x}{n}$$

Media Aritmética

La forma más simple para definir la media es decir que, es el cálculo de sumar todos los datos de nuestra variable y dividirlos entre el total de datos que tenemos disponibles. Por ejemplo si pedimos a 5 personas que nos digan cuánto dinero tienen el bolsillo en este momento y nos dicen: Q12; Q20; Q15; Q11 y Q17 y ahora debemos reportarlo a un grupo de personas, para simplificar el conjunto de datos podemos calcular la media sumando los datos $12+20+15+11+17 = 75$ y lo dividimos entre $5 = 15$. Así que podemos RESUMIR el grupo de datos con la media y decir que el grupo de entrevistados en promedio tiene Q15.

La media tiene algunas propiedades interesantes y con aplicaciones prácticas.

- Si a todos los valores se les suma una misma cantidad, la media resultante quedará aumentada en la misma cantidad.
- Si todos los valores son multiplicados por la misma cantidad, la media resultante quedará multiplicada por la misma cantidad.



- La suma de los valores de las desviaciones respecto a la media deben sumar 0. En nuestro ejemplo al restarle cada valor a la media (15 menos 12, 15 menos 20...) los valores serían 3,-5,0,4-2, y la suma de ellos es 0.
- Es sensible a datos extremos en ambos límites.

Cálculo: El cálculo de la media es simple, ya lo empezamos a describir desde su definición y podemos decir que la media se calcula a través de la adición de todos los valores recabados y dividirlo entre el total de sumandos.

La expresión matemática de la media :

13

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + \dots + a_n}{n}$$

Expresión matemática de la media.

Donde a es cada valor o registro y n es el total de registros en nuestra serie de datos.

Un dato importante que debemos considerar, es que la “tendencia central” de la media está afectada por valores extremos. Lo podemos deducir al analizar la forma en la cual se calcula y es muy importante entender esta situación. Veamos un ejemplo parecido al que hicimos antes. Le preguntamos a 5 personas que nos indiquen cuánto dinero llevan en los bolsillos y responden Q.12, Q.20, Q.15, Q.11, Q.200 (solamente cambiamos el último dato de 17 a 200).

Calculemos ahora: la suma es 258 y ahora la media es 51.6 así que redondeamos a 52. Nuestra media es de Q.52.00 sin embargo vale la pena preguntarse si realmente Q.52.00 representa bien al grupo. Para ello complementaremos más adelante con las medidas de variabilidad.

Uso de Excel: Veamos entonces cómo podemos calcular la media, haciendo uso de la herramienta de Microsoft Excel, esto nos será particularmente útil cuando tengamos listados de números muy grandes. Vamos a usar la misma tabla de datos que usamos para calcular la moda.

14

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	21	30	50	3	23	14	24	41	15	50	23	10	37	46	26	14	47	10	7	46	47	12	33	33	31	46	33	46	29	25
2	45	16	10	9	33	15	34	9	20	25	10	7	47	21	15	34	3	23	34	41	20	22	42	43	47	50	40	26	34	10
3	3	9	13	43	6	4	37	5	33	7	23	21	43	7	42	14	30	20	20	20	34	32	41	17	4	13	48	48	10	23
4	15	20	18	27	16	33	33	17	40	38	13	44	50	24	6	10	23	20	30	5	50	21	4	21	47	16	41	14	40	24
5	34	47	40	48	47	4	27	49	13	15	40	35	34	19	72	43	40	37	48	19	79	34	30	37	34	12	77	40	5	34
6	47	44	48	20	4	25	31	28	4	79	30	34	10	30	14	7	8	28	5	17	43	38	34	14	37	14	48	77	47	34
7	30	13	17	9	15	44	40	42	49	43	18	45	30	32	50	6	45	45	37	7	17	71	49	29	8	23	43	9	34	19
8	7	46	16	19	3	43	28	24	32	3	40	27	50	36	13	36	35	49	30	48	34	25	12	50	39	14	8	18	40	25
9	5	18	22	9	45	7	13	23	47	5	8	45	10	9	46	26	23	23	23	17	10	33	43	17	27	20	27	50	45	15
10	27	50	20	31	20	30	15	41	44	32	27	30	3	46	17	32	31	13	35	38	32	16	8	35	25	12	13	50	34	11
11																														
12								46																						
13																														
14																														
15																														

Calcular media.

Los datos se verán en nuestra hoja de trabajo de Excel tal y cómo se aprecian en la siguiente imagen y al igual que hicimos con la moda, utilizaremos la función fx para entrar a la pantalla de fórmulas en Excel. En la imagen se muestra en un círculo rojo.

15

Archivo Inicio Insertar Diseño de página Fórmulas Datos Referencias Vista

Cortar Copiar Copiar formato Portapapeles

Calibri 11 A⁺ A⁻

B *I* U

--	--

Ajustar texto Combinar y centrar

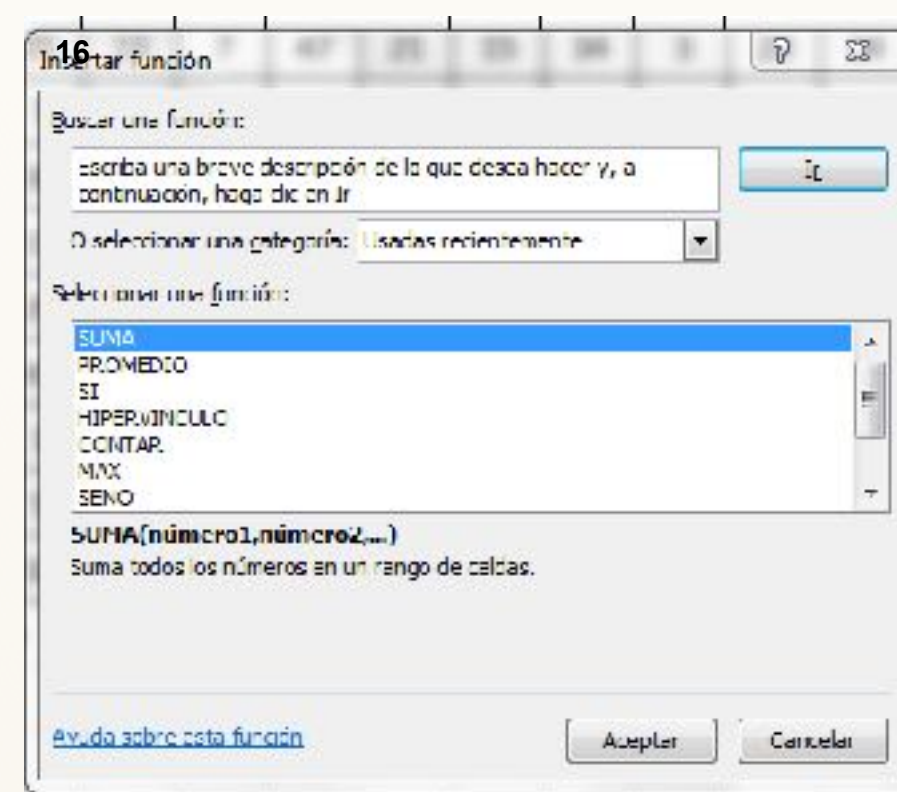
Formato Alineación

	W17																														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	21	50	50	3	23	14	24	41	15	50	28	10	37	46	28	14	47	10	7	46	47	12	33	33	31	46	33	46	29	25	
2	46	16	10	9	32	15	34	9	29	25	19	7	47	21	15	34	3	23	34	41	29	22	47	43	47	50	49	28	34	13	
3	3	9	13	43	6	4	37	5	33	7	28	24	48	7	42	14	30	26	20	25	34	32	41	17	4	13	48	48	10	28	
4	15	26	18	27	15	38	35	17	45	38	15	44	50	14	6	15	25	20	30	5	50	21	4	21	47	16	41	14	45	24	
5	33	47	36	48	47	4	27	49	13	15	46	35	34	19	22	41	45	37	48	19	29	34	30	32	38	12	27	36	5	28	
6	47	44	38	20	4	25	31	28	9	29	36	39	10	30	14	7	8	28	5	17	41	38	38	13	32	14	38	27	42	33	
7	30	13	17	9	15	44	40	42	49	43	18	46	30	32	50	6	46	46	37	7	17	21	49	29	8	23	43	9	34	19	
8	7	46	16	19	3	43	28	24	32	3	40	27	50	36	13	36	36	46	30	48	34	25	12	50	39	14	8	18	40	25	
9	5	18	22	9	45	7	38	23	47	5	8	45	10	9	46	26	23	25	23	17	19	33	48	37	27	29	27	50	45	36	
10	27	50	29	51	29	39	15	41	44	52	27	39	5	46	17	52	51	18	35	38	52	18	8	55	26	12	15	50	54	11	
11																															
12																															
13																															
14																															

Ejemplo de datos, calcular media Paso 1.

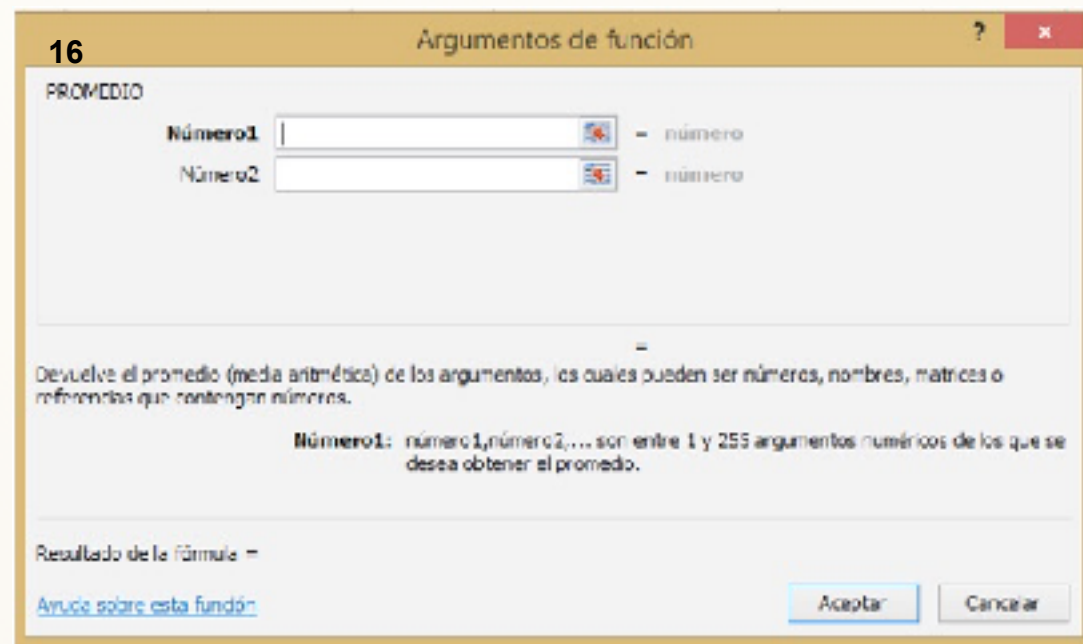
Antes de hacer clic en el área de funciones, posicionaremos el cursor en la casilla en la cual deseamos el resultado, al hacer click obtendremos lo siguiente:

Como podemos notar, automáticamente aparece entre las opciones “promedio” ya que es una de las más utilizadas. Esa es la que debemos seleccionar y presionar “aceptar” (si no apareciera la opción solamente debe escribir “promedio” en el área de “Buscar una función”).



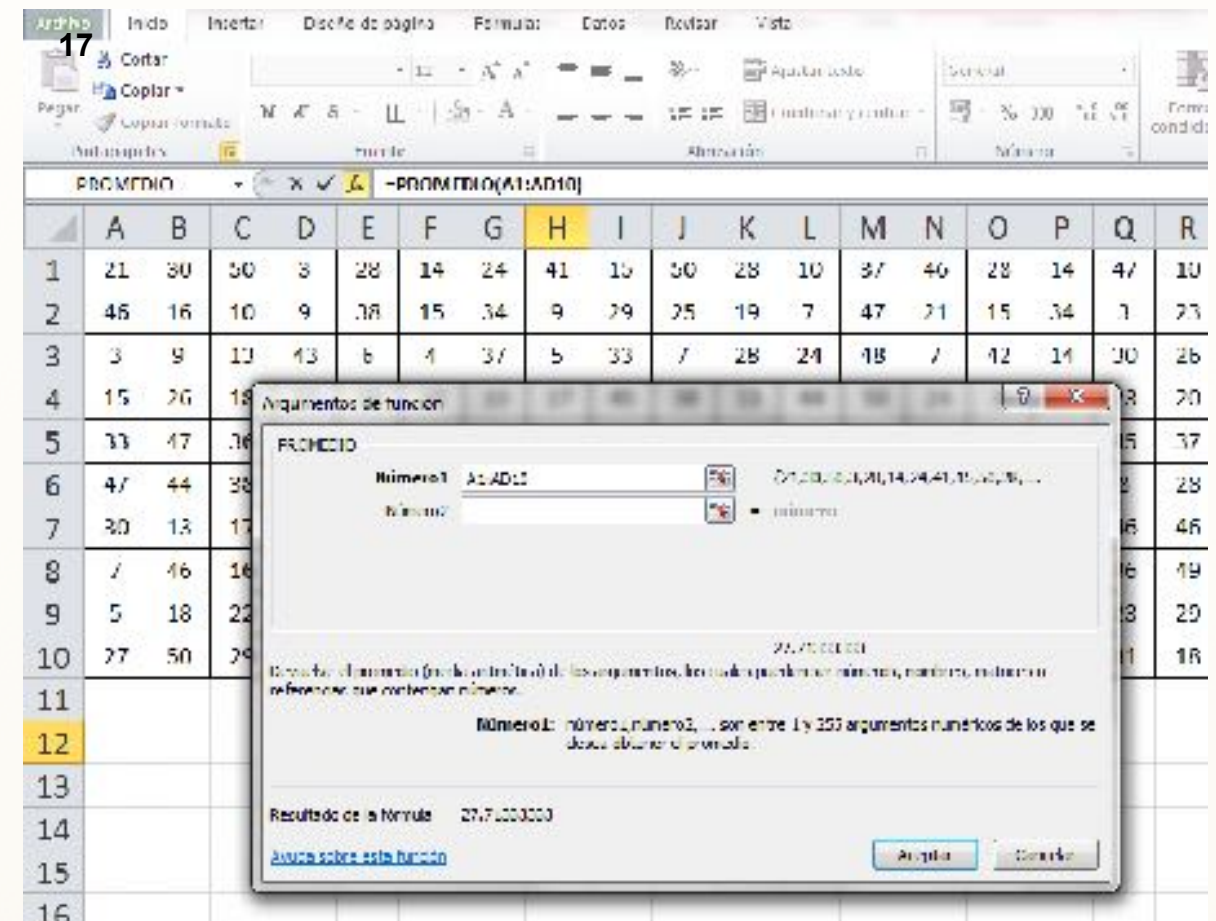
Ejemplo de datos, calcular media. Paso 2.

Al “*aceptar*” nos genera la pantalla para ingresar los argumentos a la función, aquí debemos seleccionar con el mouse toda el área en donde se encuentran nuestros datos. Esto se hace colocando el cursor tal y como aparece el ejemplo y luego con el mouse seleccionamos todos los datos de nuestra hoja de trabajo.



Ejemplo de datos, calcular media. Paso 3.

Si usted tiene más de dos áreas de trabajo de datos, entonces puede seleccionar el Número 2 y repetir el procedimiento y el programa calculará la media de todos los datos ingresados.



Ejemplo de datos, calcular media. Paso 4.

Al hacer clic en aceptar, obtendremos la media o promedio de nuestros datos en la celda en la que estábamos posicionados.

En este caso entonces la media es 27.713.

[illegible]

136

**Nota:**

Cuando tenemos un grupo de datos y lo seleccionamos con el mouse, abajo en la derecha Excel nos muestra 3 cálculos: 1) recuento de datos, es el total de datos que hay; 2) la suma; el programa suma todos los datos y genera el total y 3) ¡el promedio!



Comprobación del capítulo XII: Medidas de tendencia central

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XIII

Medidas de posición



Medidas de posición

Las medidas de posición son otra forma de representar a un conjunto de datos, al igual que en la tendencia central tiene sus ventajas y sus desventajas, en este caso la ventaja es su forma simple de cálculo y la desventaja es la pérdida de precisión.

Una característica de las medidas de posición es que requieren del ordenamiento de los datos de menor a mayor o viceversa, lo cual es fácil para conjuntos de datos pequeños pero a medida que crecen los datos el cálculo a mano puede complicarse y ser más propenso a errores de digitalización.

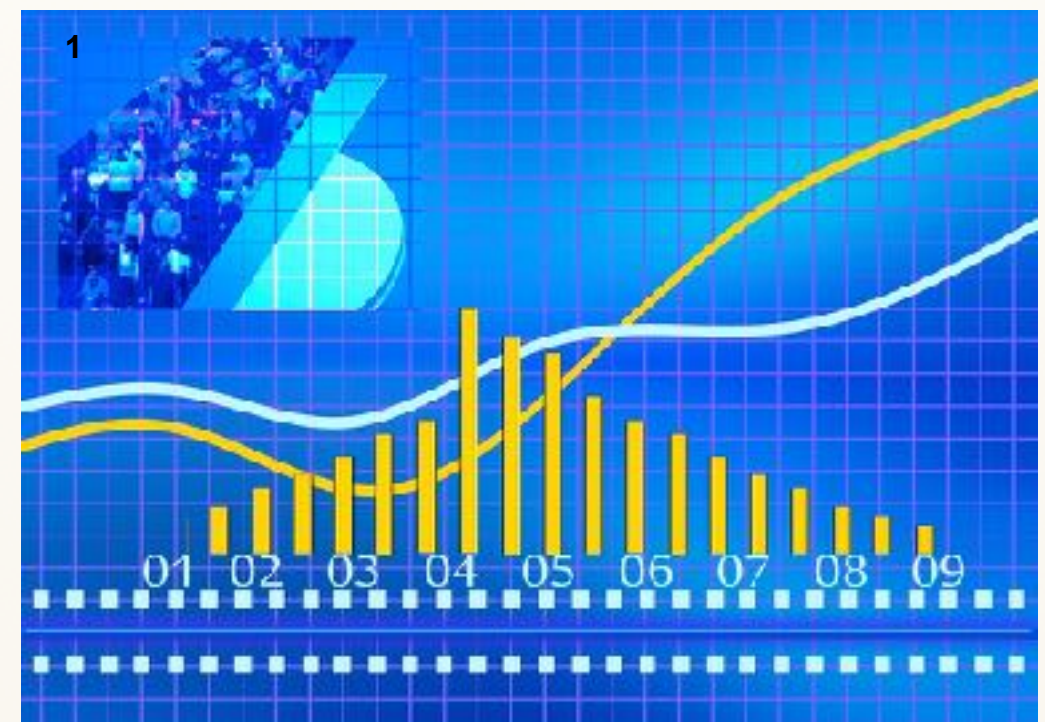
a. Mediana (Me)

Es una medida que únicamente funciona para variables cuantitativas, y permite representar al conjunto de datos de acuerdo al valor de su posición.



Definición y propiedades: La definición simple para la mediana es indicar el valor que se ubica en la posición del centro de todos los datos cuando estos se encuentran ordenados de menor a mayor o viceversa.

Otra forma de entenderlo es diciendo que la mediana es un valor que separa todos los datos en dos mitades. Una mitad de los datos serán menores que él y la otra mitad de los datos mayores que él. A éste punto medio, se le llama mediana.



La mediana

Cálculo: al igual que el cálculo de la moda, el cálculo de la mediana dependerá de si son datos agrupados o sin agrupar.

Para datos “sin agrupar” el procedimiento para obtener la mediana tiene a su vez dos métodos que dependen de si la cantidad de datos es par o impar.

Lo más simple es cuando tenemos datos IMPARES, después de ordenar los datos, únicamente debemos encontrar el dato que se ubica en la posición de en medio, es decir, cuando quedan la misma cantidad de datos antes y después de dicha posición.

Supongamos que tenemos estos datos 4, 5, 7, 9, 11. Son 5 datos (número impar) y fácilmente vemos que la tercera posición divide el grupo en 2 y está ocupada por el número 7. Esa es nuestra mediana.

Cuando la cantidad de datos es muy grande, el cálculo se puede hacer de la siguiente manera

Si $x_1, x_2, x_3, \dots, x_n$ son los datos de una muestra ordenada, y n es impar, la mediana de los datos toma el valor del dato en la posición, porque esta es la posición central.

2 Mediana para n impar

$$M_e = X_{\left(\frac{n+1}{2}\right)}$$

Mediana para n impar

Así:

En el ejemplo $n=5$, y los datos se pueden representar como:

$$x_1 = 4, x_2 = 5, x_3 = 7, x_4 = 9, x_5 = 11$$

$(n+1) / 2 = 3$ El dato en la tercera posición es la mediana así que revisamos los datos ordenados para encontrar que $X_3 = 7$.

Cuando la cantidad de datos (n) es PAR, el cálculo varía un poco porque deberemos obtener la media de los 2 datos que ocupan la posición central. Así tenemos que,

3 Mediana para n par

$$M_e = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

Mediana para n par

Veamos un ejemplo similar al anterior, tenemos esta serie de 6 datos ordenados:

2, 4, 5, 7, 9, 11 y usaremos la fórmula.

$\frac{n}{2} = 3$, , nos indica a la posición 3, en este caso el valor de la tercera posición es $x_3 = 5$.

$\frac{n}{2} + 1 = 4$, , nos indica a la posición 4, en este caso el valor de la cuarta posición es $x_4 = 7$.

Entonces: La media o el promedio de ambos valores centrales es

$$M_e = (x_3 + x_4)/2 = (5 + 7)/2 = 6$$

Ahora bien, para datos agrupados en intervalos o clases el cálculo puede ser un poco más complicado. El primer paso es identificar el intervalo en donde se encuentra la mediana simplemente dividiendo $n/2$, lo haremos en la columna de frecuencias absolutas de la distribución de frecuencias. Como podemos notar, necesitamos acceder a nuestra tabla de distribución de frecuencias porque vamos a utilizar:

L_i = el límite inferior del intervalo en dónde se encuentra la mediana

F_{i-1} = la frecuencia acumulada anterior al intervalo de la mediana

f_i = frecuencia del intervalo de la mediana

a_i = la amplitud de la clase

La fórmula que usaremos es la siguiente:

4 Mediana

$$M_e = L_i + \left(\frac{\frac{n}{2} - F_{i-1}}{f_i} \right) * c$$

Mediana

Usemos estos datos como ejemplo:

5

Intervalos	f_i	F_i
de 1 a 3	10	10
de 4 a 6	37	47
de 7 a 9	66	113
de 10 a 12	42	155
de 13 a 15	14	169

Datos

El primer paso consiste en encontrar el intervalo o clase de la mediana, $n/2 = 169/2 = 84.5$. En nuestra distribución de frecuencias identificamos que es intervalo de 7 a 9.

Siguiente paso, operamos el numerador $84.5 - F_i = 84.5 - 47 = 37.5$.

Ahora identificamos el denominador, $f_i = 66$

Así que podemos ya operar la división: $37.5/66 = 0.56$

Ya podemos operar la fórmula completa respetando jerarquías.

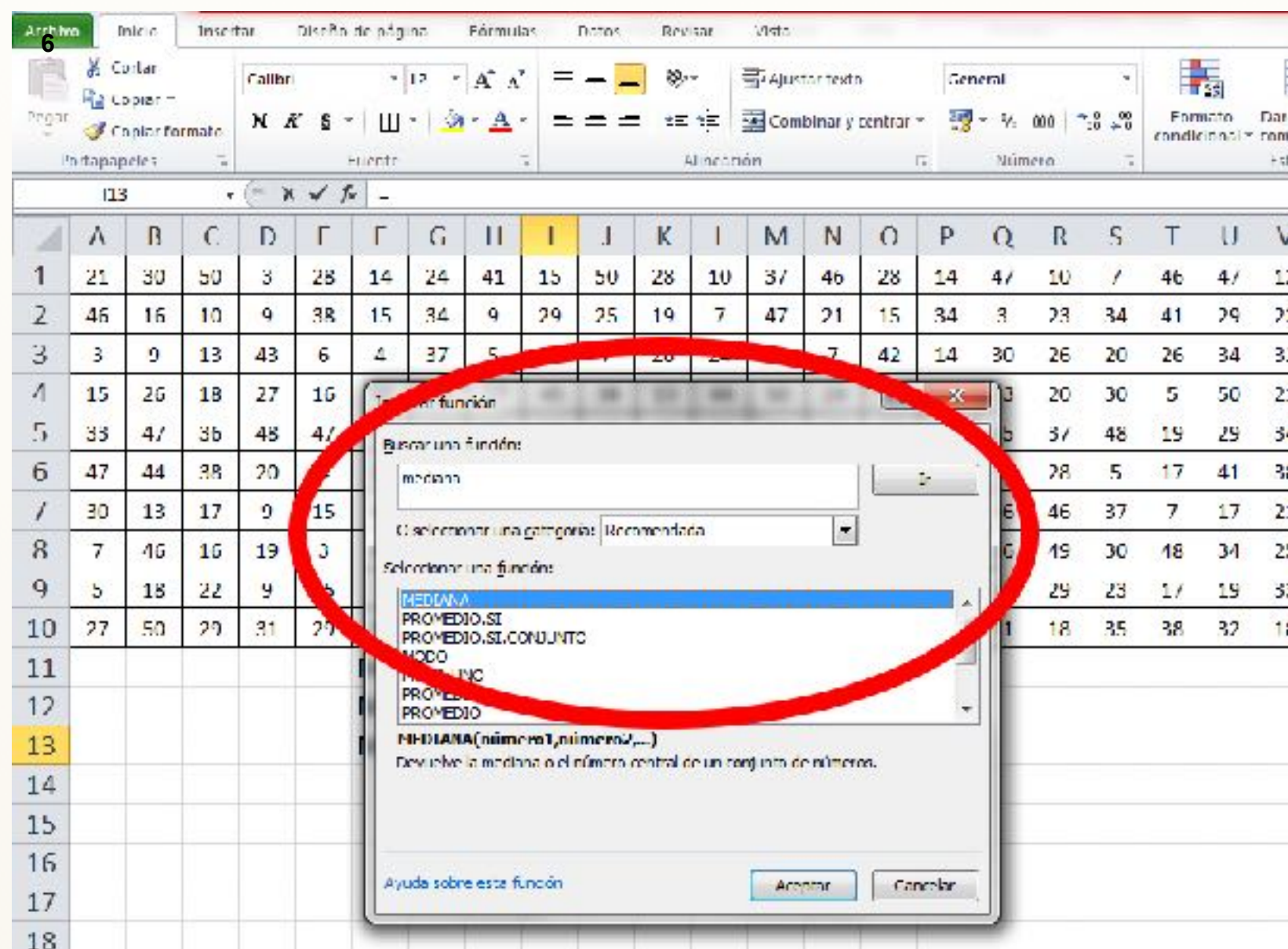
$$M_e = 7 + 0.56 * 3$$

$$M_e = 8.68$$

¡Esta es la mediana!

Uso de Excel: Aprendamos ahora a utilizar Microsoft Excel para calcular la mediana. Al igual que los casos anteriores, vamos a utilizar la misma serie de datos y la misma lógica inicial.

Empezaremos por posicionar nuestro cursor en la celda dónde queremos el resultado y luego haremos click en área de fx y escribiremos “mediana” para buscar la función.



Ejemplo de datos, calcular mediana. Paso 1

The screenshot shows an Excel spreadsheet with a data table in columns A through AD and rows 1 through 14. The formula bar at the top shows the function `=MEDIANA(A1:AD10)`. The results of the calculation are shown in rows 11, 12, and 13.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	21	30	50	4	28	14	24	41	15	50	28	10	37	46	28	14	47	10	7	46	47	12	33	34	31	46	34	46	29	25
2	46	16	10	9	38	15	24	9	29	25	19	7	47	21	15	34	3	23	34	41	29	22	42	40	47	50	49	28	34	13
3	3	9	13	43	6	4	37	5	33	7	28	24	48	7	42	14	30	26	20	26	34	32	41	17	4	13	48	48	10	28
4	15	26	18	27	16	38	33	17	45	38	13	44	50	24	6	15	28	20	30	5	50	21	4	21	47	16	41	14	46	24
5	33	47	36	48	47	4	27	49	14	15	46	35	34	19	22	41	45	37	48	19	29	34	30	32	38	12	27	36	5	28
6	47	44	38	20	4	25	31	28	9	29	36	39	10	30	14	7	8	28	5	17	41	38	38	13	32	14	38	27	42	33
7	30	13	17	9	15	44	40	42	49	43	18	40	30	32	50	6	46	46	37	7	17	21	49	29	8	23	43	9	34	19
8	7	46	16	19	8	44	28	24	32	3	40	27	50	36	13	36	36	49	30	48	34	25	12	50	39	14	8	18	40	25
9	5	18	22	9	45	7	38	23	47	5	8	45	10	9	46	26	23	29	23	17	19	33	48	37	27	29	27	50	45	36
10	27	50	29	31	29	39	15	41	44	32	27	39	3	46	17	32	31	18	35	38	32	18	8	35	26	12	13	50	34	11
11																														
12																														
13																														
14																														

Ejemplo de datos, calcular media. Paso 2.

El siguiente paso requiere que seleccionemos el área de nuestros datos en la pantalla de “argumentos de función”. Existe la opción de seleccionar más de 1 grupo de datos en caso de tener la información en tablas distintas e incluso en hojas de trabajo distintas.

Después de seleccionar todo el conjunto de datos se presiona “aceptar” y obtenemos la mediana en la celda que habíamos escogido.

En este caso, la mediana es 28.5

The screenshot shows the Excel interface with the **MEDIANA** function dialog box open. The dialog box displays the following information:

- Argumentos de función:**
 - Número1:** (21,30,30,3,28,14,24,41,15,50,28,10,37,16,28,14,47,10,7,46,47,12,33,33,31,46,33,16,29,25)
 - Número2:** (vacío)
- Resultado de la fórmula:** 28.5
- Descripción:** Devuelve la mediana y el número central de un conjunto de números.
- Nota:** Número1: número1, número2,... son de 1 a 255 números, nombres, matrices o referencias que contienen números, para los cuales desea obtener la mediana.

The background shows a grid of data from columns A to AD and rows 1 to 15. The formula bar shows **=MEDIANA(A1:AD10)**.

Ejemplo de datos, calcular media. Paso 3.

b. Cuartiles y Percentiles

Definición y propiedades: Si partimos de la mediana como una medida de posición que divide nuestro conjunto de datos en dos partes iguales, los cuartiles lo dividen en cuatro partes iguales y los percentiles en 100. También se utilizan los Quintiles (5) y los Deciles (10) pero con menor frecuencia.

Cuando hablamos de cuartiles, estaremos buscando partir el conjunto de datos ordenado en cuatro partes iguales estableciendo tres valores que partan o dividan los datos; es decir que cada cuartil tendrá 25% de los datos del conjunto.

Por consiguiente es muy importante tomar en cuenta, que el cuartil 2 coincida con la mediana. La nomenclatura para el valor que separa los cuartiles es el uso de la letra “Q” seguido del corte de cuartil que representa Q1, Q2, Q3.



Entonces, sabremos que si estamos hablando del valor Q3 sabremos que debajo de ese valor está el 75% de los datos y por encima el 25%, la misma lógica aplica para Q2 y Q3.

Cálculo: Para obtener cuartiles el procedimiento es muy similar al usado para la mediana. En primer lugar debemos ordenar los datos de menor a mayor y luego determinamos la posición en donde se ubica el corte para cada cuartil.



$$\frac{k \cdot N}{4}, k = 1, 2, 3$$

k representa el corte para los cuartiles, (si quiero el Q1 entonces sustituyo k por 1, para el Q2 la sustituyo por 2 y para el tercer corte Q3 sustituyo k por 3).

N es el total de datos en la distribución de frecuencias. Este valor, dirá la posición que se encuentra el dato, los cuales deben estar ordenados de menor a mayor.

**Nota:**

En este caso dividimos en 4 porque queremos cuartiles, pero aplica la misma fórmula para quintiles: dividiendo entre 5, para deciles dividiendo entre 10 o percentiles dividiendo entre 100.

Así pues, en el ejemplo que hemos estado desarrollando tenemos 300 datos.

Aplicamos la fórmula para Q1 así que:

$$1 * 300 / 4 = 75$$

$$Q1 = X(75).$$

Significa que en nuestros datos ordenados de mayor a menor, el que ocupa la posición 75 es Q1 y todos los valores por debajo son el 25% de los que existen en nuestros datos.

En el ejemplo siguiente la posición 75 está marcada en amarillo y sabemos entonces que el cuartil 1 es 15, si estos datos fueran notas, sabríamos que alguien que tiene 15 o menos, estará en el 1er cuartil, en otras palabras el 25% más bajo.

**Nota:**

Por el formato escrito, los datos se presentan en columnas, regularmente se encontrarían en una larga columna de 300 casos. El conteo se realiza por columnas, contando de arriba hacia abajo, desde el dato menor hasta el mayor.

Se incluyen filas y columnas en gris para facilitar el conteo.

9

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
2	3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50
4	3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50
5	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50
6	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
7	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
8	4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50
9	4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50
10	4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50

Ejemplo de datos, Q1

Para hacer el cálculo de Q2: $2 * 300 / 4 = 150$

El procedimiento es el mismo, buscamos la posición 150 en los datos ordenados para encontrar el Q2, y recordemos, es igual al valor de la mediana. Si estas fueran notas, un alumno con nota entre $Q1=15$ y $Q2=28$ está en el segundo cuartil.

10

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
2	3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50
4	3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50
5	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50
6	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
7	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
8	4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50
9	4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50
10	4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50

Ejemplo de datos, Q2

Para hacer el cálculo de de la posición de Q3:

$3 * 300 / 4 = 225$

El procedimiento es el mismo, buscamos la posición 225 en los datos ordenados para encontrar el Q3, en este caso la posición 225 le

corresponde el 39. Este es nuestro Q3, que dividirá al tercer del cuarto cuartil. Entonces, si fueran notas, todo alumno entre Q2=28 y Q3=39 está en el tercer cuartil y todos aquellos con más de 39 están el cuarto cuartil, es decir están el 25% más alto de las notas registradas.

11

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
2	3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50
4	3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50
5	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50
6	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
7	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
8	4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50
9	4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50
10	4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50

Ejemplo de datos, Q3

Cuando se trata de datos en intervalos o clases, el procedimiento también es similar al que usamos para calcular la mediana, únicamente que ahora estaremos utilizando el k como indicador de posición para ubicarnos. La fórmula es la siguiente:

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i \quad k = 1, 2, 3$$

Donde

L_i = el límite inferior del intervalo en donde se encuentra Q

F_{i-1} = la frecuencia acumulada anterior al intervalo de Q

f_i = frecuencia del intervalo de la Q

a_i = la amplitud de la clase

Así que en el ejemplo que estamos trabajando para encontrar Q_1 :

12

Intervalos	f_i	F_i
de 1 a 3	10	10
de 4 a 6	37	47
de 7 a 9	66	113
de 10 a 12	42	155
de 13 a 15	14	169

Ejemplo de datos, calcular cuartiles

Primer paso, operamos la primera parte del numerador: para ello operamos la fórmula de posición Q_1 que produce un resultado de 42.25 ese dato está en el intervalo de 4 a 6, así que podemos deducir que F_{i-1} entonces es 10.. por lo tanto el numerador es $42.25 - 10 = 32.25$.

Ahora podemos terminar el ejercicio, $L_i = 4$, nuestro $f_i = 37$ y $a_i = 3$.

Entonces $4 + (11/37) \cdot (3) \dots$ por la ley de jerarquías dividimos, multiplicamos y luego sumamos 4.86 es Q_1 .

**Nota:**

Para obtener Q2, Q3, solamente se sustituye en el numerador la k.

Uso de Excel: En Excel, los cuartiles se obtienen fácilmente, no es necesario buscar las posiciones pues el programa lo hace por nosotros y devuelve el resultado. Lo único que necesitamos hacer es conocer la mecánica para operar el programa.

**Nota:**

Para este ejemplo utilizaré otra versión de Excel, para mostrar que aunque las pantallas aparecen en lugares diferentes,

Para iniciar, haremos lo mismo que hemos realizado anteriormente, escoger una celda en blanco y colocar el cursor allí, luego ingresar al área de fórmulas fx. Cuando aparece el nuevo menú escribiremos CUARTIL para buscar la fórmula y la seleccionamos haciendo clic en “insertar función”.

13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
2	3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	3	5	7	8	10	12	13	15													36	37	38	41	43	45	46	47	48	50
4	3	5	7	8	10	12	13	15													36	37	38	41	43	45	46	47	48	50
5	3	5	7	9	10	13	14	15													36	38	39	41	43	45	46	47	49	50
6	3	5	7	9	10	13	14	15													36	38	39	41	43	45	46	47	49	50
7	3	5	7	9	10	13	14	15													36	38	39	41	43	45	46	47	49	50
8	4	5	7	9	10	13	14	15													36	38	39	41	44	46	46	47	49	50
9	4	6	7	9	11	13	15	16													36	38	40	42	44	46	46	48	49	50
10	4	6	8	9	12	13	15	16													37	38	40	42	44	46	47	48	49	50

Q1 =

Insertar función

Buscar una función:

CUARTIL

Seleccionar una categoría:

Documentos

Seleccionar una función:

CUARTIL

CUARTIL.DISC

CUARTIL.PEC

AGREGAR

CUARTIL(matriz;cuartil)

Esta función es la disponible por compatibilidad con Excel 2007 y versiones anteriores.

Devuelve el cuartil de un conjunto de datos.

Ayuda sobre esta función

Aceptar Cancelar

Ejemplo de datos, calcular cuartiles

Luego aparece una pantalla nueva, en donde debemos seleccionar toda el área en donde están nuestros datos. Allí debemos ingresar que Cuartil deseamos obtener. En este ejemplo colocaré 3.

14

Argumentos de función

MEDIANA

Número1: A1:AD10 = {21,30,50,3,28,14,24,41,15,50,28,10,37,16,28,14,47,10,7,46,47,12,33,33,31,46,33,16,29,25}

Número2: = número

25.5

Devuelve la mediana o el máximo central de un conjunto de números.

Número1: número1, número2,... son de 1 a 255 números, nombres, referencias o referencias que contienen números, para los cuales desea obtener la mediana.

Resultado de la fórmula = 25.5

[Ayuda sobre esta función](#)

Aceptar Cancelar

Ejemplo de datos, calcular cuartiles

The screenshot shows an Excel spreadsheet with data in columns A to AD and rows 1 to 10. The formula bar displays the formula `=CUARTIL(A1:AD10,3)`. The result, 39, is shown in cell AF3. The 'Generador de fórmulas' (Formula Generator) pane on the right shows the function 'CUARTIL' with the following settings:

- matriz: A1:AD10
- cuartil: 3

The result of the calculation is 39, as indicated by the 'Resultado: 39' at the bottom of the pane.

Ejemplo de datos, resultado.

El programa hará el cálculo y mostrará el resultado en la casilla que seleccionamos.

Podemos verificar que el Q3 es 39 tal y como lo hicimos con el cálculo a mano.

Percentiles

Ahora vamos a revisar cómo trabajar los percentiles, la lógica es exactamente la misma, por lo tanto no vamos a detenernos tanto en la mecánica ya que en lugar de dividir en 4 partes iguales, ahora podemos hacerlo en 100 partes iguales.

La operación es la misma fórmula que ya conocemos.

Para datos regulares:

$$\frac{k \cdot N}{100}, \quad k = 1, 2, \dots, 99$$

Así pues, si usamos nuestra misma tabla de datos y queremos el percentil 63 solamente operaremos:

$66 \cdot 300 / 100 = 198$ y esta es la posición del percentil 66. Al buscarla en nuestros datos encontramos que el percentil corresponde al 34.

16

2	4	6	8	9	12	13	15	17	17	20	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
3	4	7	8	9	12	13	15	17	18	20	22	24	25	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50
3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50
3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50
3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50
4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50
4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50

Ejemplo de datos, calcular percentil

Para datos agrupados en intervalos la fórmula es la siguiente:

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \quad k = 1, 2, \dots, 99$$

Donde:

L_i = el límite inferior del intervalo en dónde se encuentra P_k

F_{i-1} = la frecuencia acumulada anterior al intervalo de P_k

f_i = frecuencia del intervalo de la P_k

a_i = la amplitud de la clase

17

Intervalos	f_i	F_i
de 1 a 3	10	10
de 4 a 6	37	47
de 7 a 9	66	113
de 10 a 12	42	155
de 13 a 15	14	169

Ejemplo de datos, calcular percentil.

Este es el ejemplo que utilizaremos y queremos conocer el percentil 80.

Trabajamos el numerador, identificamos $N = 169$, así que $80 \cdot 169 / 100 = 135.2$.

Logramos identificar al intervalo de 10 a 12 años. Así que $F_{i-1} = 113$ y hacemos la resta $135.2 - 113 = 22.2$

Ahora podemos operar: sabemos que $L_i = 10$; $f_i = 42$ y $a_i = 3$

$$P_{k80} = 10 + (22.2/42) \cdot 3$$

$$P_{k80} = 10 + (1.58)$$

$$P_{k80} = 11.58$$

Así pues, el percentil 80 para estos intervalos de datos es 11.58 y si pensamos que estamos hablando de edades de pacientes de un pediatra, podremos concluir que 80% de sus pacientes son menores de 11 años y medio.

En Excel, el cálculo es muy sencillo, especialmente si consideramos que ya hemos trabajado mediana y cuartiles. El único detalle MUY importante que debemos considerar es que Excel entiende los percentiles como datos desde 0.1 hasta 1. Por lo tanto debemos dividir nuestro percentil entre 100 para dar la instrucción. Por ejemplo el P_{k80} que trabajamos debe pedirse como .80

Veamos la mecánica, para iniciar haremos lo mismo que siempre, colocamos el cursor en la celda en la que deseamos el resultado y entramos al panel de fórmulas. Allí buscamos percentil y obtendremos esta pantalla.

18

PERCENTIL

fx

=PERCENTIL(A1:AD10,.8)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50		
2	3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50		
3	3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50	Pk80	:=AD10,.8)
4	3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50		
5	3	5	7	9	10	13	14	15	17	18	21	3	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50		
6	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50		
7	3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50		
8	4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50		
9	4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50		
10	4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50		
11																																
12																																
13																																

Generador de fórmulas

Mostrar todas las funciones

PERCENTIL

matriz = (2;4;6;8;9;12;13;15;17;17;20;22;24...

A1:AD10

k = 0.8

.8

Resultado: 42

Listo

Percentil

19

Insertar función

Quitar una función: PERCENTIL

Seleccionar una función:

- RANGO PERCENTIL
- RANGO PERCENTIL INC
- PERCENTIL EXC
- PERCENTIL**
- PERCENTIL INC
- RANGO PERCENTIL EXC
- ARTESAN

PERCENTIL (matriz)

Esta función está diseñada para compatibilidad con Excel 2007 y versiones anteriores.

Devuelve el percentil k-ésimo de los valores de un rango.

[Ayuda sobre esta función](#)

Aceptar Cancelar

PK80 =

Generador de formulas

Aquí debemos marcar con el mouse toda el área de nuestros datos y en el área que indica k = número debemos pedir el percentil que deseamos recordando SIEMPRE que Excel necesita el dato en la expresión proporcional, es decir nuestro percentil dividido 100. Así pues, en este ejemplo colocaremos .8 en esa casilla tal y como podemos ver en la siguiente imagen.

Si nos fijamos bien, al ingresar el percentil que deseamos incluso antes de presionar la tecla enter o presionar el botón “*listo*” el programa nos indica el resultado en la esquina inferior izquierda del menú de generación de fórmulas.. el resultado es 42.

Al presionar “*enter*” el resultado aparecerá en la casilla seleccionada.



Comprobación del capítulo XIII: Medidas de posición

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XIII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XIV

Medidas de variabilidad



Medidas de variabilidad

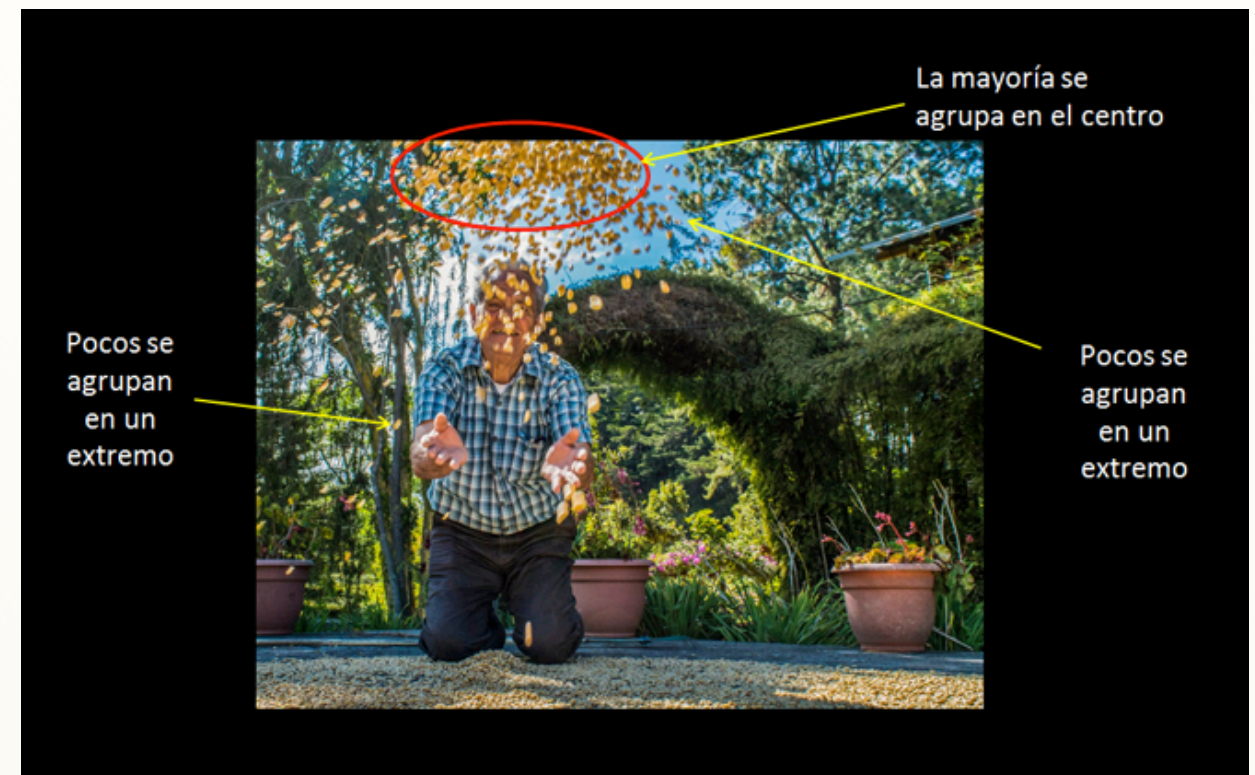
Para hablar de medidas de variabilidad debemos hacer referencia a la tendencia central.

Recordemos el caso de Manolo, el agricultor que toma una muestra de café y lo lanza al aire. Ese café muestra una tendencia central de distribución, la mayoría se agrupan por el centro y a medida que nos alejamos del centro encontramos menos granos de café. Mencionamos que esa tendencia de la foto, es la tendencia normal en la que aparecen los datos en ciencias sociales.

Ahora bien, esa tendencia central se puede medir y usar los datos para representarla sin embargo esas medidas requieren de las medidas de variabilidad para que su interpretación sea más exacta.



Veamos la imagen nuevamente:



Historia de Manolo y el café, la dispersión

La intención de las medidas que hemos analizado en esta unidad es representar un grupo a través de un dato (la media, la mediana o percentiles, por ejemplo) así que podríamos decir que en promedio el 80% de los granos de café están en el centro, o a partir de la mano derecha de Manolo inicia el percentil 90, por lo que sabemos que únicamente hay 10% más allá de su mano.

Sin embargo representar un grupo de datos únicamente por la tendencia puede ser riesgoso ya

que no podemos conocer el panorama completo. Veamos estas dos imágenes de Manolo.



Historia de Manolo y el café, la dispersión

Si nos fijamos en las fotos y asumimos que cada grano de café es la nota de un estudiante en la clase de lectura, podemos asumir que entre más a la izquierda esté el grano de café (la nota) más baja será la nota, y entre más a la derecha esté más alta será.

Entonces, les diré que el promedio de ambos resultados es 75 puntos. ¿Creen ustedes que esa media o promedio representa lo mismo? La verdad es que los representa pero de forma incompleta,

porque la foto de la izquierda tiene granos de café más concentrados en el centro y pocos en los extremos eso significaría que la mayoría de resultados están cerca de 75, pero la foto de la derecha, tiene más “DISPERSIÓN” hay datos en los extremos y eso significa que el promedio no es tan explicativo.

Para hacer el ejemplo más claro, ¿qué esperaríamos del maestro de lectura? Qué todos tengan un rendimiento similar o que algunas aprendan mucho y otros aprendan poco?

Pues para describir mejor los datos en estas situaciones, existen las medidas de dispersión.

a. Amplitud o Rango

Definición y propiedades: Por amplitud o rango nos referimos a la distancia que existe entre el dato de menor valor y el dato de mayor valor. El resultado nos permite tener una idea de la dispersión de los datos.

Cálculo: la Amplitud se denomina como A_t y se calcula restando el dato máximo menos el dato mínimo. Su expresión se ve así:

$$A_t = X_{\max} - X_{\min}$$

2	4	6	8	9	12	13	15	17	17	20	22	24	26	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	50
3	4	7	8	9	12	13	15	17	18	20	22	24	26	28	29	30	31	33	34	35	37	38	41	42	45	46	47	48	50
3	5	7	8	10	12	13	15	17	18	21	22	24	27	28	29	30	31	33	34	36	37	38	41	43	45	46	47	48	50
3	5	7	8	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	37	38	41	43	45	46	47	48	50
3	5	7	9	10	13	14	15	17	18	21	3	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	48	50
3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
3	5	7	9	10	13	14	15	17	18	21	23	24	27	28	29	30	32	33	34	36	38	39	41	43	45	46	47	49	50
4	5	7	9	10	13	14	15	17	19	21	24	24	27	28	29	30	32	33	34	36	38	39	41	44	46	46	47	49	50
4	6	7	9	11	13	15	16	17	19	22	24	25	27	28	29	30	32	33	34	36	38	40	42	44	46	46	48	49	50
4	6	8	9	12	13	15	16	17	19	22	24	25	28	28	29	31	32	34	35	37	38	40	42	44	46	47	48	49	50

En estos datos calcularíamos $50 - 2 = 48$.

Así pues, si alguien nos dice que la media o promedio de estos datos es 45 y su amplitud 48, nosotros sabremos que existen muchos datos cercanos a 45, pues el promedio nos indica el centro de los valores y la amplitud de 48 nos dice el rango en el que se encuentran todos los datos, describe que tan “regados” se encuentran éstos.

Uso de Excel: El uso de Excel en este cálculo realmente es el de una calculadora normal. NO existe función para obtener el rango pues su cálculo es simple.

b. Desviación media

Definición y propiedades: Se conoce también como “desviación promedio”, “desviación absoluta promedio” o simplemente “Desviación Media” y se entiende por ella el promedio de la diferencia absoluta de cada dato y la media del conjunto de datos.



Muy importante tener en cuenta el término “diferencia absoluta” eso significa que no usamos datos negativos, si la resta genera un -5 el valor absoluto es 5.

Entonces sabemos que la desviación media es otra forma que nos permite conocer qué tanto se alejan los datos de la media.

Se expresa como:

$$D_{\bar{x}}$$

Cálculo: su operación es bastante simple con pocos datos, simplemente se requiere restarle la media aritmética a cada valor y luego obtener la media aritmética de esas diferencias. Su expresión matemática se ve así:

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{N}$$

Donde x_i es cada valor en nuestro registro de datos.

Donde $\bar{x}$ es la media aritmética de nuestros datos.

Haremos un ejemplo y para ello usaremos un grupo de datos obtenidos de la medición de la estatura de los 20 aspirantes al equipo de baloncesto de la ciudad.

Sabemos además que la estatura promedio (media aritmética) actual es de 1.92, así que vamos a revisar los datos:

Estatura en mls		Estatura en mls	
Aspirante 1	1.90	Aspirante 11	1.93
Aspirante 2	1.90	Aspirante 12	1.91
Aspirante 3	1.85	Aspirante 13	1.90
Aspirante 4	1.78	Aspirante 14	1.84
Aspirante 5	1.98	Aspirante 15	1.88
Aspirante 6	1.86	Aspirante 16	1.83
Aspirante 7	1.79	Aspirante 17	1.84
Aspirante 8	1.85	Aspirante 18	1.87
Aspirante 9	1.82	Aspirante 19	1.85
Aspirante 10	1.87	Aspirante 20	1.80

Media aritmética

Como podemos ver, tenemos el registro de 20 aspirantes con su respectiva estatura así que la operación que requerimos primero es calcular la media aritmética tal y como ya aprendimos así que obtenemos el promedio de 1.8625. Así que nos toca restar a cada dato el valor de la media aritmética y realizar la resta, luego obtenemos su valor absoluto.

5	Estatura en mts	Media aritmética	Diferencia ⁺	Diferencia absoluta ⁺
Aspirante 1	1.9	1.85	0.04	0.04
Aspirante 2	1.9	1.85	0.04	0.04
Aspirante 3	1.85	1.85	-0.01	0.01
Aspirante 4	1.78	1.85	-0.08	0.08
Aspirante 5	1.98	1.85	0.12	0.12
Aspirante 6	1.86	1.85	0	0
Aspirante 7	1.79	1.85	-0.07	0.07
Aspirante 8	1.85	1.85	-0.01	0.01
Aspirante 9	1.82	1.85	-0.04	0.04
Aspirante 10	1.87	1.85	0.01	0.01
Aspirante 11	1.93	1.85	0.07	0.07
Aspirante 12	1.91	1.85	0.05	0.05
Aspirante 13	1.9	1.85	0.04	0.04
Aspirante 14	1.84	1.85	-0.02	0.02
Aspirante 15	1.88	1.85	0.02	0.02
Aspirante 16	1.83	1.85	-0.03	0.03
Aspirante 17	1.84	1.85	-0.02	0.02
Aspirante 18	1.87	1.85	0.01	0.01
Aspirante 19	1.85	1.85	-0.01	0.01
Aspirante 20	1.8	1.85	-0.05	0.05

Diferencia absoluta



Nota:

Al obtener las diferencias se puede comprobar una de las propiedades de la media ya que su suma será 0.

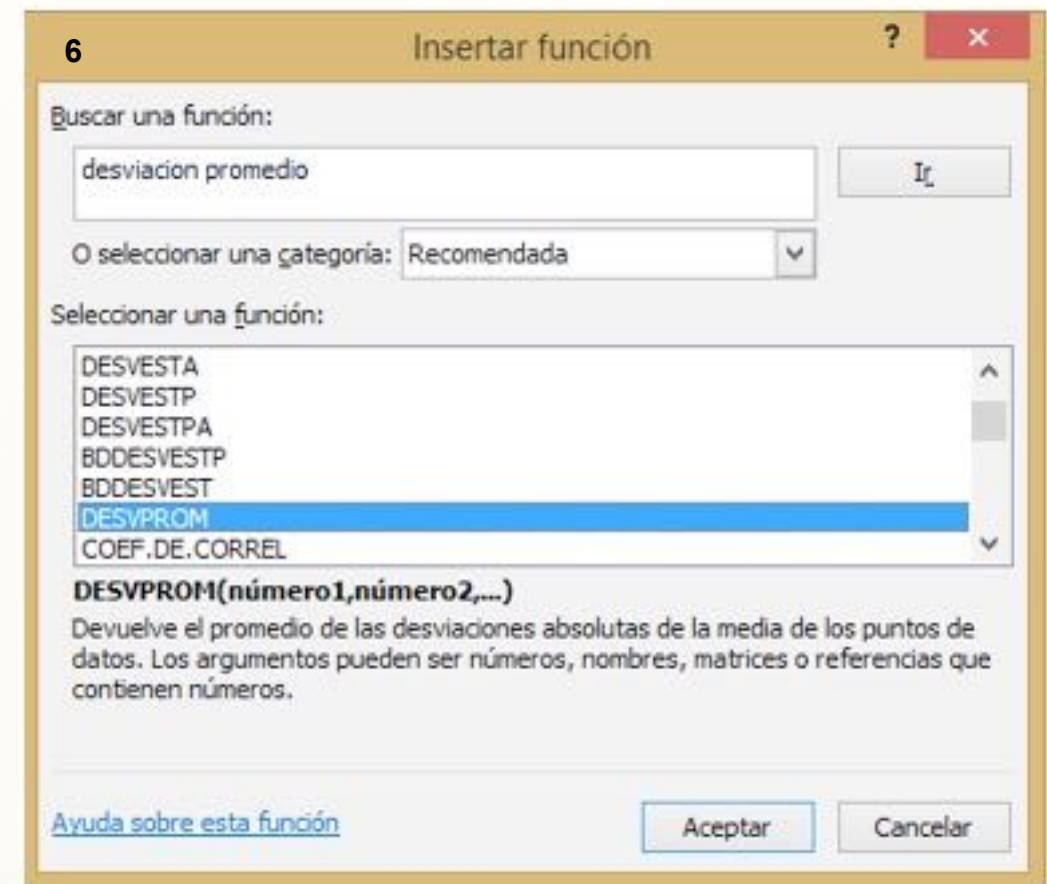
El último paso es obtener la media aritmética de las diferencias absolutas, en este caso así que usamos los conocimientos adquiridos anteriormente y obtenemos la media de las diferencias absolutas de 0.0375.

Este valor es la desviación media

$$D_{\bar{x}} = 0.0375$$

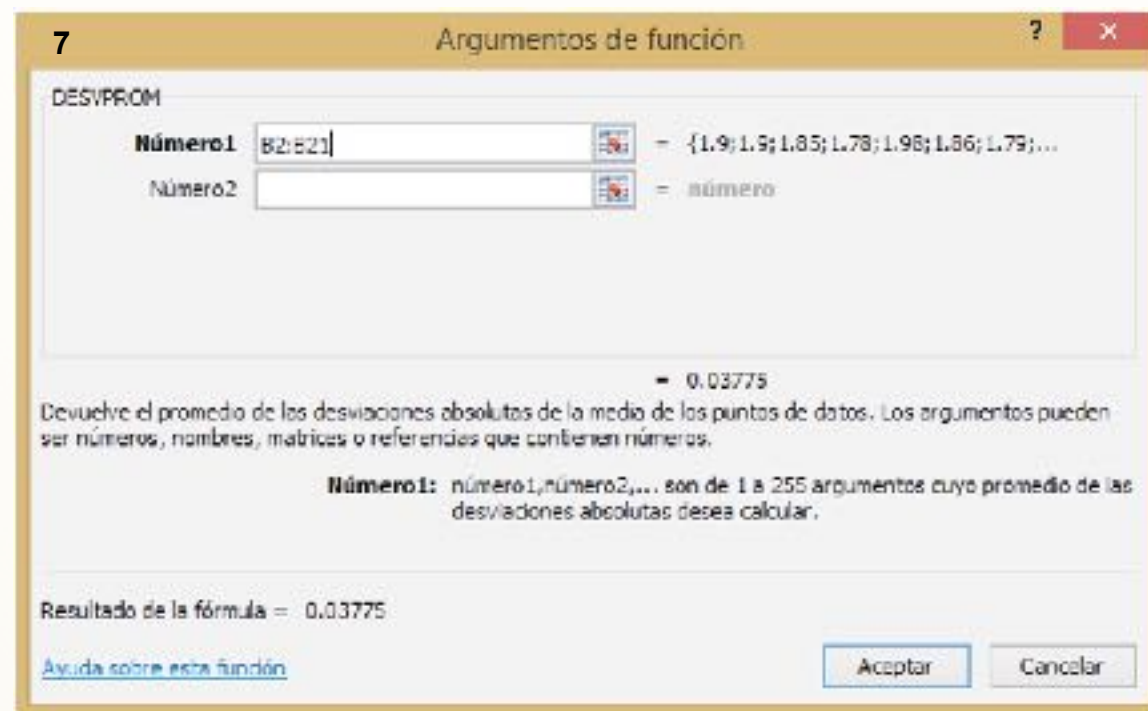
Uso de Excel: Después de hacer los cálculos paso por paso y de la misma forma en la que hemos trabajado anteriormente, podemos saber que Excel nos simplifica la vida para hacer estos cálculos y esto es especialmente importante cuando tenemos grandes cantidades de registros.

Importante tomar en cuenta que Excel conoce al cálculo como Desviación Promedio, así que seguiremos los pasos de los capítulos anteriores, ingresando al área de funciones fx y buscamos desviación promedio, la encontraremos como DESVPROM.



Insertar función

Al seleccionarla aparecerá el menú de argumentos para la función en donde debemos seleccionar el área completa de nuestros datos, aquí ya podemos ver el resultado en la esquina inferior izquierda, sin embargo al presionar “aceptar” obtendremos el resultado en la casilla que hemos seleccionado.



Argumentos de función.

Debido al redondeo, el resultado puede variar levemente en algunas decimales.

c. Desviación estándar

También conocida como “desviación típica” por excelencia es la medida de variabilidad más usada. Al inicio de la unidad decíamos que los promedios o medias son las medidas más comunes en el léxico estadístico y todos nos hemos acostumbrado a reportar el promedio de notas, el promedio de estaturas, el promedio del tiempo, el promedio de la distancia, en fin.. la media aritmética es la estrella.

Sin embargo, también hemos mencionado que podemos describir de mejor forma los datos si aunados a un dato que nos indique la tendencia, tenemos información del grado de dispersión de los datos.

Para esta situación el mundo de las matemáticas y estadísticas indica que siempre que reportemos una media/promedio, DEBEMOS reportar su desviación estándar.

Definición y propiedades:

Entenderemos por desviación estándar como una medida de qué tanto se alejan de la media aritmética, los datos de una distribución. Es la variación o dispersión de datos alrededor de su promedio.



En términos matemáticos, es la raíz cuadrada de la varianza, sin embargo no entraremos en los detalles para diferenciarlas pues en ciencias sociales se utiliza la desviación estándar.

Se representa con la letra griega sigma “ σ ” o simplemente con una “s” y tiene algunas propiedades importantes de mencionar:

- Siempre será un valor igual o mayor que cero.
- Si a todos los valores se les suma el mismo valor, la desviación no cambia.
- Si todos los valores son multiplicados por el mismo valor, la desviación no cambia.
- Se puede calcular la desviación total de varias distribuciones individuales.

Cálculo:

El cálculo de la desviación se puede volver un poco engorroso si se hace completamente a mano, basta con ver la fórmula para asustarse:

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N}}$$

Sin embargo vamos a trabajar por partes y desarrollando a partir de los conocimientos que hemos adquirido para ver que no estamos hablando en chino y que con la ayuda de una computadora podemos resolver esta telaraña.

Tendremos que entender que la desviación estándar es la raíz cuadrada de la media aritmética del cuadrado de las diferencias absolutas... al verlo así podemos deducir que el inicio del cálculo es igual al de la desviación media, posteriormente elevaremos al cuadrado las diferencias y obtendremos el promedio.

Veamos el mismo ejemplo que usamos anteriormente sobre las estaturas de los aspirantes al equipo de baloncesto:

8	Estatura en mts.	Media aritmética	Diferencia
Aspirante 1	1.9	1.86	0.04
Aspirante 2	1.9	1.86	0.04
Aspirante 3	1.85	1.86	0.01
Aspirante 4	1.78	1.86	-0.08
Aspirante 5	1.98	1.86	0.12
Aspirante 6	1.86	1.86	0
Aspirante 7	1.79	1.86	-0.07
Aspirante 8	1.85	1.86	-0.01
Aspirante 9	1.82	1.86	-0.04
Aspirante 10	1.87	1.86	0.01
Aspirante 11	1.93	1.86	0.07
Aspirante 12	1.91	1.86	0.05
Aspirante 13	1.9	1.86	0.04
Aspirante 14	1.84	1.86	-0.02
Aspirante 15	1.88	1.86	0.02
Aspirante 16	1.83	1.86	-0.03
Aspirante 17	1.84	1.86	-0.02
Aspirante 18	1.87	1.86	0.01
Aspirante 19	1.85	1.86	0.01
Aspirante 20	1.8	1.86	-0.06

Cuadrado de diferencias.

Como podemos ver aquí tenemos nuevamente la tabla de datos hasta el punto de calcular la diferencia. Ahora bien, en lugar de calcular el valor absoluto de dicha diferencia, lo que haremos será calcular el cuadrado de cada una de las diferencias y sumaremos esa columna.

9	Estatura en mts	Media aritmética	Diferencia	Cuadrado de diferencias
Aspirante 1	1.9	1.86	0.04	0.0016
Aspirante 2	1.9	1.86	0.04	0.0016
Aspirante 3	1.85	1.86	-0.01	0.0001
Aspirante 4	1.78	1.86	-0.08	0.0064
Aspirante 5	1.98	1.86	0.12	0.0144
Aspirante 6	1.86	1.86	0	0
Aspirante 7	1.79	1.86	-0.07	0.0049
Aspirante 8	1.85	1.86	-0.01	0.0001
Aspirante 9	1.82	1.86	-0.04	0.0016
Aspirante 10	1.87	1.86	0.01	0.0001
Aspirante 11	1.93	1.86	0.07	0.0049
Aspirante 12	1.91	1.86	0.05	0.0025
Aspirante 13	1.9	1.86	0.04	0.0016
Aspirante 14	1.84	1.86	-0.02	0.0004
Aspirante 15	1.88	1.86	0.02	0.0004
Aspirante 16	1.83	1.86	-0.03	0.0009
Aspirante 17	1.84	1.86	-0.02	0.0004
Aspirante 18	1.87	1.86	0.01	0.0001
Aspirante 19	1.85	1.86	-0.01	0.0001
Aspirante 20	1.8	1.86	-0.06	0.0036
SUMATORIA				0.0457

La desviación estándar es la raíz cuadrada de la media aritmética del cuadrado de las diferencias absolutas...

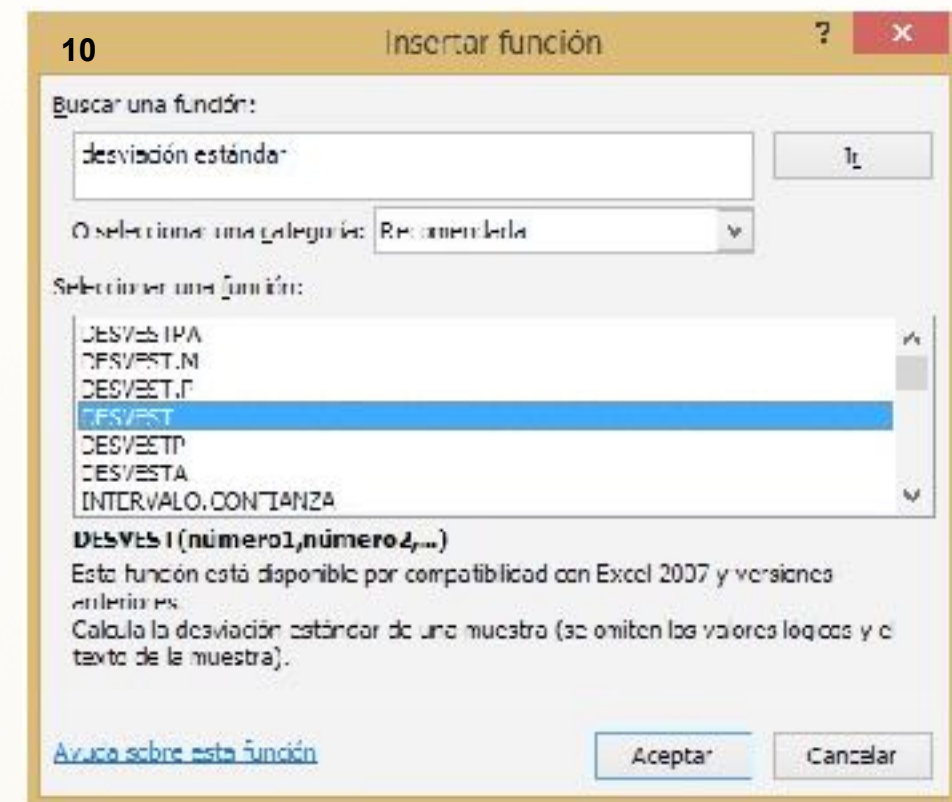
Entonces, tenemos la sumatoria de los cuadrados de las diferencias entre el dato observado y la media. Ahora debemos dividirlo en N-1 (es una corrección estadística para asegurar la representatividad de los datos).

Entonces $N=20 - 1$ nos queda 19 como denominador.

.0457 dividido entre 19 = 0.0024

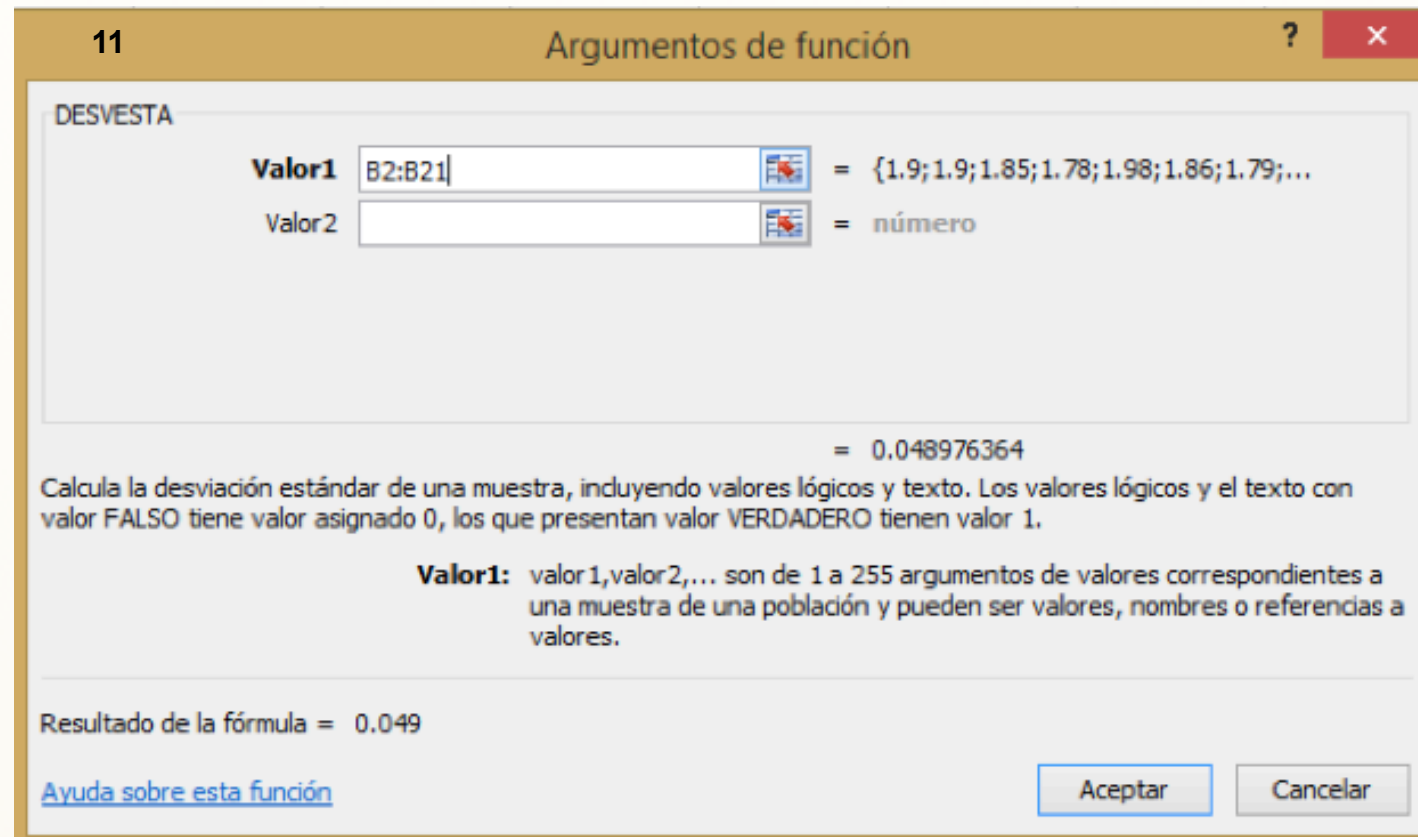
El último paso para obtener la Desviación Estándar es obtener la raíz cuadrada de este número = 0.049... Esta es la desviación estándar.

Uso de Excel: Agradeceremos a Excel por simplificarnos la vida y haremos el cálculo de la forma más sencilla y con la lógica que hemos ya adquirido en este recorrido. Ingresaremos al área de funciones fx y buscaremos “*desviación estándar*”.



Insertar función

Encontraremos la función DESVEST que será la que necesitaremos para trabajar. Al hacer aceptar ingresaremos al área de argumento para la función y deberemos seleccionar el área en donde están nuestros datos (los originales, recolectados).



Argumentos de función

Desde este momento en la esquina inferior izquierda podremos ver el resultado, pero al presionar aceptar veremos el dato en la celda seleccionada. En este caso el resultado es 0.049.



Comprobación del capítulo XIV: Medidas de variabilidad

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XIV

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Unidad V

La curva normal



Capítulo XV

La curva normal como probabilidades



La curva normal como probabilidades

Esta unidad tiene como objetivo principal, conocer y comprender este concepto de la curva normal. Para este efecto debemos aclarar que se le conoce también con otros nombres, tales como Distribución normal, Campana de Gauss, Distribución Gaussiana.

Su importancia radica en que permite modelar fenómenos sociales, naturales y psicológicos y hacer cálculos al respecto de los mismos. Para muchos autores es considerada como la distribución más importante en estadística.



Su capacidad de representar variables aleatorias continuas que son producto de medición la hacen especialmente relevante para la investigación y puede clasificarse dentro de las funciones de densidad de probabilidad continua. Los cálculos de dichas funciones requieren de conocimientos de cálculo integral y salen de los alcances de esta

unidad, pero los grandes estadistas han desarrollado tablas de probabilidad y procesos simplificados que estudiaremos para poder utilizar el área debajo de la curva con fines de cálculos de probabilidad.

Para fines de conocimiento general, la distribución normal se representa de la siguiente manera:

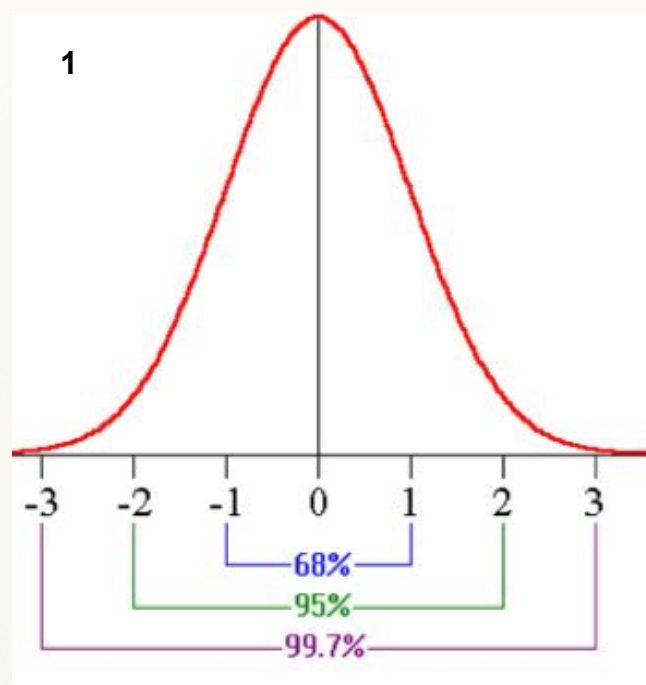
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Donde:

- X es el valor o el dato que recopilamos
- μ es la media
- σ es la desviación estándar
- e= constante en matemáticas que corresponde a 2.718
- π = constante en matemáticas que corresponde a 3.141

Históricamente podemos rastrear la curva normal hasta 1733 cuando Abraham de Moivre introdujo una aproximación a la distribución para grandes valores. Sin embargo es hasta 1809 cuando se atribuye su descubrimiento a Gauss porque la justificó rigurosamente en sus trabajos astronómicos para analizar la distribución de los errores.

En principio, la distribución se representa por la famosa campana de Gauss que representa todo los datos que recabamos en nuestra información. Su premisa es que cuando tenemos una cantidad suficientemente grande de datos, lo más probable es que tengamos una distribución de este tipo.



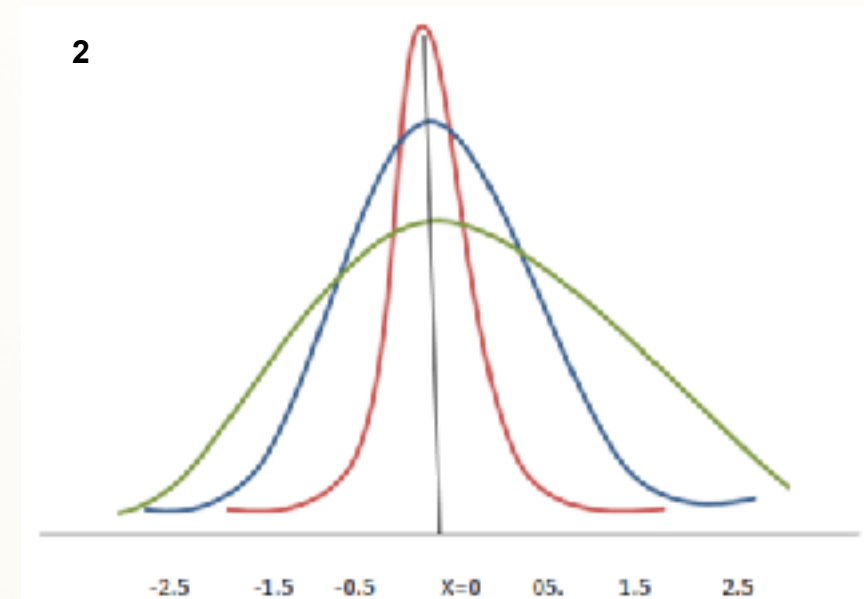
Distribución normal

En primer lugar, la media está en el centro de la campana, el punto más alto. Los puntos siguientes son desviaciones estándar, por lo que podemos deducir que la campana de Gauss nos dice lo que hemos estudiado en los capítulos anteriores. A medida que nos alejamos de la media (en desviaciones estándares) encontramos menos datos. Pensemos en la inteligencia medida por el CI (Coeficiente Intelectual), y veamos la gráfica. Al centro está el promedio que en CI es 100 por ser el promedio, la mayoría de gente está ubicada allí.

Si nos vamos a una desviación estándar hacia la derecha vemos que la altura de la curva disminuye, significa que hay menos gente en ese punto, si nos alejamos a 2 desviaciones estándar a la derecha, la curva se reduce mucho más, significa que hay menos personas aún cuyo CI se encuentre a 2 desviaciones estándar por arriba de la media. No se necesita saber exactamente el valor de la desviación estándar para poder determinar la proporción de personas que tendrán un CI determinado.

Así pues, iniciaremos por entender algunos de los conceptos importantes de la curva normal y las propiedades que tiene. Posteriormente veremos algunos cálculos que podemos realizar aprovechando las propiedades que se le atribuyen.

- Como ya hemos podido deducir, vamos a decir que la distribución normal está formada por dos parámetros, la media y la desviación estándar.
- Se representa por tener la forma de una campana simétrica y en donde también coinciden la moda y la mediana (teóricamente así es).
- Se dibuja en un plano cartesiano, y las colas de la curva nunca tocan el eje X, sino que se extienden hasta el infinito.
- Los puntos de inflexión de la curva, es decir en donde la curva cambia la tendencia de su inclinación, se encuentran a una distancia que equivale a una desviación estándar hacia ambos extremos.



Efecto de la desviación estándar en la distribución normal

imagen podemos observar el efecto de la desviación en la curva o campana de Gauss. En las tres gráficas tenemos la misma media, pero en la azul tendríamos por ejemplo una desviación estándar = 0.5, en la roja una desviación estándar 0.25 y en la verde una desviación estándar de 1.

- La curva además permite hacer interpretaciones probabilísticas. Estos temas los hemos discutido anteriormente pero vamos a volver a tocar el tema por su importancia: entre la media y una desviación estándar por ejemplo SIEMPRE se encontrará la misma proporción de datos,

•En la

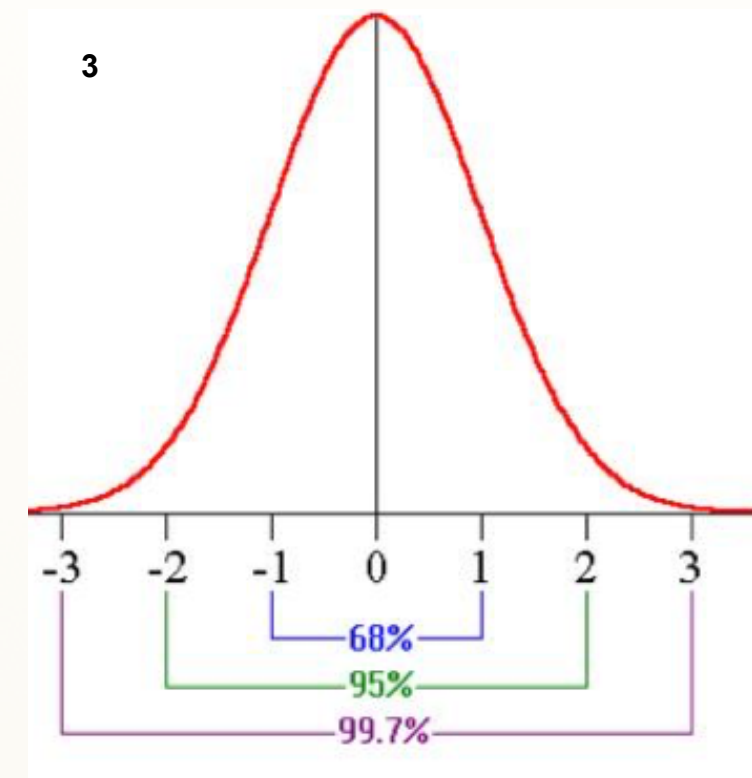
asumiendo que la cantidad de ellos es suficientemente grande para considerarse que la distribución de los datos se apega a una distribución normal. Por otro lado, si podemos decir que un evento o una distribución de datos es normal, podemos asignar probabilidades a que suceda un dato particular (probabilidad de ubicación dentro de la distribución).

Así pues, dentro de -1 desviación estándar y $+1$ desviación estándar se encuentra el 68% de los datos.



Entre -2 y $+2$ desviaciones estándar está el 95% de las probabilidades.

Entre -3 y $+3$ desviaciones estándar encontraremos prácticamente a todos los datos el 99.7% de la población.



Distribución normal

Vemos nuevamente esta gráfica que representa lo que acabamos de definir. Notemos la media $\mu=0$ y cada desviación estándar $\sigma=1$. Para entender el sentido práctico vamos a hablar de un ejemplo que me sucedió hace algunos años. Estaba en mi oficina un día y un amigo con quien solíamos correr 10km cada domingo me dijo que había una carrera de media maratón ese fin de semana.

Media maratón son 21kilómetros y el terreno en donde se correría era mucho más inclinado que donde solíamos correr. Yo no acepté la invitación de participar pues nunca había corrido más de 10km en cada entreno y no había tiempo suficiente para entrenar para 21k.

Mi amigo muy emocionado me dijo que correría y que la completaría en menos de dos horas (2:00) y estaba tan seguro en sí mismo que podíamos apostar una pizza a que lo lograría. Antes de aceptar la apuesta yo busqué la base de datos de la carrera de los años anteriores y encontré los tiempos de todos los corredores inscritos. Saqué la media y la desviación estándar para el tiempo que hicieron y obtuve estos datos, la media era 2:25 dos horas y veinticinco minutos; desviación estándar 0:32, treinta y dos minutos.

Usando el conocimiento de la distribución normal, pude hacer un cálculo simple para saber cuánto tiempo hace el 68% de la gente que corre: La media – y + una desviación estándar, es decir: entre 1:53 y 2:57.



Gente que corre.

Con esa información acepté la apuesta, porque era muy poco probable que una persona sin entrenar para una media maratón lograra hacer los tiempos promedio, por supuesto que disfruté una rica pizza gracias a mis conocimientos sobre la distribución normal y sus propiedades.

Ya que entendimos que la distribución normal puede interpretarse de una forma probabilística entonces podemos explorar las aplicaciones que tiene en la vida cotidiana, especialmente en investigadores sociales.

5

*Probabilidades*

Partamos de la idea de que sus principales parámetros son la media y la desviación estándar que pueden adquirir un número infinito de valores. Esto haría imposible calcular probabilidades específicas para cada distribución. En otras palabras, cada combinación posible de media y desviación estándar va a producir una distribución nueva.

Para ello, los grandes pensadores matemáticos idearon una forma de simplificación para nuestras vidas y la llamaron Distribución Normal Reducida o Distribución Normal Estandarizada!

Este hallazgo permite utilizar la distribución prácticamente para cualquier media y cualquier desviación estándar y por ende nos permite hacer cálculos probabilísticos.

En términos técnicos, se produce una traslación de la escala de los valores, una modificación que no afecta sus propiedades pero que permite estandarizar los valores dentro de la distribución normal.

Al resultado de la operación de la traslación le llamaremos una variable “Z” y para calcularla la fórmula que utilizaremos es la siguiente:

$$Z = \frac{X - \mu}{\sigma}$$

Donde

X es el valor o el dato que recopilamos

μ es la media

σ es la desviación estándar

En la distribución de los datos originales de nuestro estudio la media y desviación estándar pueden tener cualquier valor y a través de la transformación con la fórmula de estandarización lo que podemos lograr es una modificación cuantitativa a los datos de nuestra distribución original para que la media = 0 y que la desviación estándar = 1.





Comprobación del capítulo XV: La curva normal como probabilidades

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XV

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XVI

Usos de la curva normal



Usos de la curva normal

Estandarización de punteos: diremos que una distribución normal estándar es aquella en donde

$$\mu = 0 \text{ y } \sigma = 1$$

Vamos a hacer un ejercicio para mostrar la forma de transformar los datos y para hacerlo necesitamos partir de una distribución de frecuencias como la que vimos en la unidad 2. Vamos a suponer que queremos conocer la distribución de la cantidad de libros que han leído en su vida las personas, hacemos la pregunta a 88 personas y este es el resultado que tenemos. El primer paso será realizar la distribución de frecuencias.

1

6	11	12	13	10	11	13	14	16
7	11	12	13	10	11	13	14	17
7	11	12	13	10	11	13	14	17
8	11	12	13	10	12	13	15	18
8	11	12	13	10	12	13	15	14
8	11	12	14	10	12	13	15	12
9	11	12	14	10	12	13	15	11
9	11	12	14	10	12	13	15	9
4	11	12	14	10	12	13	16	
4	11	12	14	11	12	13	16	

Distribución de frecuencias

Recordemos que X es el dato que encontramos y f_i es la frecuencia absoluta, por lo tanto podemos saber que hay 1 persona que únicamente ha leído 6 libros en su vida, hay 5 que han leído 9, también hay 5 que han leído 15.

2

X	f_i
6	1
7	2
8	3
9	5
10	9
11	15
12	18
13	15
14	9
15	5
16	3
17	2
18	1

Dato y frecuencia

El segundo paso a realizar es obtener la media y la desviación estándar, podemos hacerlo usando de guía el capítulo 4 de medidas de tendencia central y de dispersión. **MUY IMPORTANTE**, debemos obtener media y desviación estándar de nuestra tabla de datos original, no de la distribución.



3

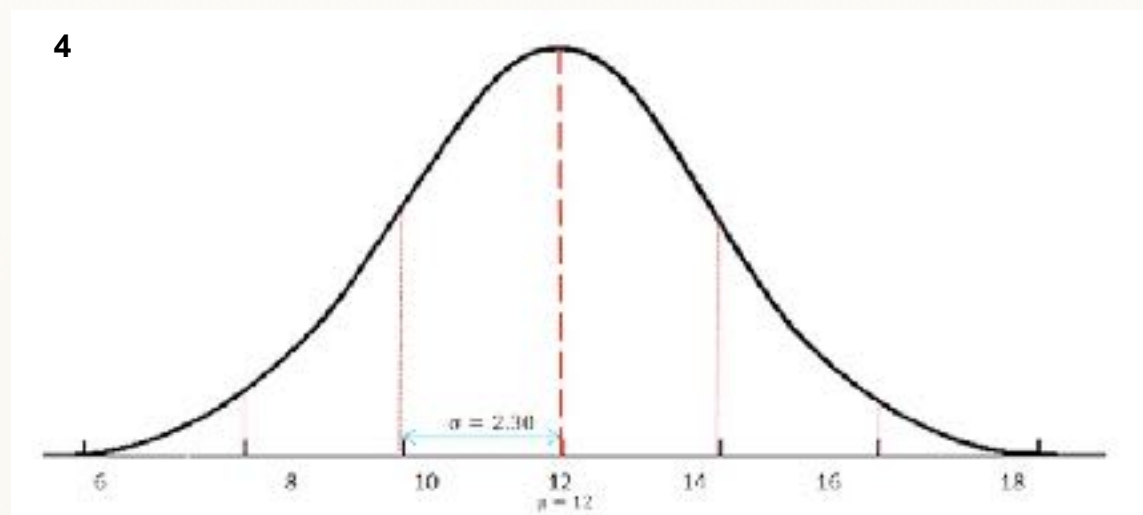
6	11	12	13	10	11	13	14	16
7	11	12	13	10	11	13	14	17
7	11	12	13	10	11	13	14	17
8	11	12	13	10	12	13	15	18
8	11	12	13	10	12	13	15	14
8	11	12	14	10	12	13	15	12
9	11	12	14	10	12	13	15	11
9	11	12	14	10	12	13	15	9
9	11	12	14	10	12	13	16	
9	11	12	14	11	12	13	16	
Media				Desviación Estándar				
O(B23:J32)								
6	11	12	13	10	11	13	14	16
7	11	12	13	10	11	13	14	17
7	11	12	13	10	11	13	14	17
8	11	12	13	10	12	13	15	18
8	11	12	13	10	12	13	15	14
8	11	12	14	10	12	13	15	12
9	11	12	14	10	12	13	15	11
9	11	12	14	10	12	13	15	9
9	11	12	14	10	12	13	16	
9	11	12	14	11	12	13	16	
Media				Desviación Estándar				
				=DESVESTA(B23:J32)				

Media y desviación estándar

Así obtenemos nuestros parámetros

$$\mu = 12 \text{ y } \sigma = 2.30$$

Con estos datos y para efecto de visualización, veremos cómo es la distribución de nuestro ejemplo, SIN estandarizar, es decir, la distribución de la encuesta que realizamos a las 88 personas.



Media y desviación estándar

Vemos la media en el centro y vemos cómo se ubican los datos que obtuvimos dentro de la distribución en cada desviación estándar. ¿qué podemos saber en este momento? Pues a grandes rasgos que el 68% de los entrevistados ha leído entre 10 y 14 libros y eso equivale

aproximadamente a 2 de cada 3 personas o 66%. También podemos decir que el 95% ha leído entre 8 y 16 libros, en otras palabras, muy pocas personas han leído menos de 8 y muy pocas han leído más de 16.

Pero la distribución normal tiene más bondades, ahora nos toca proceder a realizar la transformación de los datos hacia una distribución normal estandarizada, la obtención de cada puntaje Z nos permitirá conocer a cuántas desviaciones estándar equivale cada valor real, en nuestro ejemplo la cantidad de libros leída. Para ello usaremos nuevamente nuestra distribución de frecuencias como guía. (se modifica la posición de la tabla para hacer más visible los ejercicios).

5

X	6	7	8	9	10	11	12	13	14	15	16	17	18
f _i	1	2	3	5	9	15	18	15	9	5	3	2	1

Posición de la tabla

Usaremos la fórmula de estandarización.

6

$$Z = \frac{X - \mu}{\sigma}$$

fórmula de estandarización

El primer dato que vamos a revisar será el “6”. Así que vamos a la fórmula

$$Z = (6 - 12) / 2.30$$

$$Z = -2.6$$

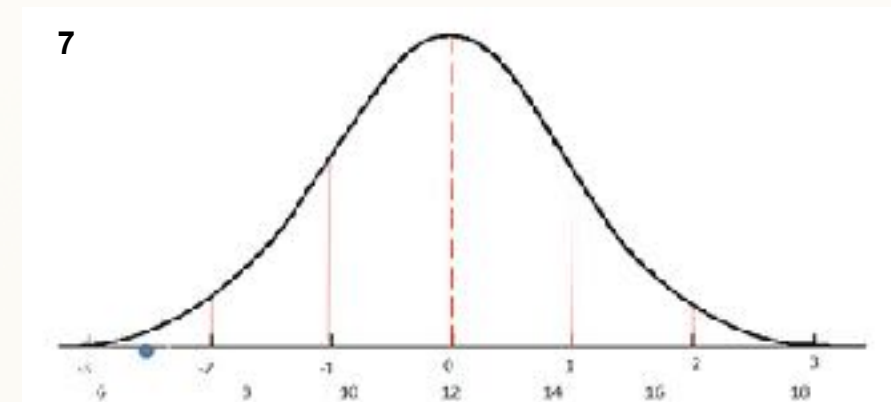
Esto significa que en una distribución normal estándar, el valor 6 equivale a -2.6... si lo vemos en la gráfica fácilmente podemos observar que la persona que dice haber leído 6 libros en su vida, se ubica a menos 2.6 desviaciones estándar de lo que leen las personas promedio (punto azul en la próxima imagen).



Nota:

Es importante darse cuenta, que la desviación estándar se ha convertido en una unidad de medida.

En esta imagen, colocamos la distribución normal estandarizada encima de la distribución normal de los datos para ver que podríamos tener cualquier valor de la media y cualquier desviación estándar y “estandarizar” esa distribución para poder trabajar probabilidades con ella.



Distribución normal estandarizada con datos del ejemplo.



Nota:

Es importante aclarar que la curva normal nunca toca el eje horizontal.

Obtención de áreas:

Para poder aprovechar más las bondades que ofrece la estandarización, tenemos que mencionar que al dibujar una campana, lo que queda dentro de la misma es un “área” un espacio debajo de la curva. Con fórmulas matemáticas se logra determinar cual es el área que hay debajo de la curva normal limitada por los valores Z de nuestro interés, y resulta que podemos relacionar el área bajo la curva con una probabilidad acumulada asociada al valor Z de nuestro interés.



Es decir, cuando la naturaleza de un evento obedezca a una distribución normal, podemos asignar una probabilidad que ese evento suceda, obteniendo el área bajo la curva normal estandarizada limitada por el o los valores Z de nuestro interés.

Los matemáticos han realizado los análisis y han generado una tabla denominada: Área debajo de la curva normal y la encontramos a disposición para poder utilizarla.

8

Cátedra: Probabilidad y Estadística
Facultad Regional Mendoza
UTN

Tabla D.5: ÁREAS BAJO LA CURVA NORMAL ESTÁNDAR



The diagrams illustrate the four types of areas under the standard normal curve:

- c.chica (z)**: The area to the right of a positive z-value.
- c.grande (z)**: The area to the left of a positive z-value.
- área central**: The area between two negative and positive z-values.
- área (0 a z)**: The area between the mean (0) and a positive z-value.

z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000	0,50	0,30854	0,69146	0,38292	0,19146
0,01	0,49601	0,50399	0,00798	0,00399	0,51	0,30503	0,69497	0,38995	0,19497
0,02	0,49202	0,50798	0,01596	0,00798	0,52	0,30153	0,69847	0,39694	0,19847
0,03	0,48803	0,51197	0,02393	0,01197	0,53	0,29806	0,70194	0,40389	0,20194
0,04	0,48405	0,51595	0,03191	0,01595	0,54	0,29460	0,70540	0,41080	0,20540
0,05	0,48006	0,51994	0,03988	0,01994	0,55	0,29116	0,70884	0,41768	0,20884
0,06	0,47608	0,52392	0,04784	0,02392	0,56	0,28774	0,71226	0,42452	0,21226
0,07	0,47210	0,52790	0,05581	0,02790	0,57	0,28434	0,71566	0,43132	0,21566
0,08	0,46812	0,53188	0,06376	0,03188	0,58	0,28096	0,71904	0,43809	0,21904
0,09	0,46414	0,53586	0,07171	0,03586	0,59	0,27760	0,72240	0,44481	0,22240
0,10	0,46017	0,53983	0,07966	0,03983	0,60	0,27425	0,72575	0,45149	0,22575
0,11	0,45620	0,54380	0,08759	0,04380	0,61	0,27093	0,72907	0,45814	0,22907
0,12	0,45224	0,54776	0,09552	0,04776	0,62	0,26763	0,73237	0,46474	0,23237
0,13	0,44828	0,55172	0,10343	0,05172	0,63	0,26435	0,73565	0,47131	0,23565
0,14	0,44433	0,55567	0,11134	0,05567	0,64	0,26109	0,73891	0,47783	0,23891
0,15	0,44038	0,55962	0,11924	0,05962	0,65	0,25785	0,74215	0,48431	0,24215
0,16	0,43644	0,56356	0,12712	0,06356	0,66	0,25463	0,74537	0,49075	0,24537
0,17	0,43251	0,56749	0,13499	0,06749	0,67	0,25143	0,74857	0,49714	0,24857
0,18	0,42858	0,57142	0,14285	0,07142	0,68	0,24825	0,75175	0,50350	0,25175
0,19	0,42465	0,57535	0,15069	0,07535	0,69	0,24510	0,75490	0,50981	0,25490

Áreas bajo la curva normal estándar.

La tabla puede parecer confusa en un inicio, pero realmente tiene una lógica bastante simple que vamos a desmenuzar a continuación para poder usarla. Lo haremos con ejemplos.

Lo primero que podemos ver es que la primer columna nos brinda los posibles valores de Z hacia abajo y luego cada fila tiene varios números, esos números son el área debajo de la curva o la probabilidad acumulada en ese punto Z. Así que solamente nos toca saber qué buscar y cuándo.



La primer columna de áreas dice: C.Chica, otras tablas le llaman de “Z a infinito” y está circulada en rojo. Nos muestra la curva normal y está marcando el área después de Z. Esto nos indica que nos sirve para responder preguntas como ¿cuántas personas han leído 14 libros o más?

En nuestro ejemplo hacemos el cálculo de Z para el valor 14. Aplicamos la fórmula

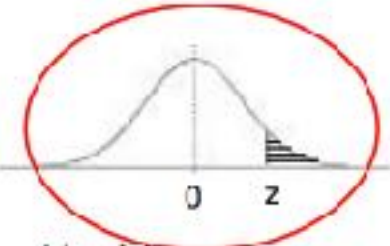
$$Z = (14-12) / 2.30$$
$$Z = 0.87$$

Ahora nos ubicaremos en la tabla para buscar este valor en la columna C. Chica

9

Cátedra: Probabilidad y Estadística
Facultad Regional Mendoza
UTN

Tabla D.5: ÁREAS BAJO LA C



c.chica (z)

c.grande (z)

z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000
0,01	0,49601	0,50399	0,00798	0,00399
0,02	0,49202	0,50798	0,01596	0,00798
0,03	0,48803	0,51197	0,02393	0,01197
0,04	0,48405	0,51595	0,03191	0,01595

Curva normal

10

z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000	0,50	0,30854	0,69146	0,38292	0,19146
0,01	0,49601	0,50399	0,00798	0,00399	0,51	0,30503	0,69497	0,38995	0,19497
0,02	0,49202	0,50798	0,01596	0,00798	0,52	0,30153	0,69847	0,39694	0,19847
0,03	0,48803	0,51197	0,02393	0,01197	0,53	0,29806	0,70194	0,40389	0,20194
0,04	0,48405	0,51595	0,03191	0,01595	0,54	0,29460	0,70540	0,41080	0,20540
0,05	0,48006	0,51994	0,03988	0,01994	0,55	0,29116	0,70884	0,41768	0,20884
0,06	0,47608	0,52392	0,04784	0,02392	0,56	0,28774	0,71226	0,42452	0,21226
0,07	0,47210	0,52790	0,05581	0,02790	0,57	0,28434	0,71566	0,43132	0,21566
0,08	0,46812	0,53188	0,06376	0,03188	0,58	0,28096	0,71904	0,43809	0,21904
0,09	0,46414	0,53586	0,07171	0,03586	0,59	0,27760	0,72240	0,44481	0,22240
0,10	0,46017	0,53983	0,07966	0,03983	0,60	0,27425	0,72575	0,45149	0,22575
0,11	0,45620	0,54380	0,08759	0,04380	0,61	0,27093	0,72907	0,45814	0,22907
0,12	0,45224	0,54776	0,09552	0,04776	0,62	0,26763	0,73237	0,46474	0,23237
0,13	0,44828	0,55172	0,10343	0,05172	0,63	0,26435	0,73565	0,47131	0,23565
0,14	0,44433	0,55567	0,11134	0,05567	0,64	0,26109	0,73891	0,47783	0,23891
0,15	0,44038	0,55962	0,11924	0,05962	0,65	0,25785	0,74215	0,48431	0,24215
0,16	0,43644	0,56356	0,12712	0,06356	0,66	0,25463	0,74537	0,49075	0,24537
0,17	0,43251	0,56749	0,13499	0,06749	0,67	0,25143	0,74857	0,49714	0,24857
0,18	0,42858	0,57142	0,14285	0,07142	0,68	0,24825	0,75175	0,50350	0,25175
0,19	0,42465	0,57535	0,15069	0,07535	0,69	0,24510	0,75490	0,50981	0,25490
0,20	0,42074	0,57926	0,15852	0,07926	0,70	0,24196	0,75804	0,51607	0,25804
0,21	0,41683	0,58317	0,16633	0,08317	0,71	0,23885	0,76115	0,52230	0,26115
0,22	0,41294	0,58706	0,17413	0,08706	0,72	0,23576	0,76424	0,52848	0,26424
0,23	0,40905	0,59095	0,18191	0,09095	0,73	0,23270	0,76730	0,53461	0,26730
0,24	0,40517	0,59483	0,18967	0,09483	0,74	0,22965	0,77035	0,54070	0,27035
0,25	0,40129	0,59871	0,19741	0,09871	0,75	0,22663	0,77337	0,54675	0,27337
0,26	0,39743	0,60257	0,20514	0,10257	0,76	0,22363	0,77637	0,55275	0,27637
0,27	0,39358	0,60642	0,21284	0,10642	0,77	0,22065	0,77935	0,55870	0,27935
0,28	0,38974	0,61026	0,22052	0,11026	0,78	0,21770	0,78230	0,56461	0,28230
0,29	0,38591	0,61409	0,22818	0,11409	0,79	0,21476	0,78524	0,57047	0,28524
0,30	0,38209	0,61791	0,23582	0,11791	0,80	0,21186	0,78814	0,57629	0,28814
0,31	0,37828	0,62172	0,24344	0,12172	0,81	0,20897	0,79103	0,58206	0,29103
0,32	0,37448	0,62552	0,25103	0,12552	0,82	0,20611	0,79389	0,58778	0,29389
0,33	0,37070	0,62930	0,25860	0,12930	0,83	0,20327	0,79673	0,59346	0,29673
0,34	0,36693	0,63307	0,26614	0,13307	0,84	0,20045	0,79955	0,59909	0,29955
0,35	0,36317	0,63683	0,27366	0,13683	0,85	0,19766	0,80234	0,60467	0,30234
0,36	0,35942	0,64058	0,28115	0,14058	0,86	0,19489	0,80511	0,61021	0,30511
0,37	0,35569	0,64431	0,28862	0,14431	0,87	0,19215	0,80785	0,61570	0,30785
0,38	0,35197	0,64803	0,29605	0,14803	0,88	0,18943	0,81057	0,62114	0,31057
0,39	0,34827	0,65173	0,30346	0,15173	0,89	0,18673	0,81327	0,62653	0,31327

Valores

Y encontramos que el área debajo de la curva es .1921, este valor puede representarse en porcentajes multiplicándolo por 100 así que tenemos que el 19.21% de los sujetos encuestados ha leído 14 o más libros.

Otra forma de preguntar este mismo ejemplo sería: ¿cuál es la probabilidad de encontrar a alguien que haya leído 14 o más libros?

Ahora bien, un investigador podría estar interesado en preguntar, ¿cuánta gente ha leído menos de 14 libros? O cuál es la probabilidad de encontrar a alguien que haya leído un máximo de 14 libros en su vida?

Para esto usaremos la segunda columna de la tabla C. grande que en otras tablas podría encontrarse como “menos infinito a Z”.

11

UTN

Tabla D.5: ÁREAS BAJO LA



z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000
0,01	0,49601	0,50399	0,00798	0,00399
0,02	0,49202	0,50798	0,01596	0,00798
0,03	0,48803	0,51197	0,02393	0,01197
0,04	0,48405	0,51595	0,03191	0,01595

Menos infinito a Z

El procedimiento nuevamente inicia por calcular Z, en este caso ya sabemos que el valor Z para 14 es 0.87 así que nos ubicamos en la tabla nuevamente pero ahora en la segunda columna.

$$Z = (14-12) / 2.30$$
$$Z = 0.87$$

12

z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000	0,50	0,30854	0,69146	0,38292	0,19146
0,01	0,49601	0,50399	0,00798	0,00399	0,51	0,30503	0,69497	0,38995	0,19497
0,02	0,49202	0,50798	0,01596	0,00798	0,52	0,30153	0,69847	0,39694	0,19847
0,03	0,48803	0,51197	0,02393	0,01197	0,53	0,29806	0,70194	0,40389	0,20194
0,04	0,48405	0,51595	0,03191	0,01595	0,54	0,29460	0,70540	0,41080	0,20540
0,05	0,48006	0,51994	0,03988	0,01994	0,55	0,29116	0,70884	0,41768	0,20884
0,06	0,47608	0,52392	0,04784	0,02392	0,56	0,28774	0,71226	0,42452	0,21226
0,07	0,47210	0,52790	0,05581	0,02790	0,57	0,28434	0,71566	0,43132	0,21566
0,08	0,46812	0,53188	0,06376	0,03188	0,58	0,28096	0,71904	0,43809	0,21904
0,09	0,46414	0,53586	0,07171	0,03586	0,59	0,27760	0,72240	0,44481	0,22240
0,10	0,46017	0,53983	0,07966	0,03983	0,60	0,27425	0,72575	0,45149	0,22575
0,11	0,45620	0,54380	0,08759	0,04380	0,61	0,27093	0,72907	0,45814	0,22907
0,12	0,45224	0,54776	0,09552	0,04776	0,62	0,26763	0,73237	0,46474	0,23237
0,13	0,44828	0,55172	0,10343	0,05172	0,63	0,26435	0,73565	0,47131	0,23565
0,14	0,44433	0,55567	0,11134	0,05567	0,64	0,26109	0,73891	0,47783	0,23891
0,15	0,44038	0,55962	0,11924	0,05962	0,65	0,25785	0,74215	0,48431	0,24215
0,16	0,43644	0,56356	0,12712	0,06356	0,66	0,25463	0,74537	0,49075	0,24537
0,17	0,43251	0,56749	0,13499	0,06749	0,67	0,25143	0,74857	0,49714	0,24857
0,18	0,42858	0,57142	0,14285	0,07142	0,68	0,24825	0,75175	0,50350	0,25175
0,19	0,42465	0,57535	0,15069	0,07535	0,69	0,24510	0,75490	0,50981	0,25490
0,20	0,42074	0,57926	0,15852	0,07926	0,70	0,24196	0,75804	0,51607	0,25804
0,21	0,41683	0,58317	0,16633	0,08317	0,71	0,23885	0,76115	0,52230	0,26115
0,22	0,41294	0,58706	0,17413	0,08706	0,72	0,23576	0,76424	0,52848	0,26424
0,23	0,40905	0,59095	0,18191	0,09095	0,73	0,23270	0,76730	0,53461	0,26730
0,24	0,40517	0,59483	0,18967	0,09483	0,74	0,22965	0,77035	0,54070	0,27035
0,25	0,40129	0,59871	0,19741	0,09871	0,75	0,22663	0,77337	0,54675	0,27337
0,26	0,39743	0,60257	0,20514	0,10257	0,76	0,22363	0,77637	0,55275	0,27637
0,27	0,39358	0,60642	0,21284	0,10642	0,77	0,22065	0,77935	0,55870	0,27935
0,28	0,38974	0,61026	0,22052	0,11026	0,78	0,21770	0,78230	0,56461	0,28230
0,29	0,38591	0,61409	0,22818	0,11409	0,79	0,21476	0,78524	0,57047	0,28524
0,30	0,38209	0,61791	0,23582	0,11791	0,80	0,21186	0,78814	0,57629	0,28814
0,31	0,37828	0,62172	0,24344	0,12172	0,81	0,20897	0,79103	0,58206	0,29103
0,32	0,37448	0,62552	0,25103	0,12552	0,82	0,20611	0,79389	0,58778	0,29389
0,33	0,37070	0,62930	0,25860	0,12930	0,83	0,20327	0,79673	0,59346	0,29673
0,34	0,36693	0,63307	0,26614	0,13307	0,84	0,20045	0,79955	0,59909	0,29955
0,35	0,36317	0,63683	0,27366	0,13683	0,85	0,19766	0,80234	0,60467	0,30234
0,36	0,35942	0,64058	0,28115	0,14058	0,86	0,19489	0,80511	0,61021	0,30511
0,37	0,35569	0,64431	0,28862	0,14431	0,87	0,19215	0,80785	0,61570	0,30785
0,38	0,35197	0,64803	0,29605	0,14803	0,88	0,18943	0,81057	0,62114	0,31057
0,39	0,34827	0,65173	0,30346	0,15173	0,89	0,18673	0,81327	0,62653	0,31327

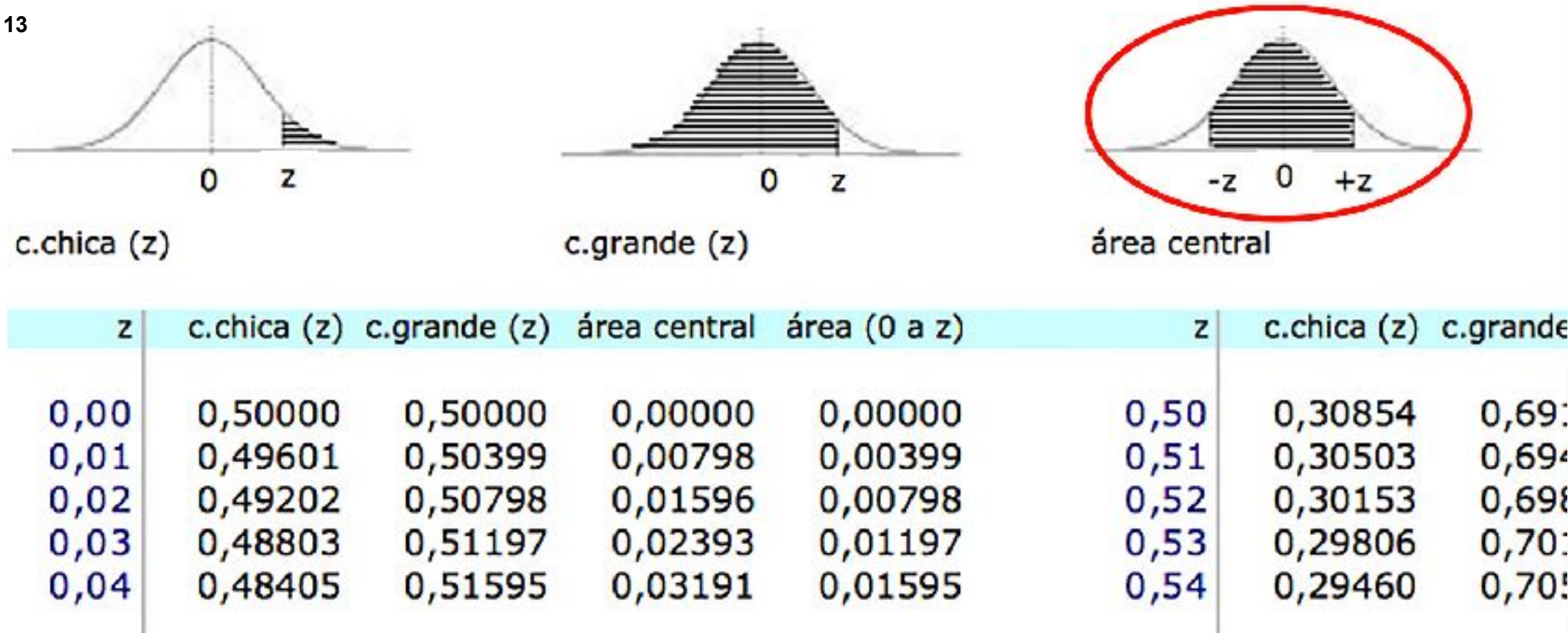
Valores

Así pues, encontramos que el área debajo de la curva o la probabilidad acumulada o la población acumulada hasta este punteo Z es 0.8078 es decir el 80.78% . Asi, la probabilidad de encontrarse a alguien que haya leído menos de 14 libros en su vida es muy alta (80.78%). Si alguien le reta a una apuesta con este grupo de gente y le dice: a que no encuentras a una persona que haya leído menos de 14 libros en su vida usted puede

escoger a cualquiera al azar teniendo el 80.78% de probabilidad de ganar.

La tercera opción es el área central. Esta es la que más hemos discutido y se usa para cuando necesitamos saber el área entre el valor negativo de un punteo Z y el valor positivo de ese mismo valor Z.

13



Área entre valor positivo y negativo.

Cuando menciono que lo hemos discutido es porque vamos a ver el ejemplo cuando $Z = 1$. Así que queremos saber cuál es la probabilidad de encontrar a una persona que esté entre -1 desviación estándar y +1 desviación estándar.

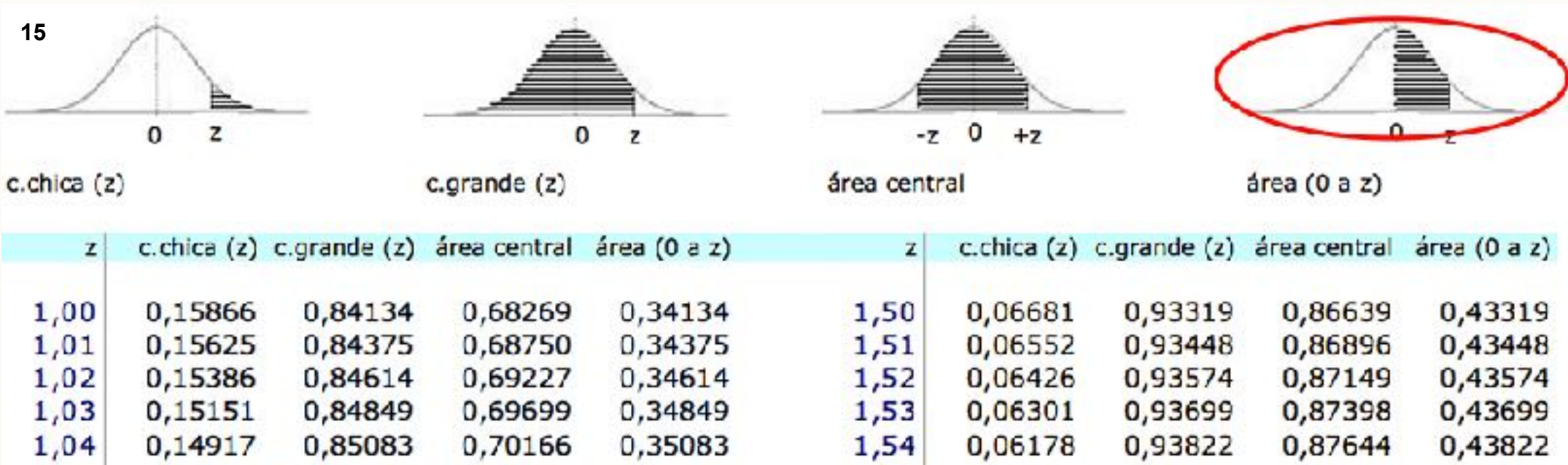
14

z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
1,00	0,15866	0,84134	0,68269	0,34134	1,50	0,06681	0,93319	0,86639	0,43319
1,01	0,15625	0,84375	0,68750	0,34375	1,51	0,06552	0,93448	0,86896	0,43448
1,02	0,15386	0,84614	0,69227	0,34614	1,52	0,06426	0,93574	0,87149	0,43574
1,03	0,15151	0,84849	0,69699	0,34849	1,53	0,06301	0,93699	0,87398	0,43699
1,04	0,14917	0,85083	0,70166	0,35083	1,54	0,06178	0,93822	0,87644	0,43822
1,05	0,14686	0,85314	0,70628	0,35314	1,55	0,06057	0,93943	0,87886	0,43943
1,06	0,14457	0,85543	0,71086	0,35543	1,56	0,05938	0,94062	0,88124	0,44062
1,07	0,14231	0,85769	0,71538	0,35769	1,57	0,05821	0,94179	0,88358	0,44179
1,08	0,14007	0,85993	0,71986	0,35993	1,58	0,05705	0,94295	0,88589	0,44295
1,09	0,13786	0,86214	0,72429	0,36214	1,59	0,05592	0,94408	0,88817	0,44408
1,10	0,13567	0,86433	0,72867	0,36433	1,60	0,05480	0,94520	0,89040	0,44520
1,11	0,13350	0,86650	0,73300	0,36650	1,61	0,05370	0,94630	0,89260	0,44630
1,12	0,13136	0,86864	0,73729	0,36864	1,62	0,05262	0,94738	0,89477	0,44738
1,13	0,12924	0,87076	0,74152	0,37076	1,63	0,05155	0,94845	0,89690	0,44845
1,14	0,12714	0,87286	0,74571	0,37286	1,64	0,05050	0,94950	0,89899	0,44950
1,15	0,12507	0,87493	0,74986	0,37493	1,65	0,04947	0,95053	0,90106	0,45053
1,16	0,12302	0,87698	0,75395	0,37698	1,66	0,04846	0,95154	0,90309	0,45154
1,17	0,12100	0,87900	0,75800	0,37900	1,67	0,04746	0,95254	0,90508	0,45254
1,18	0,11900	0,88100	0,76200	0,38100	1,68	0,04648	0,95352	0,90704	0,45352
1,19	0,11702	0,88298	0,76595	0,38298	1,69	0,04551	0,95449	0,90897	0,45449
1,20	0,11507	0,88493	0,76986	0,38493	1,70	0,04457	0,95543	0,91087	0,45543
1,21	0,11314	0,88686	0,77372	0,38686	1,71	0,04363	0,95637	0,91273	0,45637
1,22	0,11123	0,88877	0,77754	0,38877	1,72	0,04272	0,95728	0,91457	0,45728
1,23	0,10935	0,89065	0,78130	0,39065	1,73	0,04182	0,95818	0,91637	0,45818
1,24	0,10749	0,89251	0,78502	0,39251	1,74	0,04093	0,95907	0,91814	0,45907

Probabilidad entre una desviación estándar y otra.

La respuesta es el 68.26% de la población está entre -1 y + 1 desviaciones estándar.

La última opción que nos brinda la tabla se llama de “0 a Z” o de “media a Z” y es para calcular el área o probabilidad entre la media y un puntaje Z.



Media a Z

El tipo de pregunta que podríamos responder aquí es así: ¿Cuál es la probabilidad de encontrar a una persona que haya leído entre 12 y 14 libros?

El primer paso es calcular la Z para el dato que no es la media, ya que la media sabemos que es $Z=0$.

$$Z = (14-12) / 2.30$$

$$Z = 0.87$$

Con el valor Z calculado lo ubicaremos en la tabla...

16 z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000	0,50	0,30854	0,69146	0,38292	0,19146
0,01	0,49601	0,50399	0,00798	0,00399	0,51	0,30503	0,69497	0,38995	0,19497
0,02	0,49202	0,50798	0,01596	0,00798	0,52	0,30153	0,69847	0,39694	0,19847
0,03	0,48803	0,51197	0,02393	0,01197	0,53	0,29806	0,70194	0,40389	0,20194
0,04	0,48405	0,51595	0,03191	0,01595	0,54	0,29460	0,70540	0,41080	0,20540
0,05	0,48006	0,51994	0,03988	0,01994	0,55	0,29116	0,70884	0,41768	0,20884
0,06	0,47608	0,52392	0,04784	0,02392	0,56	0,28774	0,71226	0,42452	0,21226
0,07	0,47210	0,52790	0,05581	0,02790	0,57	0,28434	0,71566	0,43132	0,21566
0,08	0,46812	0,53188	0,06376	0,03188	0,58	0,28096	0,71904	0,43809	0,21904
0,09	0,46414	0,53586	0,07171	0,03586	0,59	0,27760	0,72240	0,44481	0,22240
0,10	0,46017	0,53983	0,07966	0,03983	0,60	0,27425	0,72575	0,45149	0,22575
0,11	0,45620	0,54380	0,08759	0,04380	0,61	0,27093	0,72907	0,45814	0,22907
0,12	0,45224	0,54776	0,09552	0,04776	0,62	0,26763	0,73237	0,46474	0,23237
0,13	0,44828	0,55172	0,10343	0,05172	0,63	0,26435	0,73565	0,47131	0,23565
0,14	0,44433	0,55567	0,11134	0,05567	0,64	0,26109	0,73891	0,47783	0,23891
0,15	0,44038	0,55962	0,11924	0,05962	0,65	0,25785	0,74215	0,48431	0,24215
0,16	0,43644	0,56356	0,12712	0,06356	0,66	0,25463	0,74537	0,49075	0,24537
0,17	0,43251	0,56749	0,13499	0,06749	0,67	0,25143	0,74857	0,49714	0,24857
0,18	0,42858	0,57142	0,14285	0,07142	0,68	0,24825	0,75175	0,50350	0,25175
0,19	0,42465	0,57535	0,15069	0,07535	0,69	0,24510	0,75490	0,50981	0,25490
0,20	0,42074	0,57926	0,15852	0,07926	0,70	0,24196	0,75804	0,51607	0,25804
0,21	0,41683	0,58317	0,16633	0,08317	0,71	0,23885	0,76115	0,52230	0,26115
0,22	0,41294	0,58706	0,17413	0,08706	0,72	0,23576	0,76424	0,52848	0,26424
0,23	0,40905	0,59095	0,18191	0,09095	0,73	0,23270	0,76730	0,53461	0,26730
0,24	0,40517	0,59483	0,18967	0,09483	0,74	0,22965	0,77035	0,54070	0,27035
0,25	0,40129	0,59871	0,19741	0,09871	0,75	0,22663	0,77337	0,54675	0,27337
0,26	0,39743	0,60257	0,20514	0,10257	0,76	0,22363	0,77637	0,55275	0,27637
0,27	0,39358	0,60642	0,21284	0,10642	0,77	0,22065	0,77935	0,55870	0,27935
0,28	0,38974	0,61026	0,22052	0,11026	0,78	0,21770	0,78230	0,56461	0,28230
0,29	0,38591	0,61409	0,22818	0,11409	0,79	0,21476	0,78524	0,57047	0,28524
0,30	0,38209	0,61791	0,23582	0,11791	0,80	0,21186	0,78814	0,57629	0,28814
0,31	0,37828	0,62172	0,24344	0,12172	0,81	0,20897	0,79103	0,58206	0,29103
0,32	0,37446	0,62552	0,25103	0,12552	0,82	0,20611	0,79389	0,58778	0,29389
0,33	0,37070	0,62930	0,25860	0,12930	0,83	0,20327	0,79673	0,59346	0,29673
0,34	0,36693	0,63307	0,26614	0,13307	0,84	0,20045	0,79955	0,59909	0,29955
0,35	0,36317	0,63683	0,27366	0,13683	0,85	0,19766	0,80234	0,60467	0,30234
0,36	0,35942	0,64058	0,28115	0,14058	0,86	0,19489	0,80511	0,61021	0,30511
0,37	0,35569	0,64431	0,28862	0,14431	0,87	0,19215	0,80785	0,61570	0,30785
0,38	0,35197	0,64803	0,29605	0,14803	0,88	0,18943	0,81057	0,62114	0,31057
0,39	0,34827	0,65173	0,30346	0,15173	0,89	0,18673	0,81327	0,62653	0,31327

Calcular la Z

Al ubicarnos en la columna, vemos que el área debajo de la curva entre la media y $Z = 0.87$ es de 0.3078, es decir que el 30.78% de la población encuestada se encuentra allí y esa es la probabilidad de encontrarse a alguien que haya leído entre 12 y 14 libros en su vida.

Obtención de rangos percentiles desde la curva normal:

Cuando hablamos de percentiles en estadística muchas personas suelen asustarse, aunque ya lo discutimos en capítulos anteriores vamos a volver a decir que los percentiles son 99 valores que parten los datos en 100 partes iguales.

Al trasladar este conocimiento a la distribución normal, recordaremos que la mediana equivale al percentil 50 = Cuartil 2 = decil 5 y que en la curva normal mediana, moda y media son el mismo valor. Así que en la media de la distribución está el percentil 50.



Ahora bien, cuando tengamos una distribución normal no estandarizada y necesitemos encontrar los percentiles, el proceso que se debe realizar es primero tipificar para encontrar el valor de probabilidad y luego hacer el proceso inverso a la estandarización. Lo que sucede es que la naturaleza del percentil inmediatamente hace referencia al área debajo de la curva que queda detrás de dicho valor. Un percentil 75 significa que desde ese punto hay 75% de datos debajo y por lo tanto para poder ubicarlo necesitamos calcularlo.

Por razones prácticas dejaremos las fórmulas matemáticas complejas y utilizaremos la fórmula simplificada para resolver lo que acabamos de describir.

$$P_{\alpha} N(\mu, \sigma) = \mu + P_{\alpha} Z_{(0,1)} \sigma$$

Donde:

$P \propto N(\mu, \sigma)$ se refiere a el percentil de una distribución normal No estandarizada, μ es la media

$P \propto Z(0,1)$ se refiere al percentil de la distribución normal SI estandarizada $\sigma =$ desviación estándar

Veamos entonces el sentido práctico del tema con el ejemplo que estamos trabajando. Supongamos que queremos saber el valor del percentil 75 de nuestros datos.

La media sabemos que es 12, la desviación estándar 2.3, lo que necesitamos hacer es ubicar el percentil en una distribución normal estandarizada y para ello recurriremos a la tabla pero la usaremos de forma inversa, porque el percentil representa el área y lo que buscaremos será la Z.

17



z	c.chica (z)	c.grande (z)	área central	área (0 a z)	z	c.chica (z)	c.grande (z)	área central	área (0 a z)
0,00	0,50000	0,50000	0,00000	0,00000	0,50	0,30854	0,69146	0,38292	0,19146
0,01	0,49601	0,50399	0,00798	0,00399	0,51	0,30503	0,69497	0,38995	0,19497
0,02	0,49202	0,50798	0,01596	0,00798	0,52	0,30153	0,69847	0,39694	0,19847
0,03	0,48803	0,51197	0,02393	0,01197	0,53	0,29806	0,70194	0,40389	0,20194
0,04	0,48405	0,51595	0,03191	0,01595	0,54	0,29460	0,70540	0,41080	0,20540
0,05	0,48006	0,51994	0,03988	0,01994	0,55	0,29116	0,70884	0,41768	0,20884
0,06	0,47608	0,52392	0,04784	0,02392	0,56	0,28774	0,71226	0,42452	0,21226
0,07	0,47210	0,52790	0,05581	0,02790	0,57	0,28434	0,71566	0,43132	0,21566
0,08	0,46812	0,53188	0,06376	0,03188	0,58	0,28096	0,71904	0,43809	0,21904
0,09	0,46414	0,53586	0,07171	0,03586	0,59	0,27760	0,72240	0,44481	0,22240
0,10	0,46017	0,53983	0,07966	0,03983	0,60	0,27425	0,72575	0,45149	0,22575
0,11	0,45620	0,54380	0,08759	0,04380	0,61	0,27093	0,72907	0,45814	0,22907
0,12	0,45224	0,54776	0,09552	0,04776	0,62	0,26763	0,73237	0,46474	0,23237
0,13	0,44828	0,55172	0,10343	0,05172	0,63	0,26435	0,73565	0,47131	0,23565
0,14	0,44433	0,55567	0,11134	0,05567	0,64	0,26109	0,73891	0,47783	0,23891
0,15	0,44038	0,55962	0,11924	0,05962	0,65	0,25785	0,74215	0,48431	0,24215
0,16	0,43644	0,56356	0,12712	0,06356	0,66	0,25463	0,74537	0,49075	0,24537
0,17	0,43251	0,56749	0,13499	0,06749	0,67	0,25143	0,74857	0,49714	0,24857
0,18	0,42858	0,57142	0,14285	0,07142	0,68	0,24825	0,75175	0,50350	0,25175
0,19	0,42465	0,57535	0,15069	0,07535	0,69	0,24510	0,75490	0,50981	0,25490

Escogimos la columna de C-grande porque estamos buscando el área debajo de la curva que contenga el 75% de los datos, y así ubicar la Z correspondiente. En este caso la Z que corresponde es 0.68. Tenemos entonces el valor para completar la fórmula.

$$P_{\alpha} N(\mu, \sigma) = \mu + P_{\alpha} Z(0,1) \sigma$$

$$P_{75} N(12, 2.3) = 12 + 0.68 (2.3) = 13.56$$

Es decir, que las personas que leen 13 libros en su vida, están en el percentil 75.



Comprobación del capítulo XVI: Usos de la curva normal

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XVI

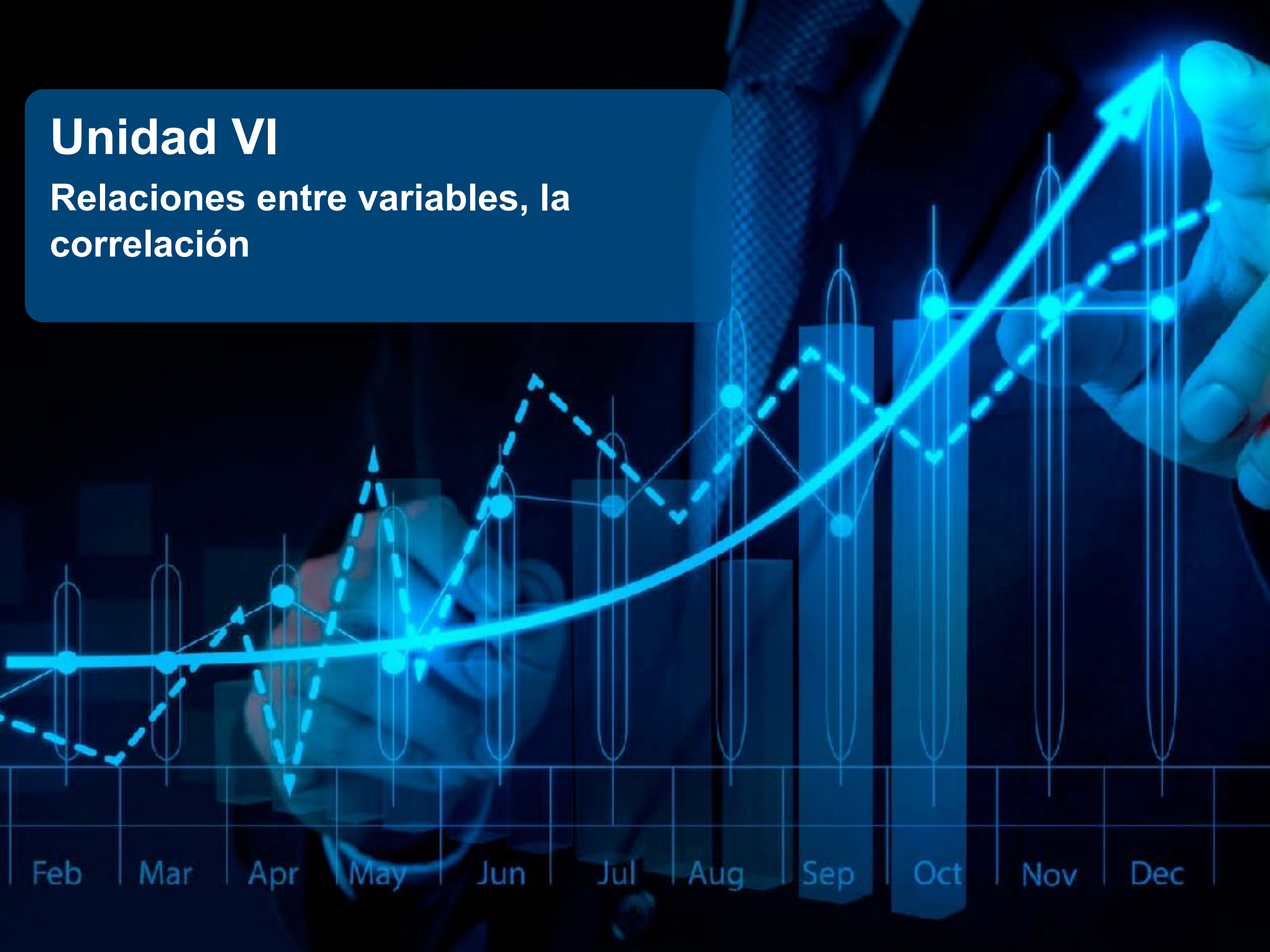
Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Unidad VI

Relaciones entre variables, la correlación



Se dedica al estudio del correcto uso de la correlación entre dos o más variables.

Es importante que discutamos en esta unidad sobre la relación entre las variables. En algunas ocasiones es común tomar decisiones porque el “sentido común” indica cierta lógica y cuando estudiamos ese sentido común, muchas veces nos damos cuenta que estábamos equivocados.



Sentido común

En mis clases de la universidad suelo preguntar el primer día del semestre :¿ qué variable es la que más afecta al rendimiento académico? Todos tenemos una opinión basada en nuestra experiencia y cuando cada uno responde, usa dicha experiencia para emitir su juicio. Sin embargo, luego de revisar todas las respuestas les digo: Qué evidencia científica tienen para decirme que esas variables se relacionan? Y son muy pocos los estudiantes que pueden citar estudios que mencionen dicha relación, así que se vuelve la primera tarea que les asigno y cada quién debe escribir un ensayo dónde planteen sus creencias y luego muestren evidencia para justificar o refutar lo que creen.

Esta unidad se enfoca en las relaciones de variables y cómo interpretarlas. El primer tema que debemos discutir es el de la **CORRELACIÓN**.

Por correlación entenderemos a una medida de asociación entre variables cuantitativas que determinan la fuerza de asociación entre ellas. En otras palabras, es la fuerza de relación entre dos variables.



Muy importante es tomar en cuenta, que el hecho de encontrar dos variables correlacionadas, no significa que una explique a la otra o que una prediga la otra y esta distinción es de vital importancia en estadística para evitar tomar decisiones basadas en correlaciones.

El término “Correlación Espuria” hace referencia a la asociación de variables debido a situaciones completamente independientes. Veremos dos ejemplos de este tipo de correlación.

- Resulta que existe una laguna donde las cigüeñas suelen descansar en una época del año. Frente a la laguna está un hospital de maternidad. Un estudiante curioso decidió contar la cantidad de cigüeñas en la laguna durante 30 días seguidos y también registró el número de nacimientos en ese hospital durante el mismo período de tiempo.

Luego buscó la correlación de los datos y encontró que las variables están fuertemente asociadas. Sin conocimiento de estadística alguien podría atreverse a decir que entre más cigüeñas llegan más niños nacen. Pero nosotros sabemos que es una correlación espuria.



Correlación espuria, ejemplo cigüeña.

- En Guatemala en una evaluación nacional, además de pruebas de matemática y lectura, preguntaban a los niños qué electrodomésticos hay en su hogar. Una analista decidió contar cuántos electrodomésticos había en la casa de cada estudiante y luego buscó la correlación con el rendimiento académico. Encontró que entre más electrodomésticos mejor era el rendimiento en matemáticas.

Nuevamente en este ejemplo, sin conocer estadística alguien podría pensar... dejemos de invertir en libros, computadoras, capacitación, mejor regalemos tostadores para que lleven a su casa. Pero no, nosotros sabemos que es una relación espuria y que atrás de la cantidad de electrodomésticos existe otra variable llamada nivel socioeconómico.

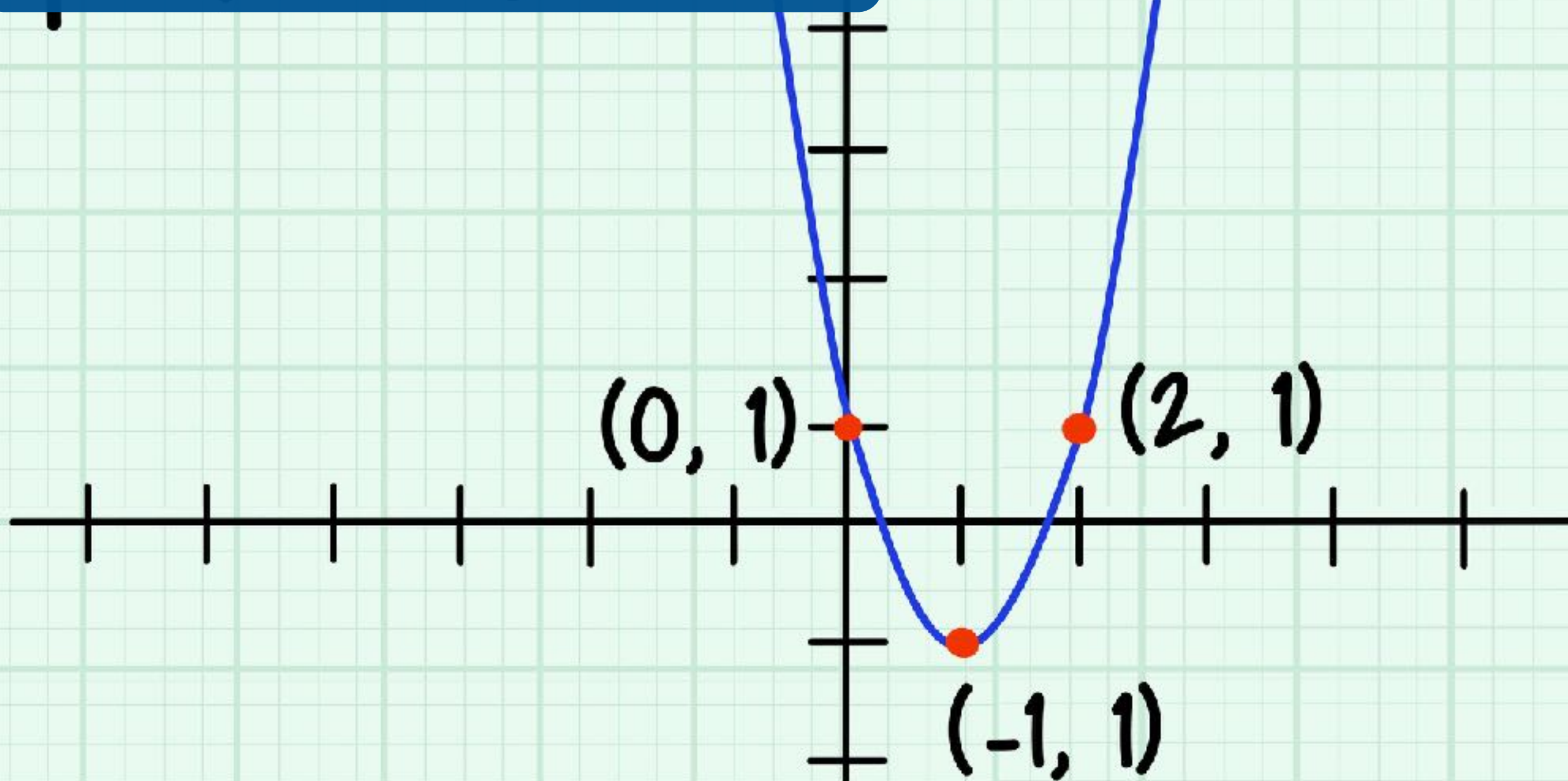


Correlación espuria, ejemplo electrodomésticos.

Capítulo XVII

El diagrama de dispersión

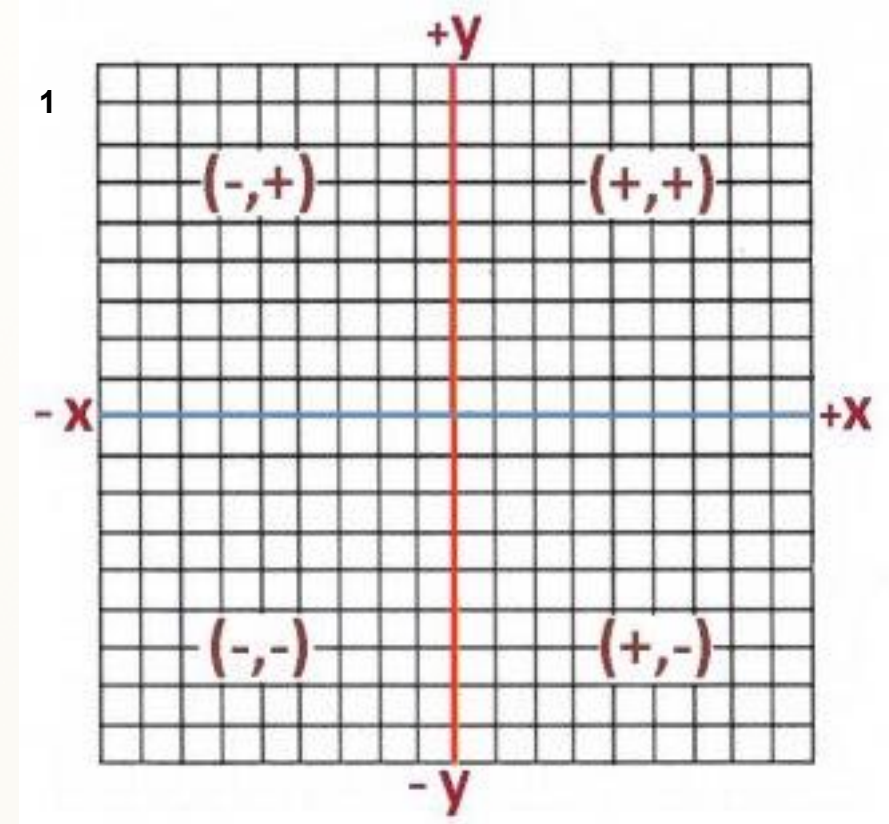
$$y = (x - 1)^2$$



El diagrama de dispersión

Es una herramienta matemática muy útil para observar de forma gráfica la relación entre variables. Utiliza el plano cartesiano para mostrar los valores de cada variable en los ejes X y Y y usando dichos valores para localizar dentro del plano la posición de cada caso medido.

En la siguiente imagen se observa un ejemplo del plano cartesiano, para los diagramas de dispersión se utiliza únicamente el Cuadrante I.



Plano cartesiano

a. Relación de las variables y su significado

La mayor virtud del diagrama de dispersión es su utilidad para mostrar relaciones de variables cuando una de ellas está bajo control del investigador. A esta variable se le denomina “independiente” y se coloca en el eje X, el eje horizontal. La otra variable, la que medimos, se coloca en el



eje Y del plano, el eje vertical y se le denomina como la variable “dependiente”.

Cuando se desean mostrar relaciones sin variable dependiente e independiente, no es importante cuál de las dos variables se coloca en qué posición.

Probablemente el aspecto más importante es que al considerar la elaboración de una gráfica de dispersión las unidades de medida deben ser las mismas.

b. Elaboración en Excel

Para elaborar los gráficos de dispersión en Excel, vamos a pensar en un ejemplo y lo haremos juntos paso a paso. Recordemos que el diagrama sirve para representar relaciones de variables.

Pensaremos en un estudio en el ámbito laboral. Resulta que la encargada de recursos humanos está interesada en conocer cuánto duermen los empleados de una institución, porque ha leído que la productividad en el trabajo se relaciona con la cantidad de horas que se duermen.

Así que decide investigar si esa relación es cierta. Para ello, empieza a pedirle a todos los empleados que al llegar a la empresa anoten en un listado la cantidad de horas que durmieron, esta variable será la independiente en el estudio (eje x) ya que no podemos “controlarla” o “medirla” y queremos saber si ella se relaciona con las horas productivas de trabajo.

Para medir la variable dependiente (eje Y) le pide al encargado de informática que monitoree el tiempo que cada persona dedica a realizar tareas relacionadas con la oficina, de esta forma tendrá la medida de “horas productivas de trabajo”.

Luego de cierto tiempo de recabar información, la encargada de recursos humanos tiene una tabla de datos con esta información.

2

Horas de sueño	Horas productivas de trabajo
5	5
7	3
6	3
6	5
4	5
10	8
4	6
7	6
4	3
9	4
6	4
8	5
9	7
8	5
8	5
9	4
4	3
9	8
4	4
8	5
7	3
7	5
4	3
9	8
7	5
7	6
7	4
10	6
5	5
5	3
7	4

Tabla de datos

Con el fin de encontrar la relación entre las variables, nos pide que le ayudemos a saber si se relacionan las variables.

El primer paso, es identificar cuál es la variable dependiente y cuál la independiente, como sugerencia puede pensarse en dos cosas:

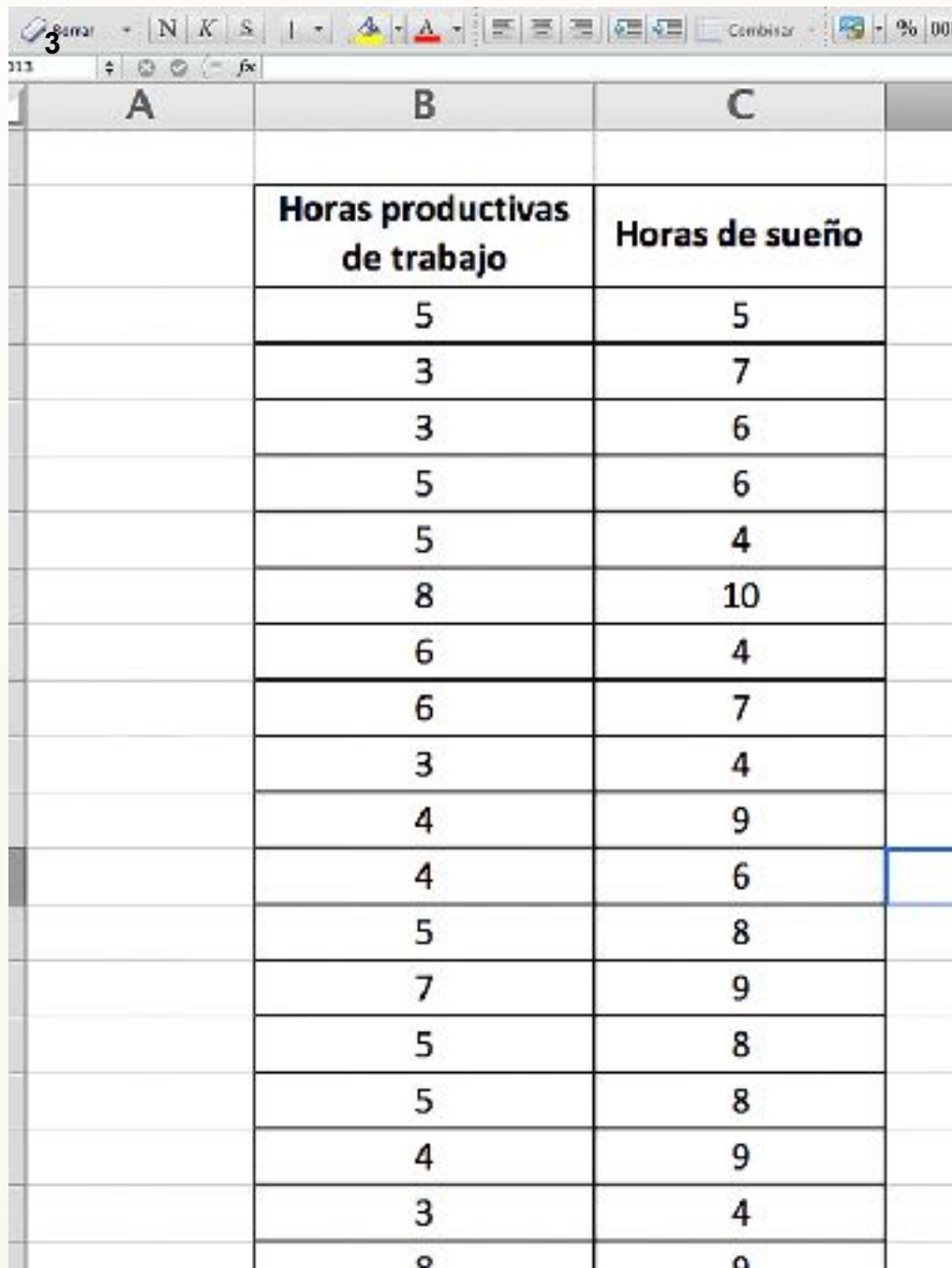


1. ¿Qué variable sucede primero? Normalmente la que sucede primero es la independiente.
2. ¿Qué variable estoy midiendo? Normalmente la variable que medimos es la dependiente.

En el caso de nuestro ejemplo:

- Dependiente = Horas de sueño
- Independiente = Horas productivas de trabajo

Segundo paso es ordenar las variables en Excel, por programación, el programa está diseñado para tomar la primera variable en las columnas como la variable Independiente (eje X) y la segunda en aparecer como la Dependiente (Eje Y), así que debemos asegurarnos que estén colocadas así.



The image shows a spreadsheet with three columns labeled A, B, and C. Column B is titled 'Horas productivas de trabajo' and column C is titled 'Horas de sueño'. The data is as follows:

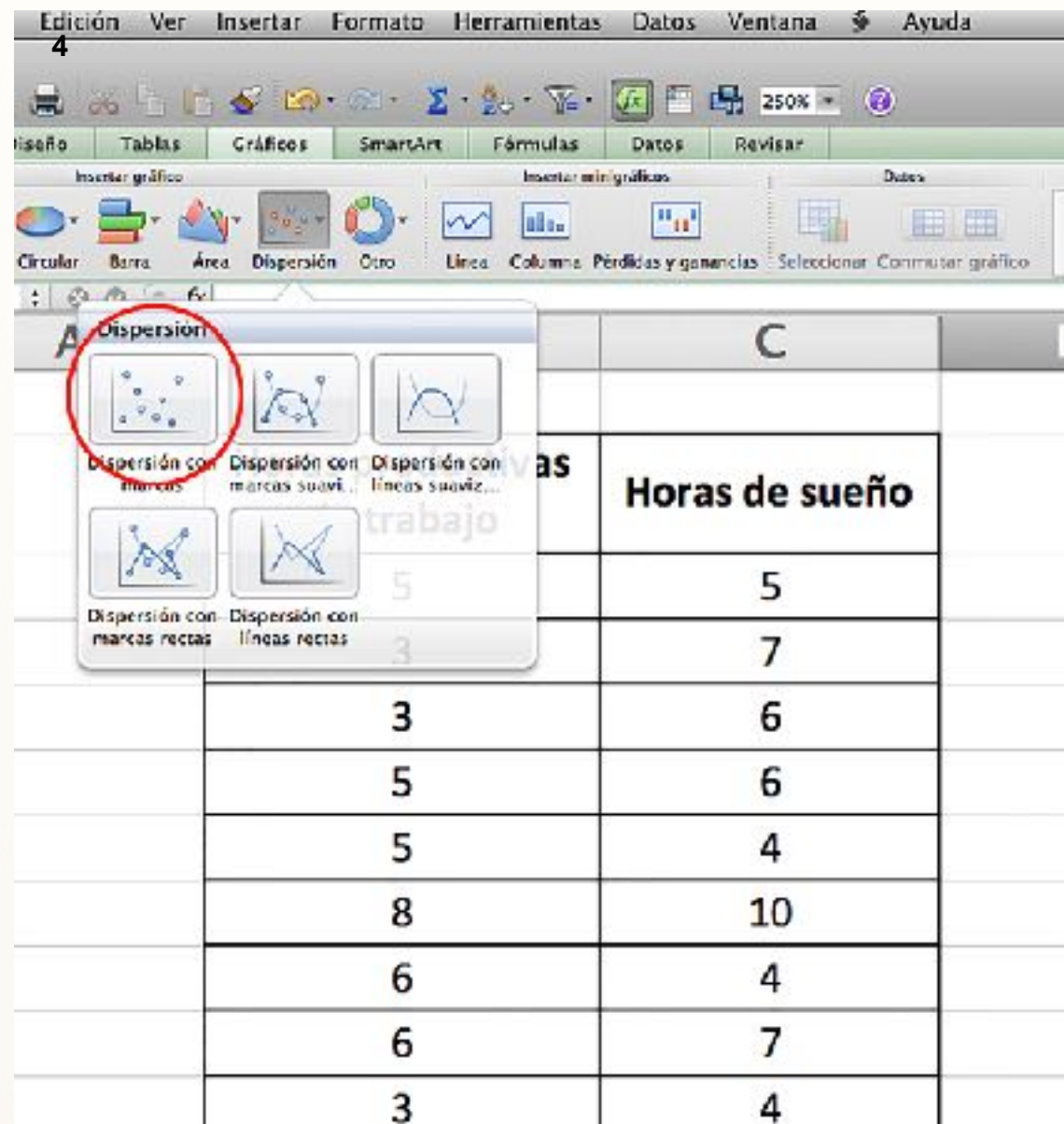
	B	C
	Horas productivas de trabajo	Horas de sueño
	5	5
	3	7
	3	6
	5	6
	5	4
	8	10
	6	4
	6	7
	3	4
	4	9
	4	6
	5	8
	7	9
	5	8
	5	8
	4	9
	3	4
	2	0

Orden de variables

Podemos ver en este caso que están organizadas en este sentido. La columna C es la variable “horas de sueño” que representa la Dependiente y se constituirá como el eje Y de nuestro plano cartesiano.

La columna B es la variable “horas productivas de trabajo” que representa la Independiente y constituirá el eje X del plano cartesiano.

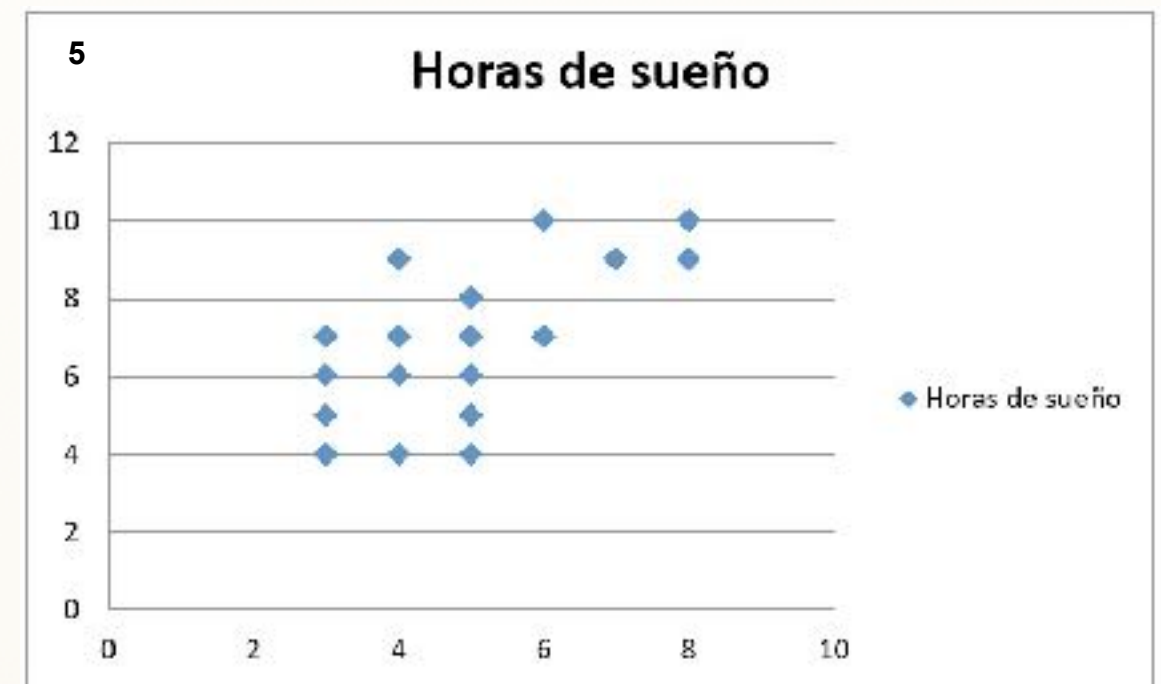
El tercer paso es seleccionar las dos columnas de datos. (Tener en cuenta que se deben seleccionar los datos y no las columnas enteras con datos vacíos). Ubicar la pestaña “INSERTAR” y allí buscar las gráficas de dispersión, tal y como lo muestra el círculo rojo en la imagen. Automáticamente aparecerá la gráfica.



Dispersión

El programa generará entonces el diagrama de dispersión de acuerdo a los datos que nosotros seleccionamos, deberemos tener en cuenta que Excel selecciona los parámetros por un proceso interno y no necesariamente son los parámetros que mejor describen a nuestros datos.

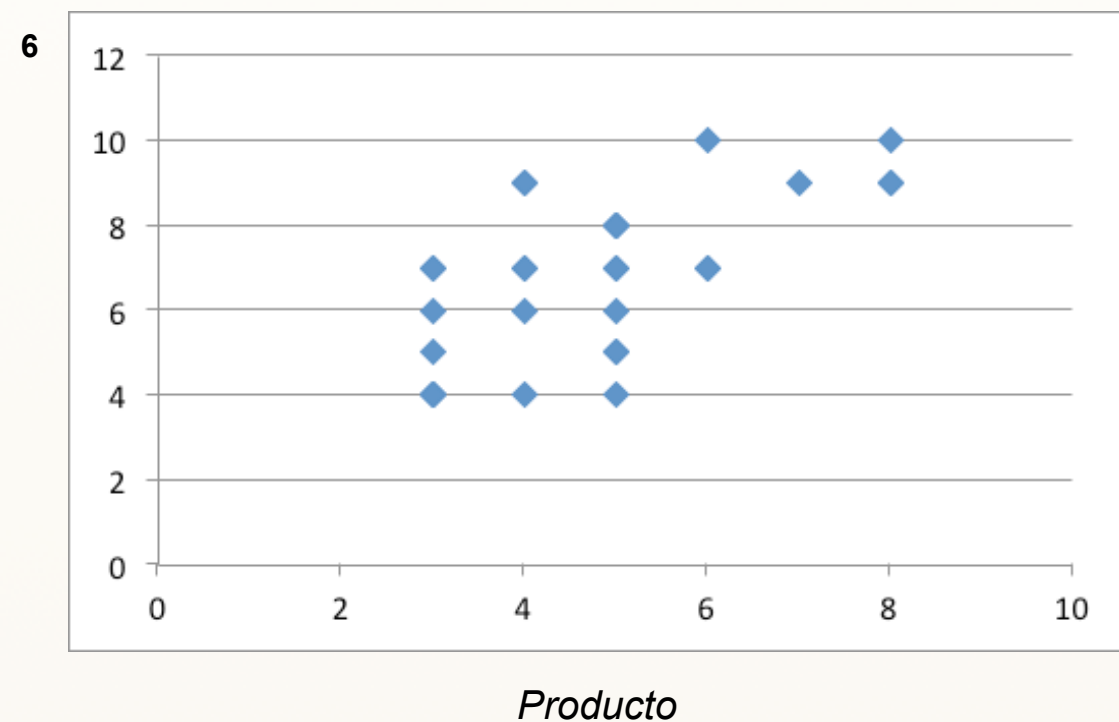
Veamos en el ejemplo el resultado que genera Excel como producto:



Dispersión

Aunque en principio, la gráfica puede funcionarnos podemos editarla para que represente mejor los datos que tenemos. Por ejemplo, no siempre aparecen los rótulos para los ejes o aparecen sin nombre, el eje Y inicia en 0 y nuestros datos inician en 4 por lo que queda un espacio vacío. El eje X tiene una situación similar, así que procederemos a editar la gráfica.

Primera edición: Eliminar etiquetas, nos colocamos sobre las etiquetas “Horas de sueño” y presionamos la tecla “DELETE”. El producto se ve así:



Segunda edición: Eje Y. Necesitamos que represente mejor a nuestros datos, las horas de sueño en nuestra investigación inician en 4 y llegan hasta 10, por lo que no es necesario tener tanto espacio en blanco en el extremo superior.

Vamos a hacer clic derecho sobre el eje Y para luego seleccionar la opción “dar formato al eje”.



Al hacerlo aparecerá un nuevo menú con varias opciones de modificación, nos interesa la unidad máxima y la podemos modificar a 10 que es el dato más alto que tenemos. La unidad mínima es recomendable dejarla en 0 para representar bien el eje cartesiano, aunque es cuestión de estilo personal.

La tercera edición corresponde al Eje X. Haremos el mismo procedimiento pero en esta ocasión cambiaremos otros detalles.

Vamos a modificar las dos áreas marcadas. La primera porque nuestros datos muestran que el máximo de horas trabajadas es 8, así que no es necesario tener espacio hacia la derecha. El segundo porque corresponde al intervalo que muestra el eje X y deseamos que muestre cada opción de la variable y no que lo resuma de 2 en 2.

El resultado deja la gráfica como se muestra a continuación...

8

Opciones del eje

Mínima: ☒ Automático ☐ Fija 0.0

Máxima: ☐ Automática ☒ Fija 8.0

Unidad mayor: ☐ Automática ☒ Fija 1

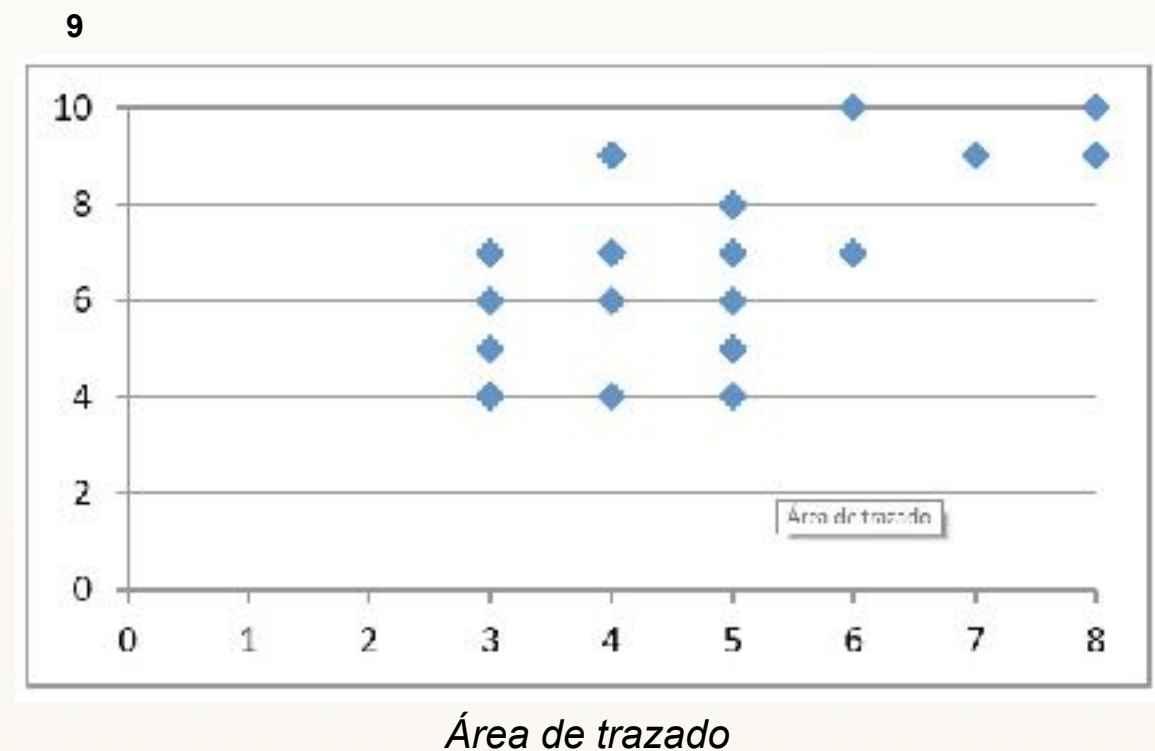
Unidad menor: ☐ Automática ☒ Fija 0.4

☐ Valores en orden inverso

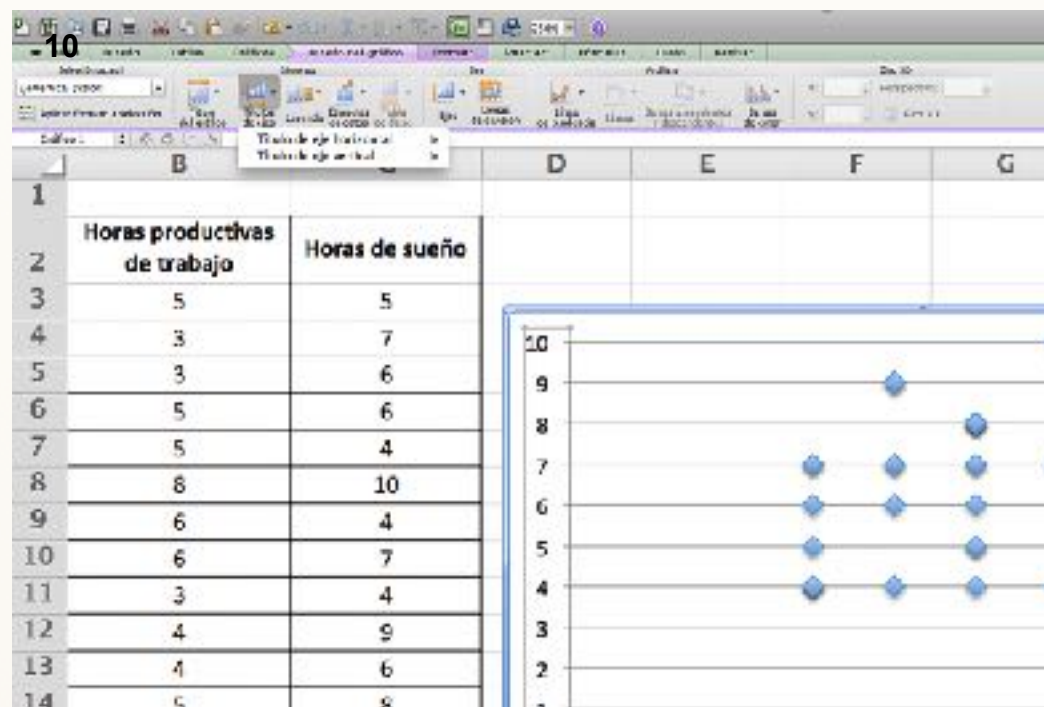
☐ Escala logarítmica Base: 10

Opciones del Eje.

Un máximo de 10 horas de sueño y un máximo de 8 horas productivas de trabajo. El espacio de la izquierda, en blanco, se debe a la ausencia de datos y el investigador debe decidir si requiere usarlo o no. En este caso, cabría la posibilidad de que haya gente que no haya dormido y que haya gente que solamente sea productiva 1 hora. No es el caso de nuestros datos, pero podría ser, así que se decide dejar dichos rangos en la gráfica.

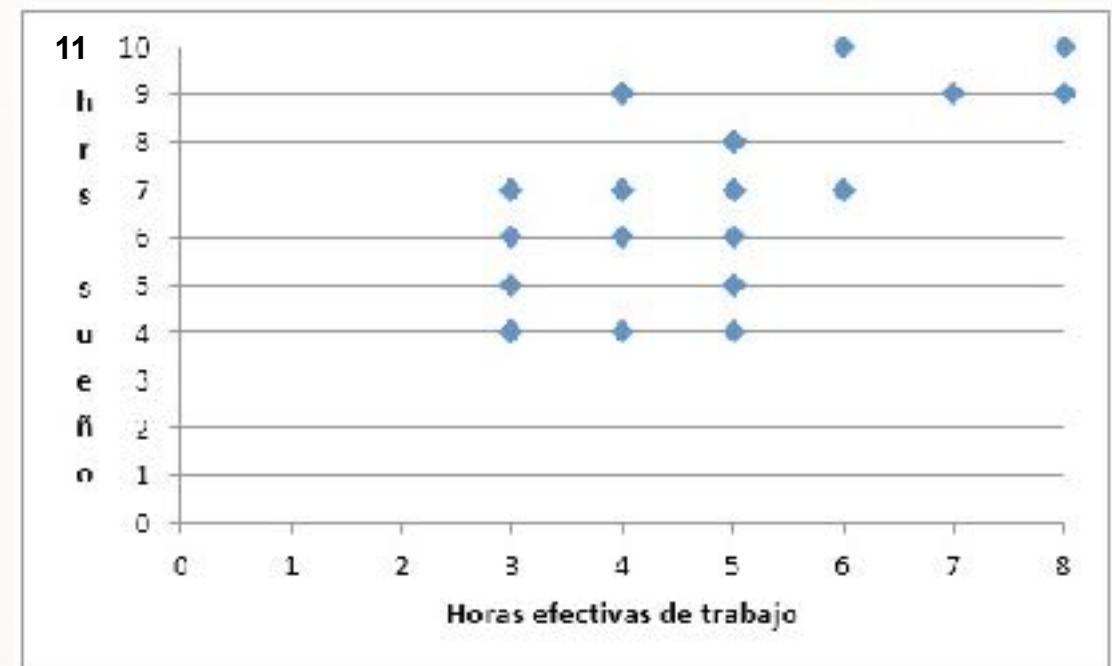


La cuarta edición corresponde a añadir títulos de ejes. Para ello debemos hacer clic en la gráfica y ubicar la pestaña de “presentación”, allí ubicaremos la opción “rótulos de ejes” y podemos seleccionar el eje horizontal y el vertical...



Títulos de Ejes

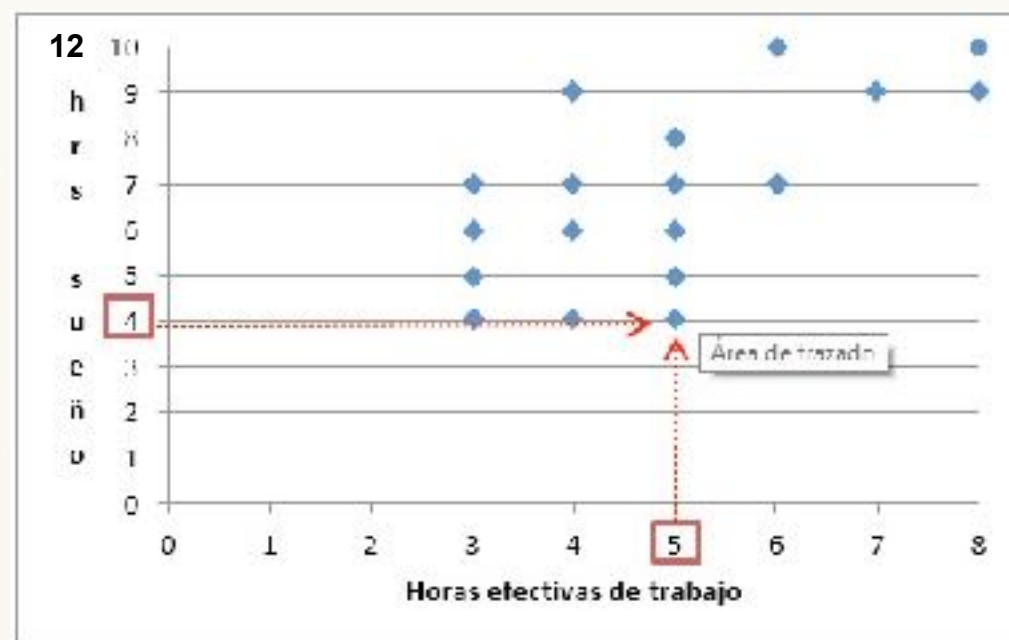
El producto de esta edición es una gráfica con los ejes bien identificados como se puede apreciar.



Producto

c. Interpretación

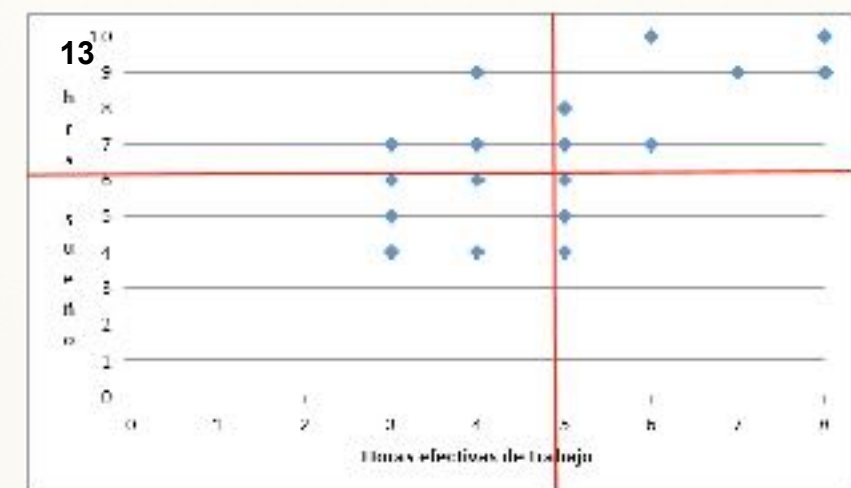
La interpretación visual de la gráfica se realiza siguiendo la lógica del plano cartesiano. Cada punto en la gráfica representa un caso (los datos de una persona en ambas variables) y para graficarlo se ubica primero En el eje Y está la variable dependiente, la “horas de sueño” el 0 hasta abajo y 10 arriba y luego se busca en el eje X el valor de la variable independiente “horas efectivas de trabajo”. Así se llena la gráfica que podemos ver. Por ejemplo, si Rodrigo González dijo que durmió 5 horas de trabajo y su reporte de horas laborales efectivas dice que trabajó 4, podemos encontrarlo en la gráfica...



Datos de una persona en ambas variables

En este caso, allí localizaríamos los datos de Rodrigo González.

Así que siguiendo la lógica que mencionamos al ver la gráfica lo que necesitamos identificar es cuál es la relación de las variables y al parecer entre más horas de sueño de una persona, más horas de efectivas de trabajo.



Relación entre variables.

Una forma en la que se puede pensar la interpretación es dividiendo nuestro cuadrante en 4 y viendo qué significan los puntos que quedan en cada cuadrante y dónde hay más datos.

En este ejemplo, el cuadrante superior derecho significa más horas de sueño (porque está arriba) y más horas de trabajo productivo (porque está más a la derecha) y vemos que hay bastantes datos allí.

También vemos el cuadrante inferior izquierdo, pocas horas de sueño y pocas horas productivas, también hay varios datos allí. Los dos cuadrantes adicionales significan muchas horas de sueño y poca productividad (cuadrante esquina superior izquierda); o poco sueño y mucha productividad y vemos que tienen pocos datos (cuadrante esquina inferior derecha).

Así pues, podemos de forma gráfica, considerar que parece haber una relación positiva (si una sube, la otra también) entre las horas de sueño y las horas productivas de trabajo.

Para poder corroborarlo estadísticamente, aprenderemos el coeficiente de correlación.



Comprobación del capítulo XVII: El diagrama de dispersión

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XVII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XVIII

Coeficiente de correlación lineal



Coeficiente de correlación lineal

a. Definición y propiedades

Generalmente se usa el término correlación para referirse a la correspondencia o vinculación existente entre dos situaciones, cosas o personas. En estadística es una medida de la fuerza y la dirección que existe entre dos variables.

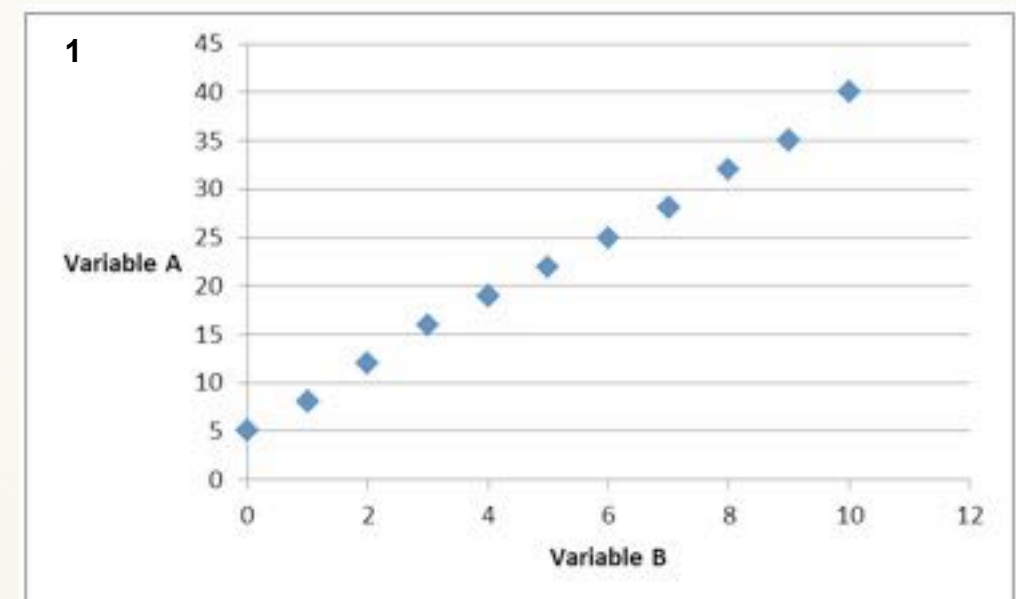
En otras palabras, podremos decir que las variables se correlacionan cuando una de ellas varía en función de la otra. Por ejemplo, supongamos que tenemos una variable A y una variable B. Diremos que están correlacionadas cuando al modificar A, B sufre también una modificación en cualquier sentido.

El producto del cálculo de una correlación es el llamado “coeficiente de correlación” que se identifica en los textos como “r”. Varía dentro de un rango de -1.0 hasta +1.0 y su interpretación depende de la cercanía a los polos. Es decir, existe más correlación al acercarse a -1.0 o +1.0



Sus propiedades principales son las siguientes:

- La correlación se mantiene estable aunque cambie la escala de medición las dos variables.
- $-1 \leq r \leq 1$ el coeficiente de correlación es número real comprendido entre -1 y +1.
- Un coeficiente cercano a 0 implica correlaciones débiles y cercano a -1,+1 implica correlaciones fuertes.
- Si $r=1$, el diagrama de dispersión es una línea recta de puntos.



Correlación cercana a 1

b. Cálculo

El cálculo a mano de una correlación puede considerarse complejo, especialmente con cantidades de datos importantes. La fórmula para calcular la correlación es la siguiente:

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2] [n \sum y_i^2 - (\sum y_i)^2]}}$$

En el numerador multiplicamos el total de nuestros datos por la sumatoria del producto entre los valores de las dos variables y luego lo restamos al producto de la sumatoria de cada variable.

En el denominador calcularemos la raíz cuadrada del producto entre: el total de datos por la sumatoria del cuadrado de la variable x menos la sumatoria de x, elevada el cuadrado y la sumatoria del cuadrado de la variable y menos la sumatoria de Y elevada el cuadrado.

Haremos un ejercicio con datos tomados de un ejemplo clásico de la relación entre estatura y peso. Supondremos que medimos y pesamos a 10 personas y estos son sus datos:

2

Altura X	Peso Y
185	85
188	85
189	86
191	90
192	87
192	91
197	93
200	103
202	100
204	101

Valor X, Y

Necesitaremos crear dos columnas nuevas en donde obtengamos el cuadrado de cada valor X y cada valor Y, eso se hace multiplicando el valor por sí mismo. También obtendremos la sumatoria de cada columna.

3	Altura "X"	X cuadrado	Peso "Y"	Y cuadrado	XY
P1	185	34,225	85	7,225	15,725
P2	188	35,344	85	7,225	15,980
P3	189	35,721	86	7,396	16,254
P4	191	36,481	90	8,100	17,190
P5	192	36,864	87	7,569	16,704
P6	193	37,249	91	8,281	17,472
P7	197	38,809	93	8,649	18,321
P8	200	40,000	103	10,609	20,600
P9	202	40,804	100	10,000	20,200
P10	204	41,616	101	10,201	20,604
SUMATORIA	1,940	376,728	921	85,255	179,050

Sumatoria de columnas

1) Empezaremos a trabajar el numerador identificando que $N = 10$, la sumatoria de X por Y es 179,050.

Así que operamos N por Sumatoria $X*Y = 1,790,500$.

Ahora debemos realizar la multiplicación de la sumatoria de X (1,940) por la sumatoria de Y(921) = 1,786,740.

Así que para obtener el valor del numerador solo nos queda restar ambos números
 $1,790,500 - 1,786,740 = 3,760$ Este es el numerador.

2) Ahora calculamos el valor del denominador, siempre respetando las jerarquías de las operaciones.

Empezamos calculando $N=10$ por la sumatoria de X cuadrado: $N * 376,728 = 3,767,280$.

Ahora calculamos el cuadrado de la sumatoria de X, es decir $1940^2 = 3,763,600$.

Por último restamos ambos valores: $3,767,280 - 3,763,600 = 3,680$.

Ahora hacemos el mismo procedimiento para Y, $N=10$ por sumatoria de Y cuadrado = 852,550.

El cuadrado de la sumatoria de Y= $921^2 = 848,241$.

Restamos ambas cifras = $852,550 - 848,241 = 4,309$.

Ahora debemos multiplicar la operación de X por la de Y:

$$3,680 * 4,309 = 15,857,120$$

Ahora operamos raíz cuadrada de dicho valor. El resultado para el denominador = 3,980.

Para obtener r entonces completamos la fórmula haciendo la división.

$$r = 3,760 / 3,980$$

$$r = 0.94$$

El proceso se ve complejo, pero siguiendo un orden se puede realizar sin problemas. Ahora bien, como hemos aprendido hasta este punto Microsoft sale al rescate con su poderosa herramienta de Excel. Así que aprenderemos a hacer el proceso usando las funciones del programa pero antes veremos cómo se categoriza la correlación.

c. Categorización

La correlación se puede categorizar por su dirección y por la fuerza, al final del capítulo veremos un juego interactivo para practicar nuestra habilidad de interpretar la correlación en un diagrama de dispersión.

Para clasificar por su fuerza diversos autores proponen distintas clasificaciones ya que no hay un consenso en el mundo de las ciencias. Sin embargo creo que podemos usar esta tabla de referencia

4	-+1.0	Correlación perfecta
	-+0.85	Correlación fuerte
	-+0.50	Correlación moderada
	-+0.20	Correlación débil
	0	No hay correlación

De acuerdo a la dirección puede ser directa o inversa y nos permite determinar el sentido o la forma en que una variable está enlazada a la otra. Para entenderlo mejor veamos un par de ejemplos:

Se dice que el tiempo que una persona pase en el tráfico durante un período de 10 años está relacionado con la cantidad de accidentes de tráfico que la persona tiene en ese mismo período de tiempo.



Tráfico en Guatemala

Variable 1 = tiempo en el tráfico en los últimos 10 años.

Variable 2 = Cantidad de accidentes de tránsito que ha sufrido cada persona.

Estadísticamente podremos encontrar una **CORRELACIÓN** porque es lógico que entre más tiempo pase una persona en su vehículo entre el tráfico, tiene más probabilidad de sufrir un accidente.

Este tipo de correlación se llama **DIRECTA**, porque mientras una variable aumenta, la otra también lo hace, entre más tiempo... más accidentes.



Ahora veamos un ejemplo del tipo de correlación indirecta. Pensemos en el tiempo que una persona realiza actividades físicas durante el día y la cantidad de grasa corporal. Una nutricionista pide a un grupo de persona que registren durante un mes toda actividad física que realizan y luego toma sus medidas antropométricas para determinar el porcentaje de grasa corporal de los participantes.



Ejercicio físico

Luego de revisar los datos, encuentra que el porcentaje de grasa corporal disminuye cuando el tiempo de actividad física aumenta.

En este caso la correlación es **INDIRECTA**, mientras una variable aumenta, la otra se reduce.



En un diagrama de dispersión podemos ver la dirección de la correlación. Así se verían nuestros ejemplos.

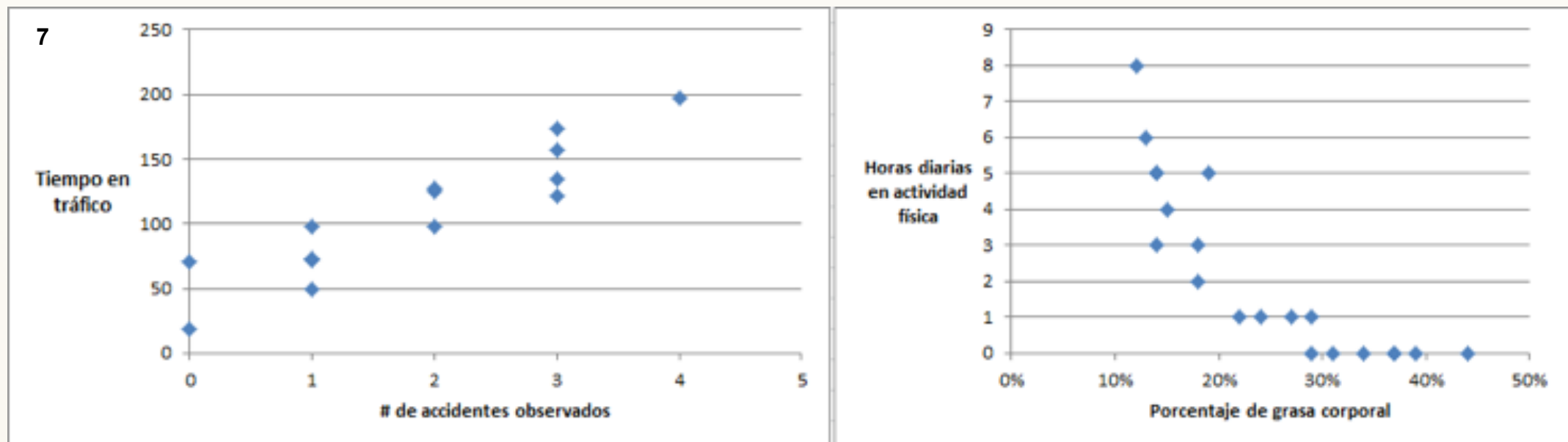


Diagrama de dispersión

Ahora bien, de acuerdo a su **FUERZA**, la correlación puede ser fuerte o puede ser débil y esta clasificación nos permite representar de mejor manera a dos variables correlacionadas.

La mejor forma de entender la situación es a través de un juego online que lamentablemente no se encuentra en español. Pero es tan simple que cualquiera puede usarlo y aprender la fuerza de la correlación.

Recordemos que es más fuerte al acercarse a -1 o $+1$ y la nube de puntos se asemeja a una línea y es más débil al acercarse a 0 y la nube de puntos parece más una nube.





Comprobación del capítulo XVIII: Coeficiente de correlación lineal

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XVIII

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Resolución de problemas usando correlación

Resolución de problemas usando correlación

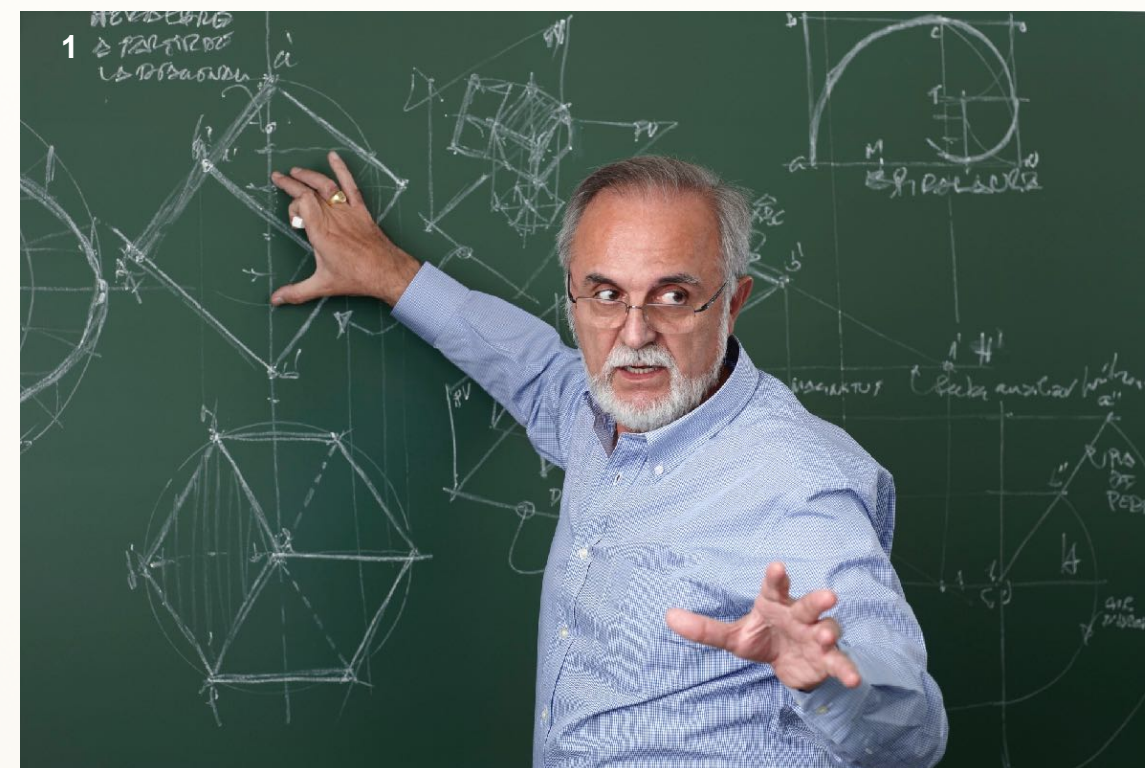
Resolución de problemas usando correlación

Para el resto de la unidad vamos a diseñar un ejercicio que nos permita practicar la correlación, entenderla mejor y hacer uso del paquete estadístico Microsoft Excel. Posteriormente aprovecharemos los datos para entender también la regresión.

Vamos a suponer que existe un director de un centro educativo X que se ha preocupado por los resultados que publica el Ministerio de Educación. Ha escuchado muchas quejas que dicen que las pruebas no son válidas porque los alumnos no se las toman en serio. Pero aun así los papás de sus alumnos les molesta ver que los resultados no son mejores.

Como el director sabe que no está solo y que hay más directores en su misma posición, decide hacer una convocatoria y reúne a varios

establecimientos con el mismo problema y deciden que la mejor forma de resolver el problema es contratando a un investigador.



Correlación cercana a 1

Encuentran a una investigadora experta en evaluación estandarizada que les sugiere realizar una evaluación estandarizada con otro instrumento, al menos un mes antes que la prueba del Ministerio y luego usar esos resultados para compararlos con los resultados que presenta el Ministerio de Educación.

Luego de hacer la muestra, determinan que necesitan 110 estudiantes escogidos al azar entre todos los establecimientos y ellos tomarán la prueba que escogieron con la investigadora.

Los resultados del estudio se presentan en una tabla de datos para los 110 estudiantes que tomaron ambas pruebas. Ahora necesitamos determinar si hay o no relación entre las pruebas. Si encontramos que No hay relación, tendremos evidencia para decir que las pruebas del Ministerio no son válidas, de lo contrario,

2

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
Prueba Investigación	50	50	50	50	53	54	56	59	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
Prueba Ministerio	59	62	52	51	49	63	63	61	70	69	64	63	64	65	65	65	63	61	63	67	64	63	64	65	65	65	65	63	61	63

	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	P41	P42	P43	P44	P45	P46	P47	P48	P49	P50	P51	P52	P53	P54	P55	P56	P57	P58	P59	P60
Prueba Investigación	61	62	62	62	62	62	63	63	63	63	63	64	71	71	73	73	74	74	74	74	74	75	77	79	79	80	80	80	80	80
Prueba Ministerio	67	64	65	61	69	60	59	64	68	67	64	65	76	73	72	81	79	70	71	79	71	76	87	74	74	80	85	80	79	83

	P61	P62	P63	P64	P65	P66	P67	P68	P69	P70	P71	P72	P73	P74	P75	P76	P77	P78	P79	P80	P81	P82	P83	P84	P85	P86	P87	P88	P89	P90
Prueba Investigación	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	81	82	82	83	83	83	84	84	85	85	85	87
Prueba Ministerio	85	84	86	79	83	85	84	86	77	85	80	79	83	85	84	86	77	85	84	83	84	79	81	86	90	87	89	80	84	85

	P91	P92	P93	P94	P95	P96	P97	P98	P99	P100	P101	P102	P103	P104	P105	P106	P107	P108	P109	P110
Prueba Investigación	87	88	90	90	90	90	90	91	91	92	93	95	95	97	98	99	99	100	100	100
Prueba Ministerio	89	92	94	91	92	91	93	90	94	90	93	94	96	95	98	94	99	97	99	100

Tabla de datos

contaremos con información científica para empezar a creer en ellas.

a. Ejercicios de cálculo

El primer punto requiere que identifiquemos nuestra variable independiente “X” y la variable dependiente “Y”. Si seguimos la sugerencia presentada anteriormente, pensamos ¿qué sucedió primero? La respuesta es la prueba de la investigadora por lo tanto es la variable independiente.



Con el fin de explorar los datos, podemos iniciar

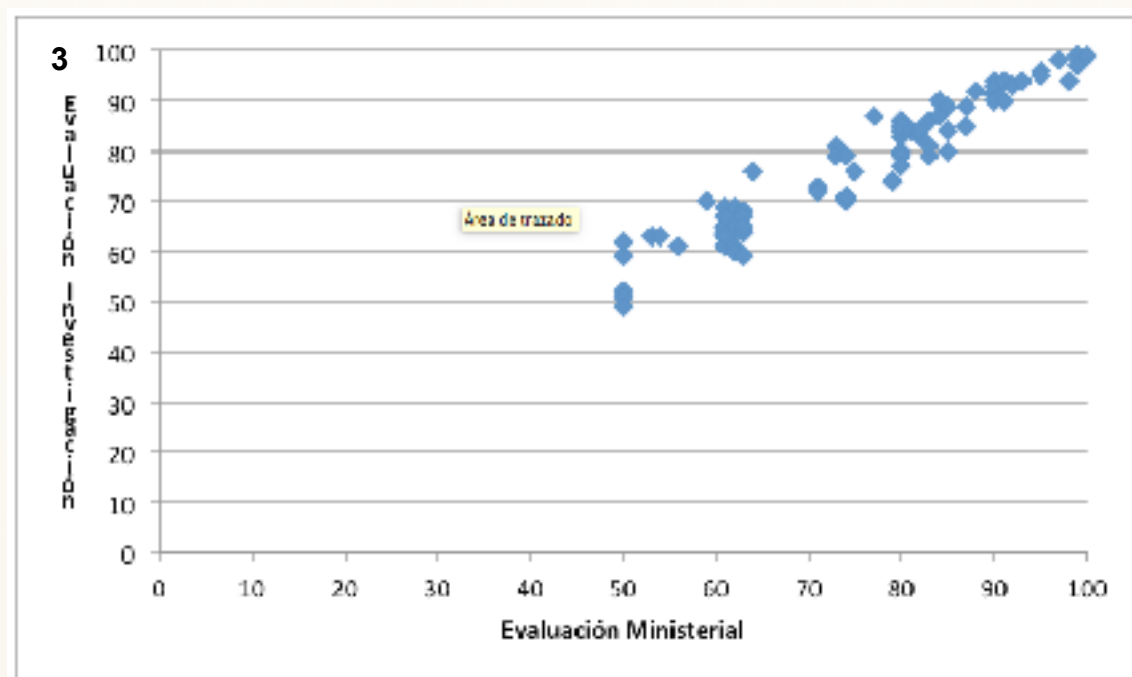


Diagrama de dispersión
por hacer el diagrama de dispersión ¿podemos

adivinar el resultado?



Ahora debemos organizar los datos necesitamos 5 columnas una para X, una para X², otra para Y, otra para Y² y una última para la XY (la multiplicación de X por Y)



Nota:

La tabla completa se encuentra en el anexo 1.

El numerador: En este caso $N=110$ y la sumatoria de $XY = 633136$ así que el producto es 69644960

Ahora realizamos la sumatoria de X por la sumatoria de $Y = 67182494$

Y entonces restamos para obtener el numerador = 2462466

El denominador: N por la sumatoria de $X^2 = 67759890$

Ahora el cuadrado de la sumatoria de $X = 65108761$

Y restamos los valores = 2651129

Repetimos el proceso pero ahora con Y
N por la sumatoria de $Y^2 = 71753220$

Ahora el cuadrado de la sumatoria de $Y = 69322276$

Y restamos los valores = 2430944

La diferencia de ambas cifras = 220185

El denominador será el resultado de multiplicar la operación de X por la de Y y al resultado sacarle la raíz cuadrada = 2538650.45

Ahora dividimos numerador /
denominador = 0.96 = r



Tenemos una correlación positiva y fuerte entre ambas variables. Lo que le indica al profesor que las pruebas que usa el ministerio de educación parecen ser válidas ya que las personas obtienen resultados correlacionados en dos pruebas diferentes que miden el mismo constructo.

b. Ejercicios en Excel

Un padre de familia ha decidido buscar financiamiento para la escuela pública de su comunidad, tiene la intención de comprar libros y darlos a los papás de los estudiantes para que tengan disponibilidad en casa.

Cuando hace la propuesta a la comunidad la mayoría de personas aplaude el esfuerzo pero hay un grupo de maestros que no están seguros que eso funcione. Ellos argumentan que solo con tener libros no va a mejorar la lectura, porque los padres necesitan saber cómo usarlos para que sus hijos refuercen lo que aprenden.

El líder de la comunidad entonces propone que realicen una investigación, hacen una encuesta con los padres de familia preguntando cuántos libros tienen en su casa y quieren contrastar la información contra el promedio final en las notas de Comunicación y Lenguaje de los estudiantes.

4

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
Cantidad de libros	41	38	67	0	49	30	40	66	38	38	40	7	53	68	63	41	27	0	54	33	31	6	40
Comunicación y Lenguaje	78	59	72	80	46	77	75	47	61	56	76	61	73	76	47	68	79	44	49	71	88	56	65

	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	A39	A40	A41	A42	A43	A44	A45	A46
Cantidad de libros	68	54	32	5	12	56	1	12	35	59	73	28	14	25	4	47	12	21	57	5	33	28	0
Comunicación y Lenguaje	61	72	66	64	48	65	47	43	73	52	90	54	55	57	55	79	82	74	57	75	83	93	42

Datos

**Nota:**

Los datos son ficticios y generados aleatoriamente para uso práctico.

Luego de hacer la investigación tienen estos datos.

Lo primero que haremos es generar e identificar las variables dependiente e independiente. En este caso ¿Qué variable medimos? El rendimiento académico, así que esta es la dependiente (Y) y ¿qué variable sucede primero? Tener una cantidad de libros en casa, es entonces la independiente (X).

Haremos entonces un diagrama de dispersión y para ello necesitamos organizar los datos en dos columnas, recordemos el orden en el que deben colocarse para que Excel reconozca la X y la Y.

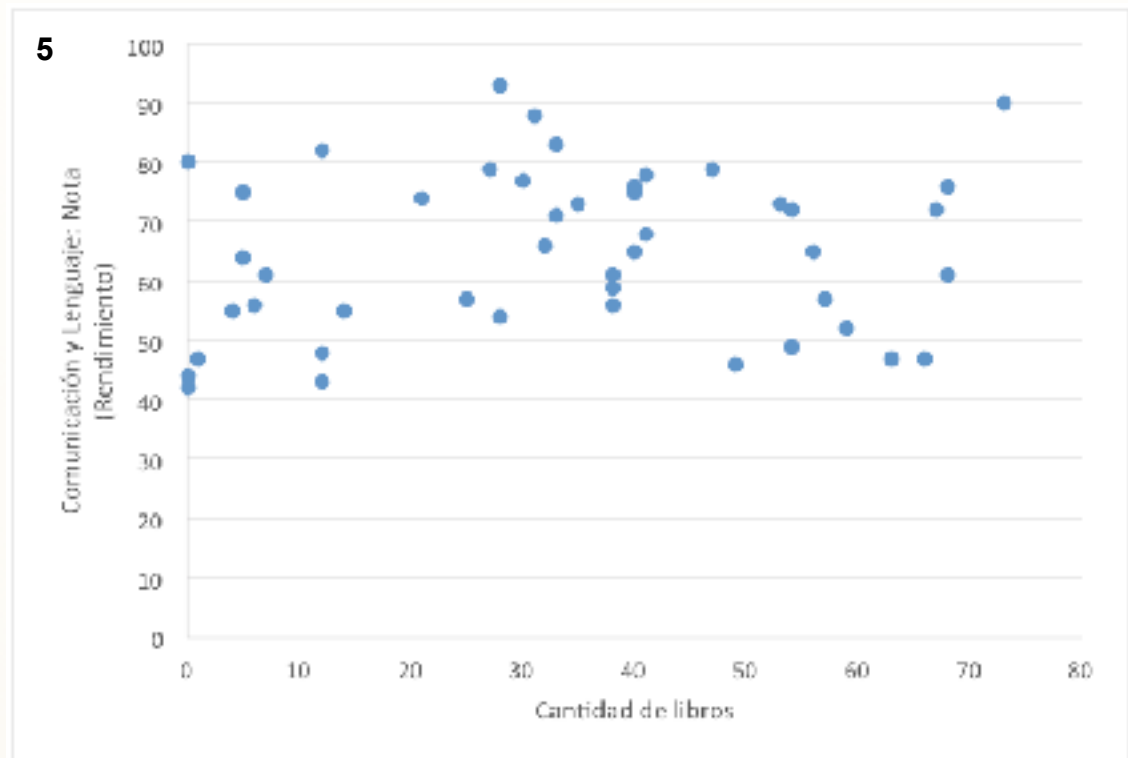


Diagrama de dispersión

Nos dirigimos a nuestros datos y buscamos el área de funciones, allí seleccionamos COEF.DE.CORREL y seleccionaremos cada columna de datos (así como aprendimos en el capítulo anterior).

Obtenemos un $r = .16$, es un valor cercano a 0 que nos dice que en este grupo de 75 estudiantes, prácticamente no existe relación entre la cantidad de libros que hay en casa y el resultado que obtienen en Comunicación y Lenguaje.

Recordemos que los datos son ficticios y no pretenden mostrar la realidad de la educación en nuestros países. Sin embargo, es posible encontrar estos datos en casos reales y en este ejemplo nos veríamos en la necesidad de escuchar más a los docentes que dicen que hay que educar a los padres para que el acceso a más libros produzca un buen resultado académico.



Comprobación del capítulo XIX: Resolución de problemas usando correlación

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XIX

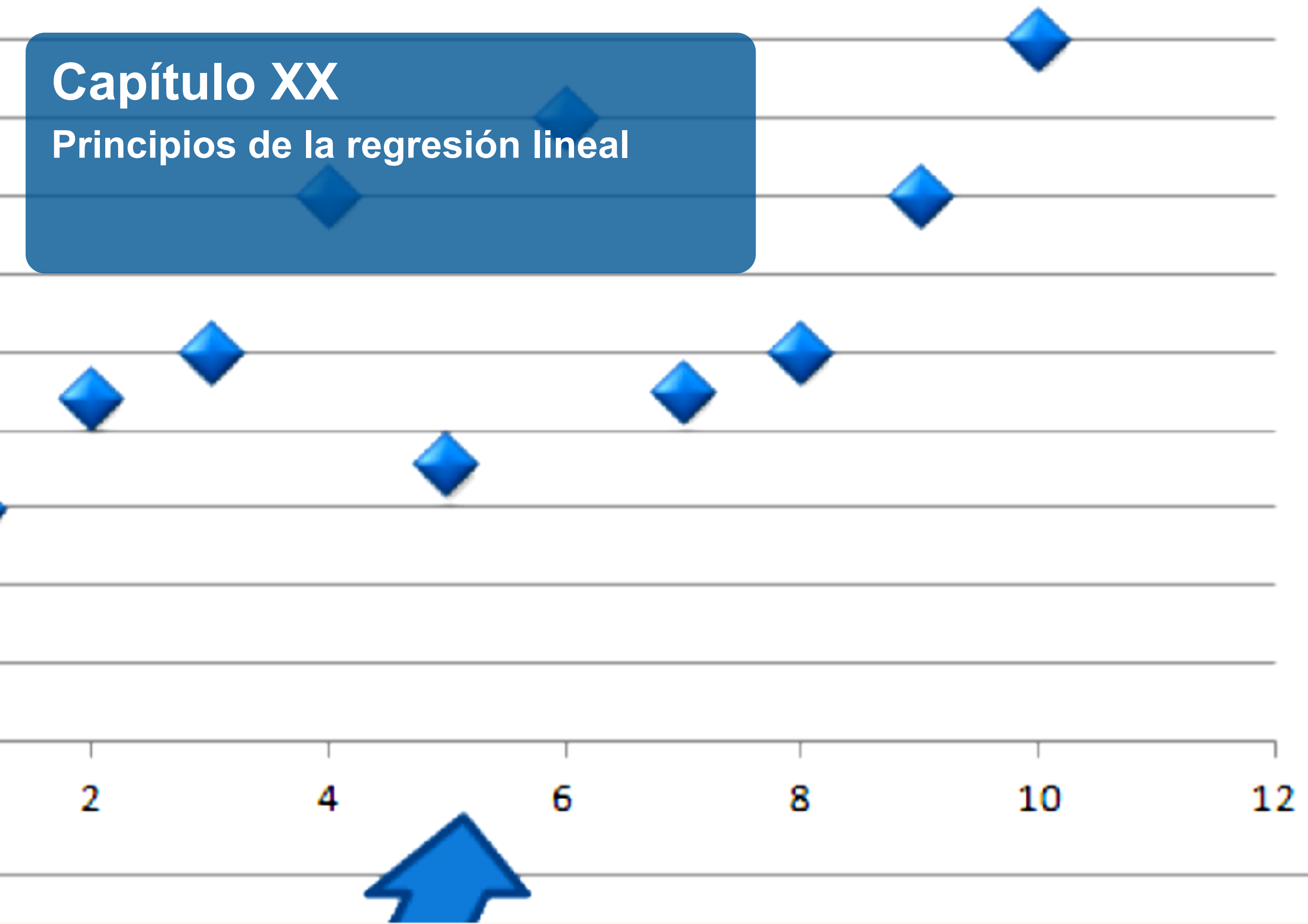
Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



Capítulo XX

Principios de la regresión lineal



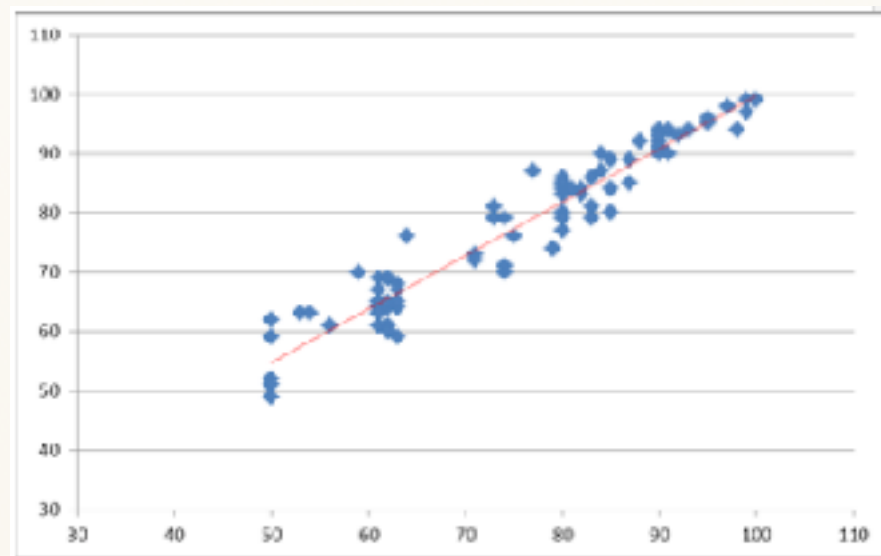
Principios de la regresión lineal

La regresión lineal, es un modelo matemático que analiza la relación entre dos o más variables.

Está estrechamente ligada al diagrama de dispersión, porque en palabras sencillas se trata de encontrar una función lineal o la línea recta que de mejor forma represente la nube de puntos. En sentido matemático hablaremos de la regresión como la fórmula que representa a dicha recta.

En la imagen podemos ver un ejemplo de línea de regresión sobre la nube de puntos.

1



Ejemplo de recta de regresión sobre nube de puntos.

a. Definición

La recta de regresión y su análisis son un modelo de predicción, se utiliza para determinar relaciones de dependencia entre las variables en estudio. Puede decirse también que es la representación matemática de la recta que mejor se ajusta a la nube de puntos.



Su representación matemática es $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

Donde:

Y_i = La variable dependiente que queremos predecir.

β_0 es un parámetro que define la intersección de la recta en el eje Y.

β_i = parámetro del modelo que representa la inclinación de la recta para la variable independiente.

X_i = valor de la variable independiente

ε = variable aleatoria que se considera como el error del modelo

b. Cálculo

Para realizar el ejercicio usaremos el mismo ejemplo que utilizamos en la correlación sobre la relación entre una prueba estandarizada diseñada y la prueba estandarizada del Ministerio de Educación.

La bondad de la regresión es que permite profundizar en los datos, porque luego de ver los resultados de la correlación, ahora queremos saber algo más de la información.



Recordemos que en la correlación únicamente determinamos que tanto se relacionaban los resultados en la prueba que hizo la investigadora con los resultados de la prueba del Ministerio de Educación.

Ahora queremos saber si podemos usar la prueba de la investigadora para predecir el resultado que se obtendría en la prueba del Ministerio.

Todos los cálculos los realizaremos en Microsoft Excel por su complejidad.

Así que el primer paso es generar el diagrama de dispersión para ello acudimos a los datos generados por los 110 alumnos evaluados...

2

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
Prueba Investigación	50	50	50	50	53	54	56	59	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
Prueba Ministerio	59	62	52	51	49	63	63	61	70	69	64	63	64	65	65	65	63	61	63	67	64	63	64	65	65	65	65	63	61	63

	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	P41	P42	P43	P44	P45	P46	P47	P48	P49	P50	P51	P52	P53	P54	P55	P56	P57	P58	P59	P60
Prueba Investigación	61	62	62	62	62	62	63	63	63	63	63	63	64	71	71	73	73	74	74	74	74	75	77	79	79	80	80	80	80	80
Prueba Ministerio	67	64	65	61	69	60	59	64	68	67	64	65	76	73	72	81	79	70	71	79	71	76	87	74	74	80	85	80	79	83

	P61	P62	P63	P64	P65	P66	P67	P68	P69	P70	P71	P72	P73	P74	P75	P76	P77	P78	P79	P80	P81	P82	P83	P84	P85	P86	P87	P88	P89	P90
Prueba Investigación	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	81	82	82	83	83	83	84	84	85	85	85	87
Prueba Ministerio	85	84	86	79	83	85	84	86	77	85	80	79	83	85	84	86	77	85	84	83	84	79	81	86	90	87	89	80	84	85

	P91	P92	P93	P94	P95	P96	P97	P98	P99	P100	P101	P102	P103	P104	P105	P106	P107	P108	P109	P110
Prueba Investigación	87	88	90	90	90	90	90	90	91	91	92	93	95	95	97	98	99	99	100	100
Prueba Ministerio	89	92	94	91	92	91	93	90	94	90	93	94	96	95	98	94	99	97	99	100

Tabla de datos

Tal y como ya aprendimos anteriormente, diseñamos nuestro diagrama de dispersión empleando todas las recomendaciones brindadas. Modificamos los ejes para representar bien los datos, usamos los rótulos de los ejes y editamos de acuerdo a nuestras preferencias.

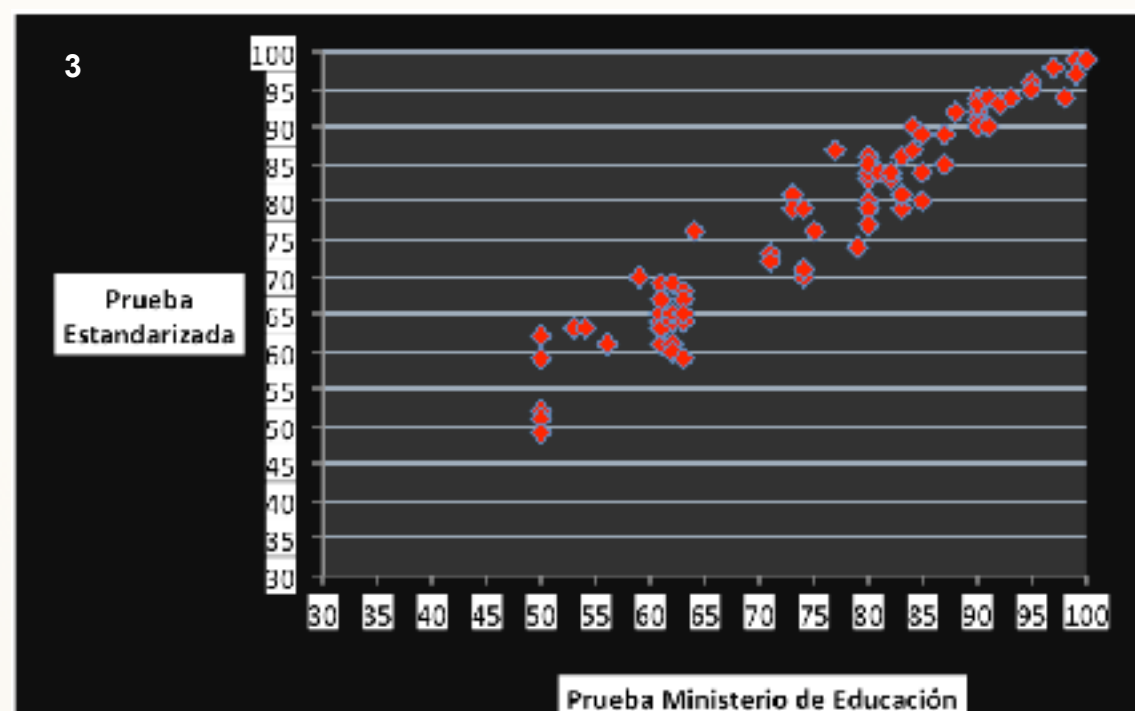
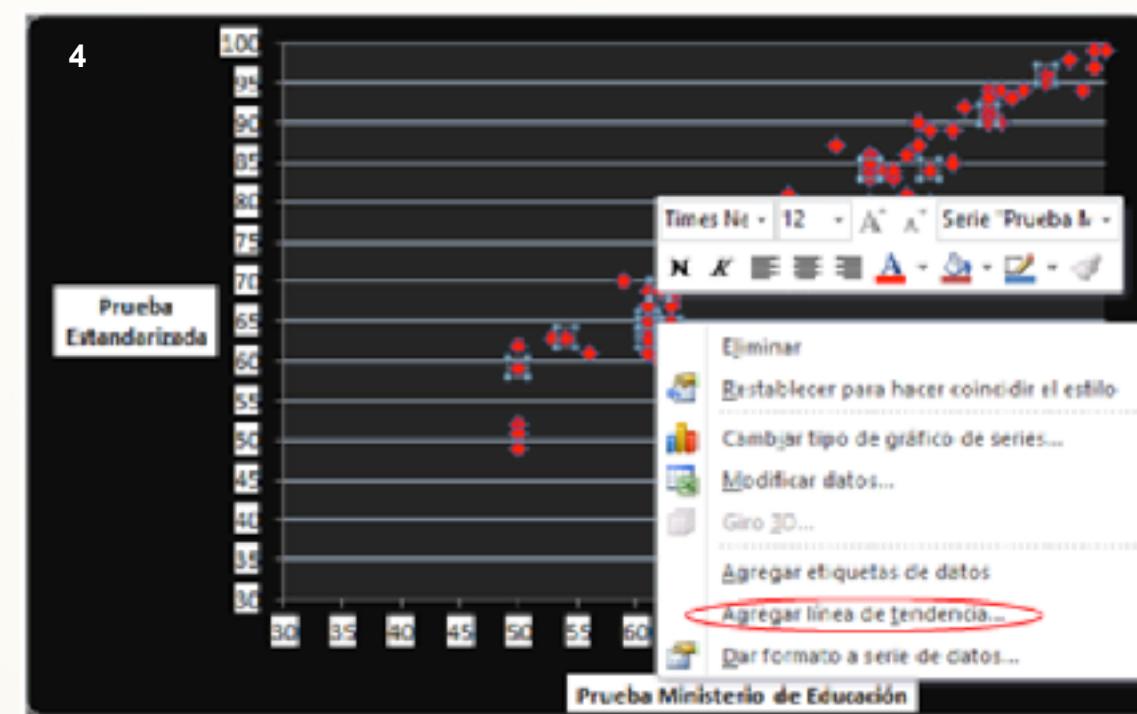


Diagrama de dispersión, ejemplo de regresión

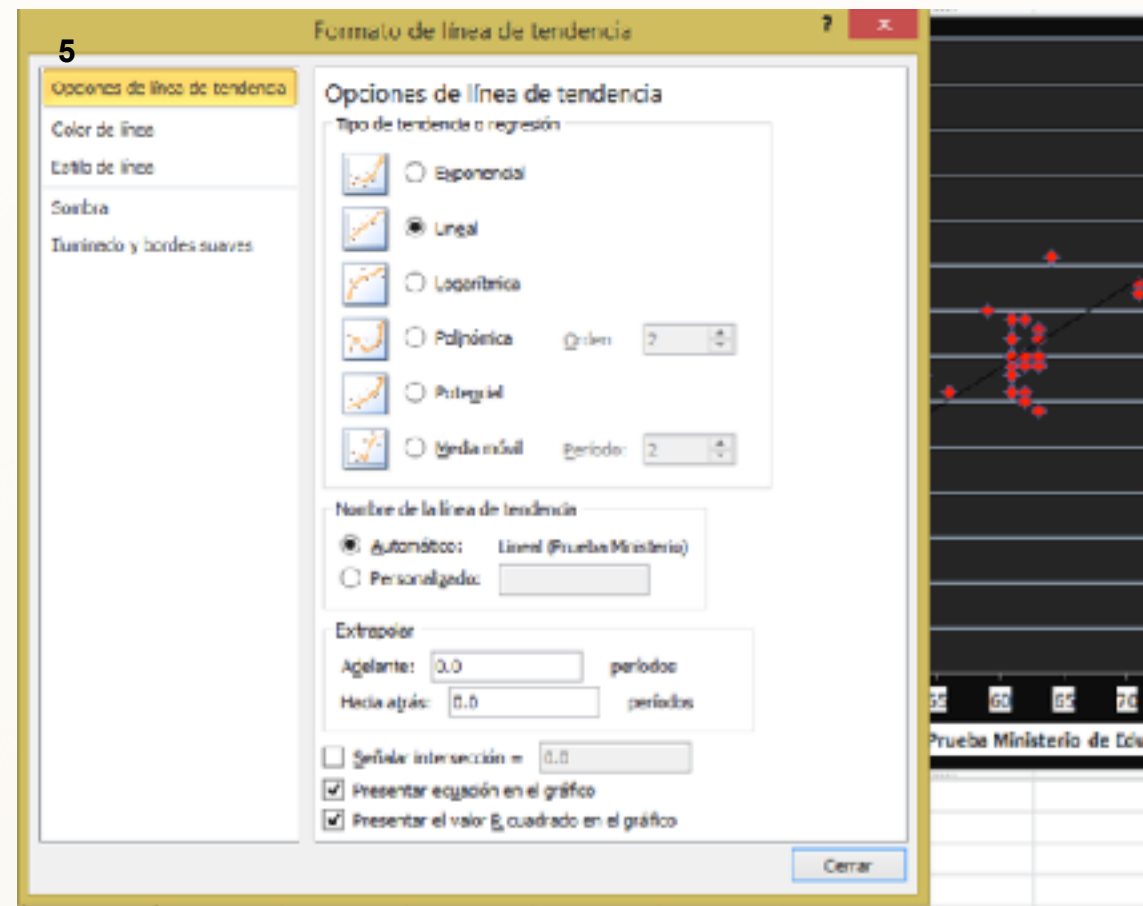
En este caso se modificó el fondo de la imagen, el fondo del área del diagrama, el color de la nube de puntos, se agregó un fondo a los ejes y a los títulos. Simplemente se personalizó la tabla sin afectar los resultados.

Ahora en Excel haremos clic derecho sobre algún punto de la nube y obtendremos un menú, allí debemos seleccionar “agregar línea de tendencia”.



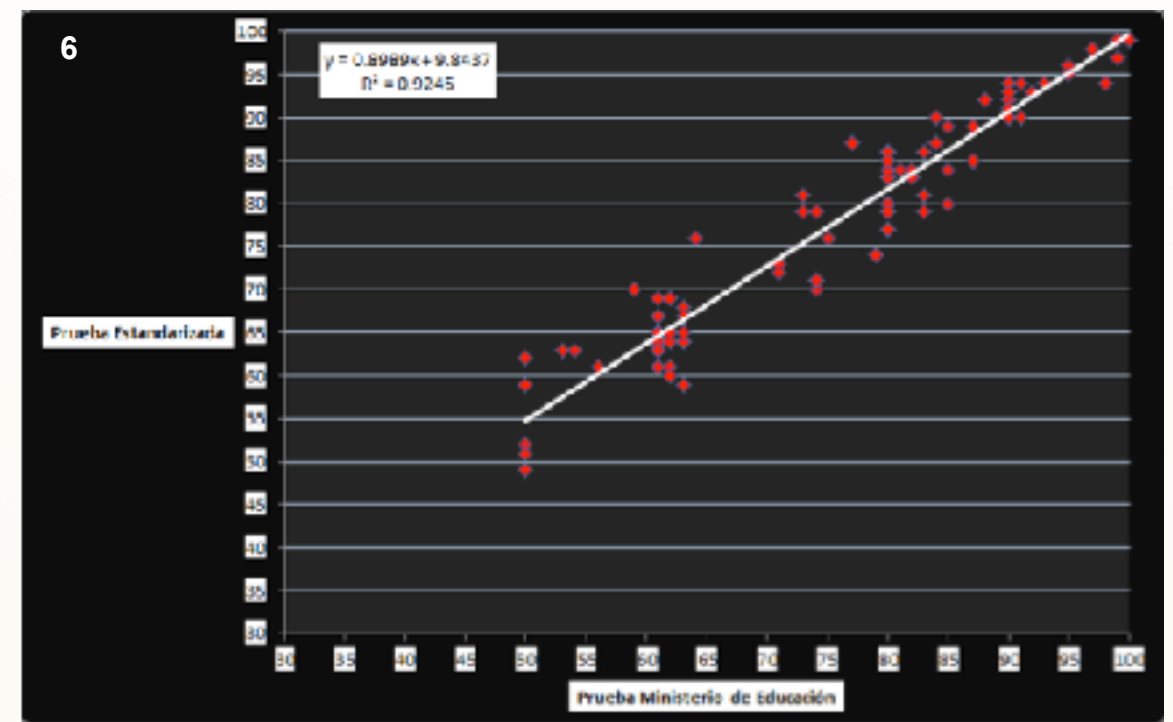
Agregar línea de tendencia

Luego de seleccionar la opción indicada en el círculo rojo, aparecerá un nuevo menú que nos permite seleccionar el tipo de línea de tendencia, en nuestro caso seleccionaremos siempre la opción “lineal”. Además, vamos a seleccionar los dos cuadros de abajo 1) “*presentar ecuación en gráfico*” y 2) “*presentar el valor R cuadrado en el gráfico*”.



Formato de línea de tendencia

Al hacer clic en la opción Cerrar, nuestro diagrama tendrá la fórmula de la regresión resuelta, el valor r^2 y la línea de tendencia marcada.



Línea de tendencia

Hasta aquí terminamos con los cálculos en Excel, ¿fácil verdad? Ahora necesitamos entender qué es lo que obtuvimos y cómo lo vamos a utilizar. Nos enfocaremos primero en la fórmula de la regresión.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

Y veamos ahora la fórmula que genera Excel...

7

$$y = 0.8989x + 9.8437$$

$$R^2 = 0.9245$$

Formula

Si hemos seguido el trabajo que estamos haciendo en este momento ya hemos entendido que lo que expresa la fórmula no es nada más que una forma de adivinar científicamente el valor de Y cuando tengamos un valor X, en nuestro caso queremos saber cuál será el punteo de una persona en la prueba del Ministerio a partir del punteo que obtiene en la prueba de la investigadora. Lo que generó Excel son los datos, los parámetros del modelo que necesitamos para hacer dichas predicciones. Así pues que ahora sabemos que:

8

$$\beta_0 = 9.8437$$

$$\beta_i = 0.8989$$

X_i = es el valor del cuál queremos predecir. Así que ahora hagamos preguntas.

Para un alumno que obtuvo 80 en la prueba de la investigadora ¿Cuál será el resultado en la prueba del Ministerio?

9

$$Y_i = \beta_0 + \beta_1 X_i$$

$$Y_{80} = 9.8437 + (0.8989 * 80)$$

$$Y_{80} = 9.8437 + 71.91$$

$$\mathbf{Y_{80} = 81.75}$$

Estos datos predicen que un alumno que saque 80 en la prueba de la investigadora obtendría 81.75 en la prueba del Ministerio. Es una prueba más de la correlación existente pero además nos permite hacer predicciones.

Ahora bien, nos vamos a fijar en la otra parte de la información que tenemos, el r^2 .

10

$$y = 0.8989x + 9.8437$$

$$R^2 = 0.9245$$

Excel interpreta la expresión en mayúsculas “R”, como la representación que ya conocemos del coeficiente de correlación “r”. Si elevamos ese coeficiente al cuadrado, lo que obtenemos es una medida que representa la cantidad de varianza que captura la variable independiente de la dependiente. Algunos autores se refieren al r^2 como la porción de Y que explica X, sin embargo hay que tener cautela al hablar de relaciones de causa y efecto, porque la regresión es un modelo matemático y para poder asegurar que existe causa y efecto se necesita diseñar un experimento controlado.



Normalmente el r^2 se usa como porcentaje, al multiplicarlo por 100, así que diríamos que para este estudio, 92% de la prueba del Mineduc puede ser explicada por el resultado en la prueba de la investigadora. Además podemos comprobarlo, si vemos el resultado de este ejercicio en el capítulo de correlación veremos que $r = .96$, si elevamos al cuadrado (o lo multiplicamos por sí mismo) obtenemos 0.92.



Con este ejercicio ficticio, podemos ver el poder de la estadística puesto en acción. Este ejercicio es una situación muy práctica y muy útil porque, si ya sabemos que las pruebas de la investigación y las del Ministerio están relacionadas significa que podemos evaluar a nuestros alumnos con la prueba de la investigación varios meses antes de la prueba del Ministerio y al hacer el análisis de regresión sabremos si tenemos que tomar medidas correctivas para preparar mejor a nuestros estudiantes.



c. Ejercicio en Excel

Retomaremos el ejercicio usado en la correlación del padre de familia que quiere apoyar dando libros a los padres de los estudiantes, antes de hacerlo deciden investigar si se relaciona la cantidad de libros en casa con las notas de Comunicación y Lenguaje.

Estos son los datos:

11

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
Cantidad de libros	41	38	57	0	49	30	40	66	38	38	40	7	53	68	53	41	27	0	54	33	31	5	40
Comunicación y Lenguaje	78	59	72	80	45	77	75	47	61	56	76	61	73	76	47	58	79	44	49	71	88	56	65

	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	A39	A40	A41	A42	A43	A44	A45	A46
Cantidad de libros	68	54	32	5	12	56	1	12	35	59	73	28	14	25	4	47	12	21	57	5	33	28	0
Comunicación y Lenguaje	61	72	66	64	48	65	47	43	73	52	90	54	55	57	55	79	82	74	57	75	83	93	42

Organización de datos para crear el diagrama de dispersión.

Así que organizaremos los datos para crear el diagrama de dispersión. Recordemos que al finalizar el diagrama necesitamos hacer click derecho sobre algún punto de la nube y solicitar la “línea de tendencia” además pediremos la ecuación y el r2.

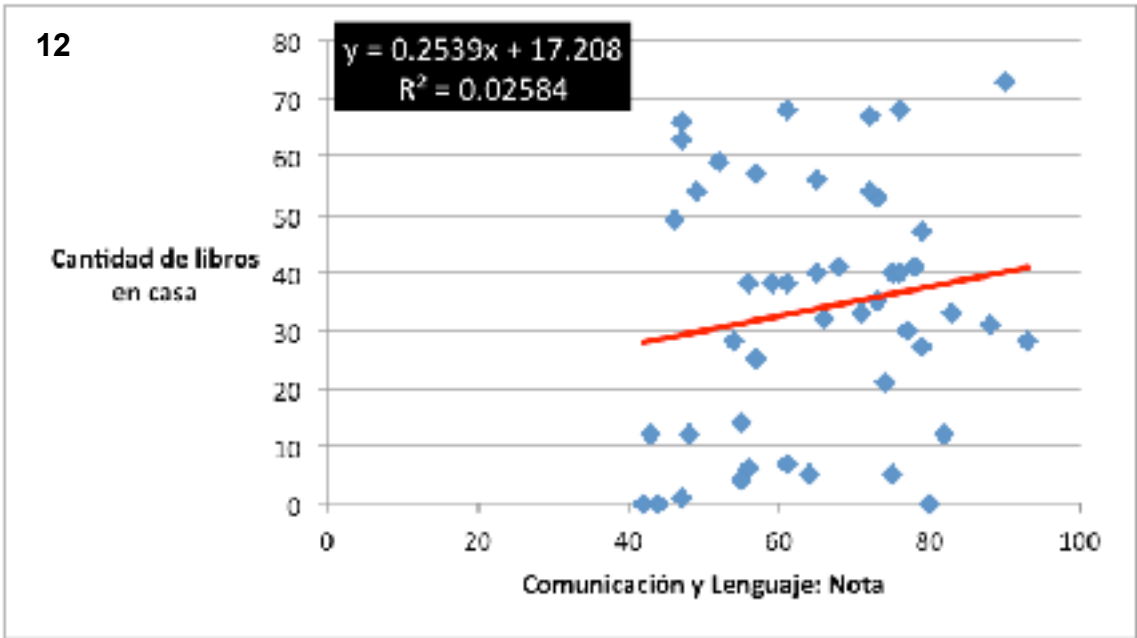


Diagrama de dispersión, Ejemplo de regresión.

Ahora bien, antes de hacer preguntas de predicción debemos tomar en cuenta el valor de r^2 en este caso 2.5% de la variación en la nota de Comunicación y Lenguaje está relacionada con la cantidad de libros en casa.

Este valor es muy pequeño. En educación, algunos expertos, los más flexibles, dicen que para considerar los resultados de una regresión es necesario que r^2 sea mayor del 18% y los más exigentes dicen que debe ser arriba de 25%.

Así que pensemos en qué significa que el r^2 sea reducido.

- Para iniciar, significa que la correlación también era baja, al parecer las variables están relacionadas de una forma leve.

- Si r^2 es pequeño, quiere decir que la variación de la variable dependiente (las notas) se debe a otras variables que no medimos, en nuestro ejemplo 2.5% se debe a la cantidad de libros (nuestra variable) entonces $100\% - 2.5\% = 97.5\%$ se debe a otras variables.
- Por último y muy importante, es que si usamos los datos “cantidad de libros” para predecir “notas en comunicación y lenguaje” es muy probable que nos equivoquemos.

Por lo tanto, cuando r^2 es menor al 18%, es mejor preguntarnos qué variables serán las que no estoy midiendo y que pueden estar relacionadas con las notas. En nuestro ejemplo pueden ser muchas, por ejemplo, el nivel educativo de los padres, el estado de nutrición del niño, si el niño trabaja por las tardes, la cantidad de horas de estudio y otras que podríamos pensar que afectan las notas.



Comprobación del capítulo XX: Principios de la regresión lineal

Si eres estudiante inscrito, ingresa en la
plataforma para resolver la

Comprobación del capítulo XX

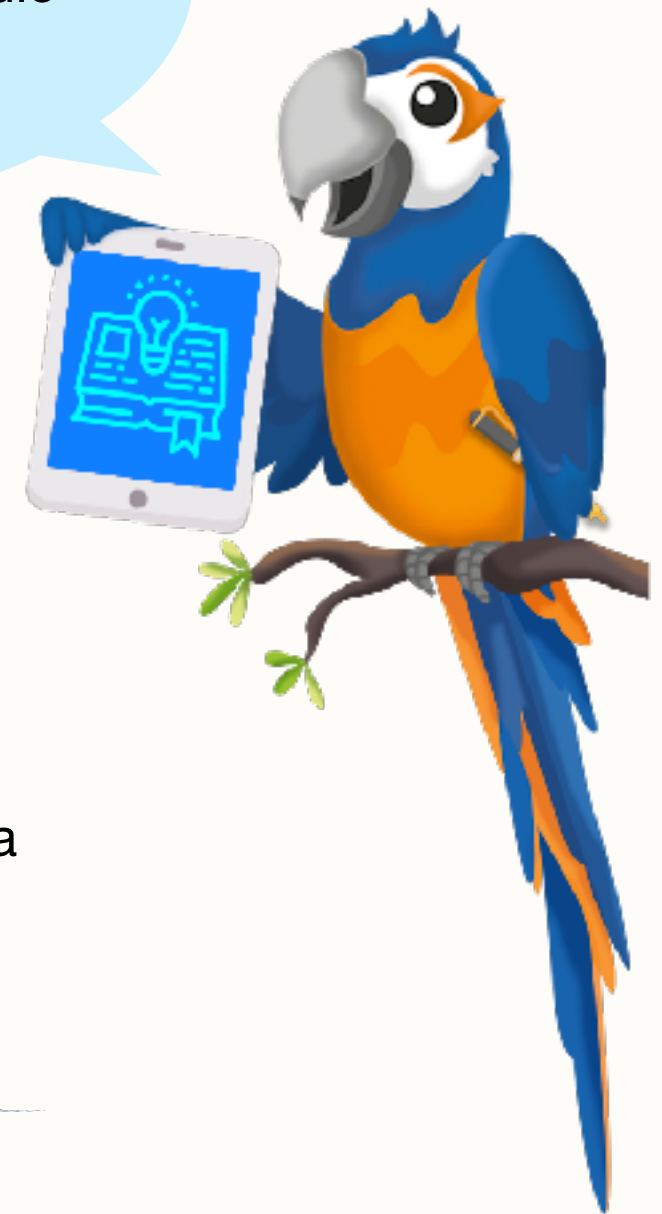
Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

secretario@uomac.net



¡Buen Trabajo!

Has completado el estudio
de tu curso



Si eres alumno registrado ingresa a la plataforma
para resolver la

Evaluación Corta de Aprendizaje (ECA)

Si tienes alguna dificultad, no dudes en
ponerte en contacto con UOMAC en:

informacion@uomac.net

Bibliografía



Niño Torres, Margarita. Metodología de la Investigación y Fundamentos de Estadística. Libro electrónico. Chile. 2011

Pérez Margarita. Estadística Descriptiva Ejercicios resueltos en Excel. Libro electrónico. México. 2013

Portilla Enrique Estadística Primer Curso. Mac Graw-Hill. México 1998

Richardson Karen. Statistics for the plain stupid. Libro electrónico. Estados Unidos. 2015

Ritchey, Ferris J. Estadística para las Ciencias Sociales. McGraw-Hill. México. Segunda Edición, 2006.

Webliografía



Documentos electrónicos

Amon Hortelano, J. (s.f.). EDUCATECA.
Recuperado el 31 de 07 de 2013, de
<http://bit.ly/2nK74IL>

Sociedad Andaluza de Educación Matemática
Thales, Facultad de Matemática. (s.f.).
SOCIEDAD ANDALUZA DE EDUCACIÓN
MATEMÁTICA THALES. Recuperado el 2013,
de Estadística Descriptiva

Universidad de Zaragoza. (s.f.). Portal
Educativo. Recuperado el 01 de Agosto de
2013, de Estadística: lecciones, apuntes,
trabajos y enlaces: [http://ciberconta.unizar.es/
docencia/estadistica/](http://ciberconta.unizar.es/docencia/estadistica/)

[http://www.ditutor.com/estadistica/
estadistica_descriptiva.html](http://www.ditutor.com/estadistica/estadistica_descriptiva.html)

[http://www.socialresearchmethods.net/kb/
statdesc.php](http://www.socialresearchmethods.net/kb/statdesc.php)

[https://www.khanacademy.org/math/probability/
descriptive-statistics](https://www.khanacademy.org/math/probability/descriptive-statistics)

[http://www.aulafacil.com/cursos/t675/ciencia/
estadisticas/estadisticas](http://www.aulafacil.com/cursos/t675/ciencia/estadisticas/estadisticas)

[http://www.uam.es/personal_pdi/economicas/
fphernan/EconometriaTII.pdf](http://www.uam.es/personal_pdi/economicas/fphernan/EconometriaTII.pdf)

[http://es.slideshare.net/hectorquintero/
presentacin-grfica-de-datos](http://es.slideshare.net/hectorquintero/presentacin-grfica-de-datos)

<https://www.youtube.com/watch?v=XQw6D5xMcTA>

<https://www.youtube.com/watch?v=fOWmk7ETxw>

http://www.vitutor.com/estadistica/descriptiva/a_5.html

http://www.vitutor.com/estadistica/descriptiva/a_4.html

www.youtube.com/watch?v=tIRgjy_xwl8

http://datateca.unad.edu.co/contenidos/208041/Modulo_EXE/leccin_5_tablas_y_grficos_expresin_visual_de_la_estadstica.html

<http://graficosestadisticos.blogspot.com/>

<https://www.youtube.com/watch?v=7wN5WSCZBXI>

Distribución Normal: <http://youtu.be/cfIM5TMTYSk>

Área bajo la Curva: <http://youtu.be/41BMX80Kirk>

https://www.youtube.com/watch?v=b_Oee84PrGg

http://www.ditutor.com/distribucion_normal/distribucion_normal.html

https://www.youtube.com/watch?v=b_Oee84PrGg

<https://www.youtube.com/watch?v=7wN5WSCZBXI>

<https://www.youtube.com/watch?v=9kRI-Zz-YCY>

<https://www.youtube.com/watch?v=41BMX80Kirk&feature=youtu.be>

http://www.uca.edu.sv/matematica/upload_w/file/REGRESION%20SIMPLE%20Y%20MULTIPLE.pdf

<https://www.youtube.com/watch?v=fNeXC8d5En8>

http://www.vitutor.com/estadistica/bi/recta_regresion.html

https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal

Anexo



X	Y			
Prueba Investigación	Prueba Ministerio	X2	Y2	XY
50	59	2500	3481	2950
50	62	2500	3844	3100
50	52	2500	2704	2600
50	51	2500	2601	2550
50	49	2500	2401	2450
53	63	2809	3969	3339
54	63	2916	3969	3402
56	61	3136	3721	3416
59	70	3481	4900	4130
61	69	3721	4761	4209
61	64	3721	4096	3904
61	63	3721	3969	3843
61	64	3721	4096	3904
61	65	3721	4225	3965
61	65	3721	4225	3965
61	65	3721	4225	3965
61	63	3721	3969	3843
61	61	3721	3721	3721
61	63	3721	3969	3843
61	67	3721	4489	4087
61	64	3721	4096	3904
61	63	3721	3969	3843
61	64	3721	4096	3904
61	65	3721	4225	3965
61	65	3721	4225	3965
61	65	3721	4225	3965
61	63	3721	3969	3843
61	61	3721	3721	3721
61	63	3721	3969	3843
61	67	3721	4489	4087
62	64	3844	4096	3968
62	65	3844	4225	4030
62	61	3844	3721	3782
62	69	3844	4761	4278
62	60	3844		3720
63	59	3969	3481	3717
63	64	3969	4096	4032
63	68	3969	4624	4284
63	67	3969	4489	4221
63	64	3969	4096	4032
63	65	3969	4225	4095
64	76	4096	5776	4864

71	73	5041	5329	5183
71	72	5041	5184	5112
73	81	5329	6561	5913
73	79	5329	6241	5767
74	70	5476	4900	5180
74	71	5476	5041	5254
74	79	5476	6241	5846
74	71	5476	5041	5254
75	76	5625	5776	5700
77	87	5929	7569	6699
79	74	6241	5476	5846
79	74	6241	5476	5846
80	80	6400	6400	6400
80	85	6400	7225	6800
80	80	6400	6400	6400
80	79	6400	6241	6320
80	83	6400	6889	6640
80	85	6400	7225	6800
80	84	6400	7056	6720
80	86	6400	7396	6880
80	79	6400	6241	6320
80	83	6400	6889	6640
80	85	6400	7225	6800
80	84	6400	7056	6720
80	86	6400	7396	6880
80	77	6400	5929	6160
80	85	6400	7225	6800
80	80	6400	6400	6400
80	79	6400	6241	6320
80	83	6400	6889	6640
80	85	6400	7225	6800
80	84	6400	7056	6720
80	86	6400	7396	6880
80	77	6400	5929	6160
80	85	6400	7225	6800
81	84	6561	7056	6804
82	83	6724	6889	6806
82	84	6724	7056	6888
83	79	6889	6241	6557
83	81	6889	6561	6723
83	86	6889	7396	7138
84	90	7056	8100	7560
84	87	7056	7569	7308
85	89	7225	7921	7565
85	80	7225	6400	6800
85	84	7225	7056	7140

87	85	7569	7225	7395
87	89	7569	7921	7743
88	92	7744	8464	8096
90	94	8100	8836	8460
90	91	8100	8281	8190
90	92	8100	8464	8280
90	91	8100	8281	8190
90	93	8100	8649	8370
90	90	8100	8100	8100
91	94	8281	8836	8554
91	90	8281	8100	8190
92	93	8464	8649	8556
93	94	8649	8836	8742
95	96	9025	9216	9120
95	95	9025	9025	9025
97	98	9409	9604	9506
98	94	9604	8836	9212
99	99	9801	9801	9801
99	97	9801	9409	9603
100	99	10000	9801	9900
8069	8326	615999	652302	633136

Anexo. Ejemplo de correlación