BI4SME

boosting business intelligence skills for SME growth

BI4SME – R2 – Materiales de Entrenamiento UNIDAD 1 - Matemáticas para BI

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

Co-funded by the European Union

Sommario

Public Licence

1. UNIDAD 1: Matemáticas para BI

1. Introducción

En general, **BI (Inteligencia de Negocios)** puede describirse como un conjunto de procesos que permiten a los profesionales utilizar los datos con la máxima productividad. BI es una combinación de varias soluciones analíticas y de informes que abarcan la minoría de datos, la visualización de datos, indicadores clave de rendimiento y más. Hoy en día, ni los estudiantes, ni los jóvenes candidatos a trabajos, ni los profesionales que ya están trabajando pueden tener éxito en el análisis de negocios sin utilizar las herramientas de matemáticas y estadísticas. La falta de experiencia en programación puede superarse con la ayuda de programas educativos y capacitaciones, pero las matemáticas y la lógica son necesarias para que los profesionales puedan ver el mundo más allá de los números, para poder describir y justificar los resultados obtenidos y tomar decisiones de gestión informadas.
Dado que el mundo de hoy se está desplazando cada vez más hacia uno basado en la gestión profesional y el análisis de datos, el desarrollo de habilidades cuantitativas para fundamentar las decisiones de gestión corporativa y construir estrategias de gestión en un mundo acelerado es relevante y necesario. Con un conocimiento básico de Microsoft Excel, comprenderás conceptos de big data y análisis estadístico.

2. Estadísticas descriptivas para entender los datos

Las **estadísticas descriptivas** son coeficientes informativos breves que resumen un conjunto de datos dado, que puede ser una representación de toda la población o una muestra de una población. Las estadísticas

descriptivas pintan una imagen de las propiedades de los datos. Las estadísticas descriptivas se dividen en medidas de tendencia central y medidas de variabilidad (dispersión). Las medidas de tendencia central incluyen la media, mediana y moda, mientras que las medidas de variabilidad incluyen la desviación estándar, varianza, variables mínimas y máximas, curtosis y sesgo.

Proporcionaremos instrucciones paso a paso para usar Excel para calcular estadísticas descriptivas para datos dados. También mostraremos cómo interpretar los resultados, determinar qué estadísticas son más aplicables a los datos dados, etc.

2.1. Estadísticas descriptivas vs. estadísticas inferenciales

Las estadísticas descriptivas tienen una función diferente a las estadísticas inferenciales, conjuntos de datos que se utilizan para tomar decisiones o aplicar características de un conjunto de datos a otro. Imagina otro ejemplo donde una empresa vende, por ejemplo, microondas. La empresa recopila datos como el recuento de ventas, cantidad promedio comprada por transacción y ventas promedio por día / por semana / por mes. Toda esta información es descriptiva, ya que cuenta una historia de lo que sucedió en el pasado. En este caso, no se está utilizando más allá de ser informativa. Supongamos que la misma empresa quiere lanzar un nuevo microondas. Recopila los mismos datos de ventas anteriores, pero elabora la información para hacer predicciones sobre lo que serán las ventas del nuevo microondas. El acto de utilizar estadísticas descriptivas y aplicar características a un conjunto de datos diferente convierte el conjunto de datos en estadísticas inferenciales (Figura 1). Ya no estamos simplemente resumiendo datos; los estamos usando para predecir lo que sucederá con un conjunto de datos completamente diferente (el nuevo microondas).

Descriptive Statistics	Inferential Statistics
It is used to describe the characteristics of either the sample or the population by using quantitative tools.	It is used to draw inferences about the population data from the sample data by making use of analytical tools.
Measures of central tendency and measures of dispersion are the most important types of descriptive statistics.	Hypothesis testing and regression analysis are the types of inferential statistics.
It is used to describe the characteristics of a known dataset.	It tries to make inferences about the population that goes beyond the known data.
Measures of descriptive statistics are mean, median, variance, range, etc.	Measures of inferential statistics are z test, f test, linear regression, ANOVA test, etc.

Figura 1. Principales diferencias entre estadísticas descriptivas e inferenciales (fuente abierta)

2.2. Entender las estadísticas descriptivas: la línea de fondo

Las **estadísticas descriptivas** son informativas y están destinadas a describir las características reales de un conjunto de datos. El propósito principal de las estadísticas descriptivas es proporcionar información sobre un conjunto de datos (Estadísticas Descriptivas).

Los tres tipos principales de estadísticas descriptivas son la distribución de frecuencia, la tendencia central y la variabilidad de un conjunto de datos (Figura 2). La distribución de frecuencia registra la frecuencia con que ocurren los datos, la tendencia central registra el punto central de distribución de los datos y la variabilidad de un conjunto de datos registra su grado de dispersión.

Descriptive Statistics	
Measures of Central Tendency	Measures of Dispersion
Mean	Range
Median	Standard Deviation
Mode	Quartile Deviation
	Variance
	Absolute Deviation

Figure 2. Figura 2. Elementos clave de las estadísticas descriptivas (fuente abierta)

Las medidas de tendencia central se centran en los valores promedio o intermedios de los conjuntos de datos, mientras que las medidas de variabilidad se centran en la dispersión de los datos. Estas dos medidas utilizan gráficos, tablas y discusiones generales para ayudar a comprender el significado de los datos analizados.

Las medidas de tendencia central describen la posición central de una distribución para un conjunto de datos. Los profesionales analizan la frecuencia de cada punto de datos en la distribución y lo describen utilizando la media, mediana o moda, que mide los patrones más comunes del conjunto de datos analizado.

Por ejemplo, la suma del siguiente conjunto de datos (5, 6, 10, 34) es 55, se puede obtener con combinaciones de Excel Alt+= o Fórmulas Autosuma Suma (Figura 3).

El promedio del siguiente conjunto de datos (5, 6, 10, 34) es 13.75 = (5+6+10+34) / 4, se puede obtener con la combinación de Fórmulas Autosuma Promedio en Excel (Figura 4).

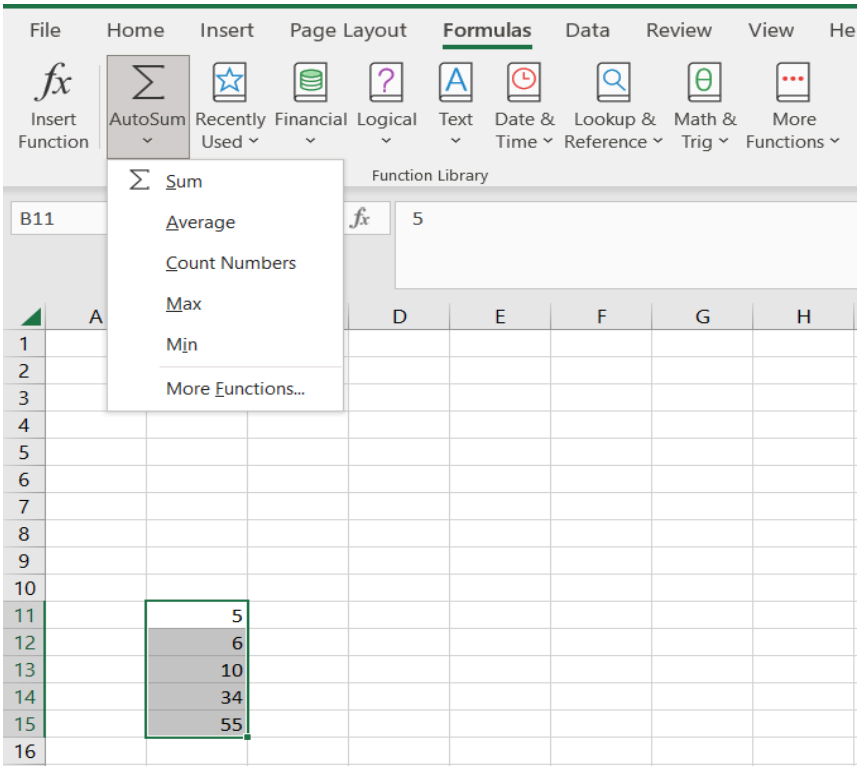


Figura 3. Pantalla de ejemplo 1

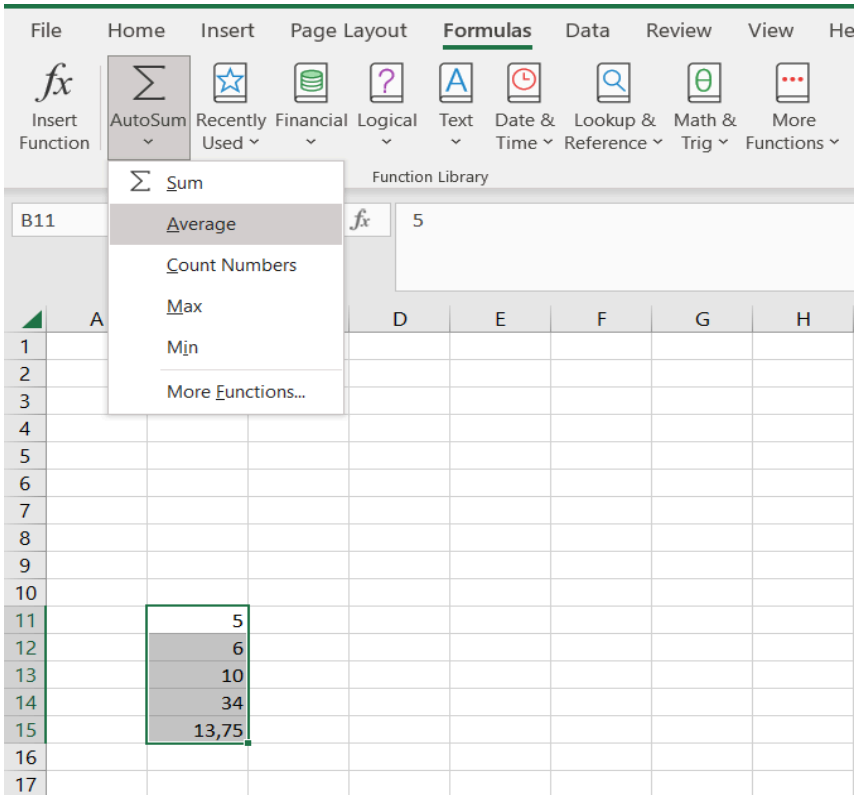


Figura 4. Pantalla de ejemplo 2

Por ejemplo, la empresa X vende algunos tipos de productos de ropa a diferentes valores (3700, 13300, 36000, 6700, 600, 17000, 8500, 4000) en $, podemos calcular la suma total (89800) y el promedio (11225) en $ (Figura 5).

L	M	N	O
	Product	Sales	
	Socks	3700	
	Shorts	13300	
	Tights	36000	
	Jerseys	6700	
	Caps	600	
	Helmets	17000	
	Vests	8500	
	Bib-Shorts	4000	
	Sum	89800	
	Average	11225	

Figura 5. Pantalla de ejemplo 3

Las personas utilizan las estadísticas descriptivas para reformular conocimientos cuantitativos difíciles de entender en un conjunto de datos grande en descripciones fáciles de digerir. El promedio de calificaciones de un estudiante (GPA), por ejemplo, proporciona una buena comprensión de las estadísticas descriptivas (por ejemplo, el promedio entre 75, 86, 92, 67, 88 es 81.6). La idea de un GPA es que toma puntos de datos de una amplia gama de exámenes, clases y calificaciones y los promedia para proporcionar una comprensión general del rendimiento académico de un estudiante. El GPA personal de un estudiante refleja su rendimiento académico promedio.

La correlación entre el gasto en educación y los resultados de las pruebas PISA se puede rastrear usando el ejemplo de los valores promedio de los resultados de la prueba de lectura PISA y los costos promedio para la educación en diferentes países (Figura 6). Como se puede ver, por ejemplo, Ucrania gasta muy poco. Sin embargo, el resultado es más alto de lo esperado para este nivel de costos. Singapur, Estonia, Hong Kong: un gran retorno por cada centavo gastado. Luxemburgo, Brunei y Qatar: gran gasto

para resultados comparables con países que gastan significativamente menos (desperdicio de dinero).

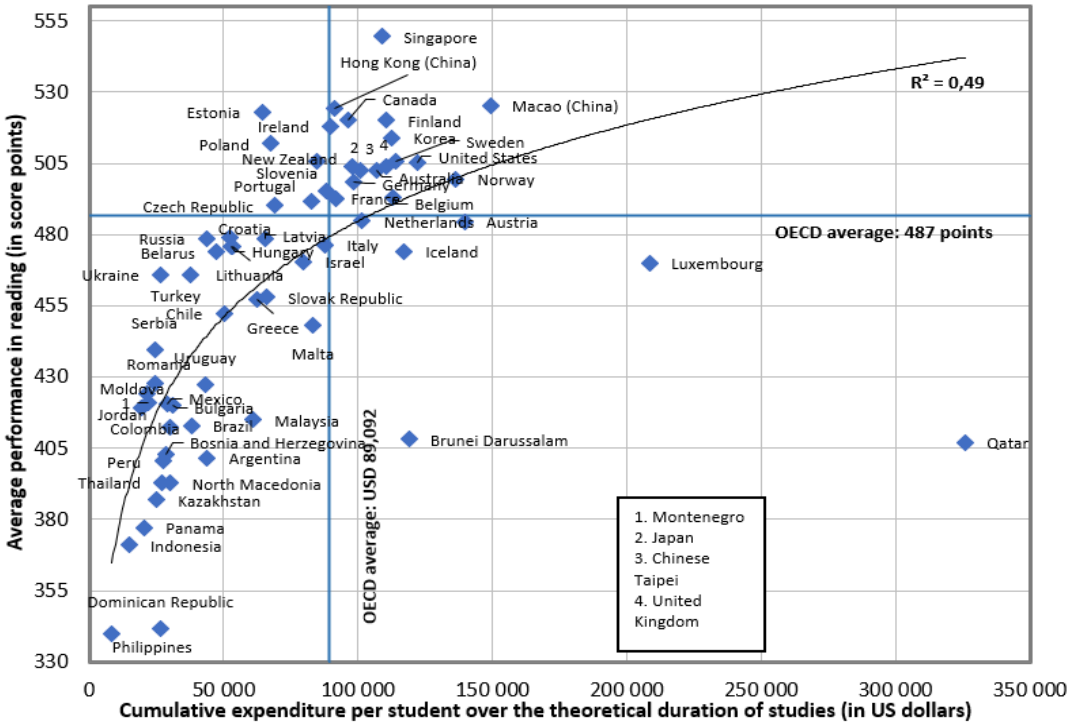


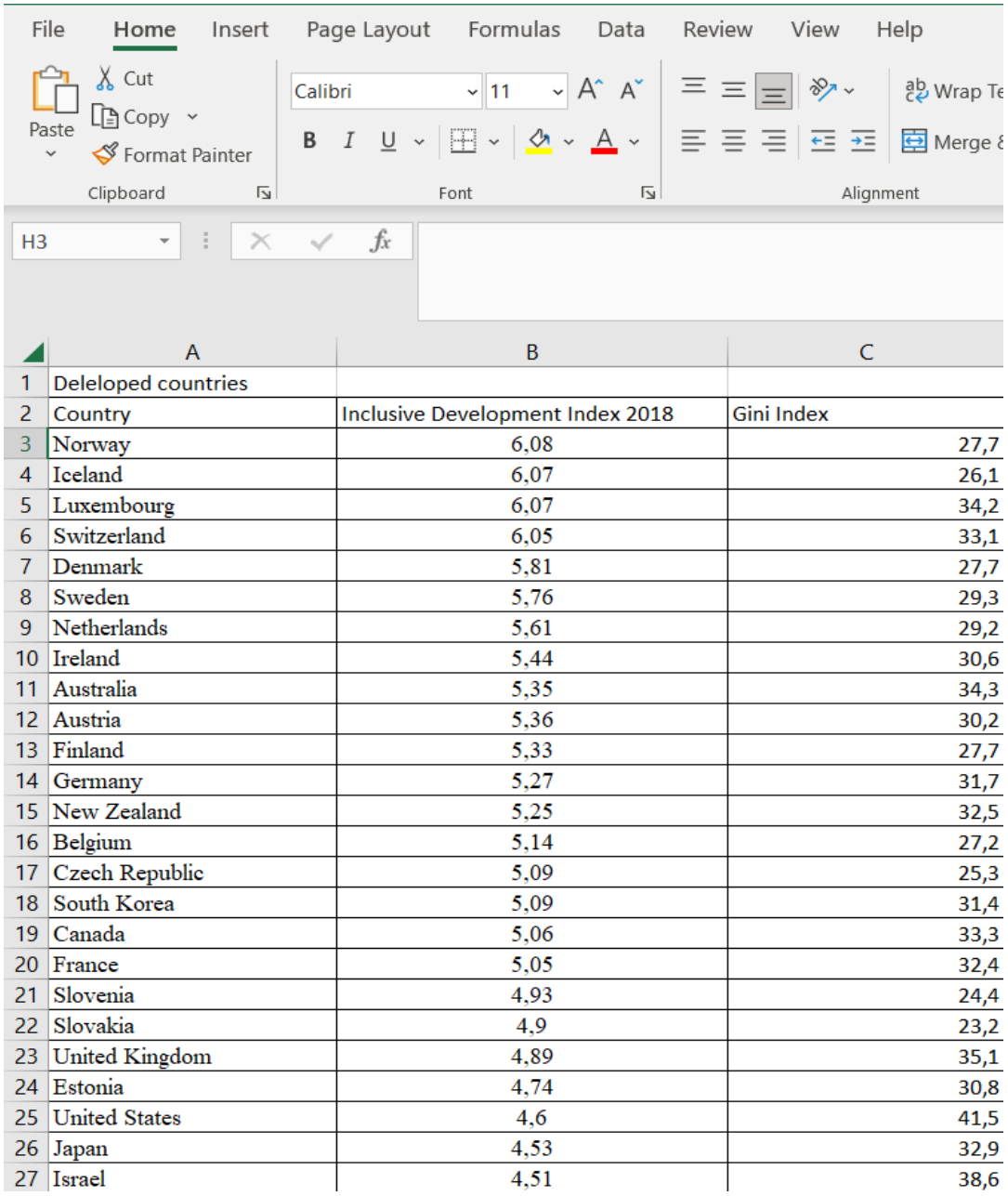
Figura 6. Correlación entre dos variables (fuente abierta)

2.3. Tablas dinámicas

Si te enfrentas a una montaña de datos en Excel y necesitas consolidarlos en un formato utilizable para que puedan filtrarse, agruparse y ordenarse fácilmente, definitivamente necesitas trabajar con tablas dinámicas.

Las tablas dinámicas son parte de la mayoría de los paquetes de software de hojas de cálculo o inteligencia empresarial. Pueden usarse con cualquier tipo de datos que se almacenen en columnas, pero son extremadamente útiles cuando los datos son tan grandes que superan lo que los filtros normales de las hojas de cálculo pueden manejar. La tabla dinámica, una vez configurada, puede usarse para agrupar, ordenar y resumir datos utilizando funciones matemáticas básicas. La tabla dinámica o tabla de resumen puede alterarse simplemente arrastrando y soltando campos gráficamente.

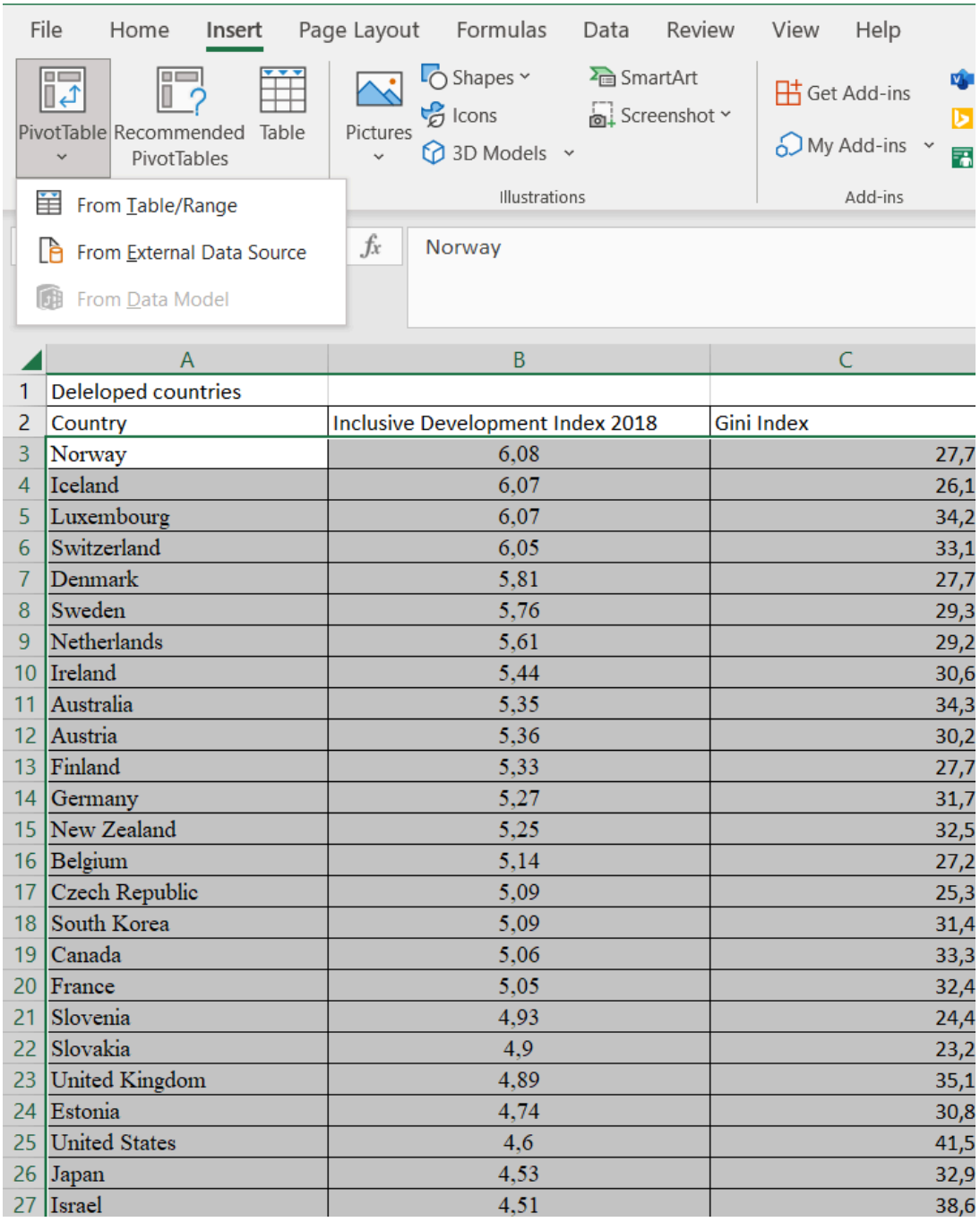
Primero necesitamos una fuente de datos que esté configurada en columnas. Por ejemplo, los datos muestran las estadísticas del IDI (Índice de Desarrollo Inclusivo) y el Índice de Gini de diferentes países para 2018 (Figura 7).



	A	B	C
1	Deleloped countries		
2	Country	Inclusive Development Index 2018	Gini Index
3	Norway	6,08	27,7
4	Iceland	6,07	26,1
5	Luxembourg	6,07	34,2
6	Switzerland	6,05	33,1
7	Denmark	5,81	27,7
8	Sweden	5,76	29,3
9	Netherlands	5,61	29,2
10	Ireland	5,44	30,6
11	Australia	5,35	34,3
12	Austria	5,36	30,2
13	Finland	5,33	27,7
14	Germany	5,27	31,7
15	New Zealand	5,25	32,5
16	Belgium	5,14	27,2
17	Czech Republic	5,09	25,3
18	South Korea	5,09	31,4
19	Canada	5,06	33,3
20	France	5,05	32,4
21	Slovenia	4,93	24,4
22	Slovakia	4,9	23,2
23	United Kingdom	4,89	35,1
24	Estonia	4,74	30,8
25	United States	4,6	41,5
26	Japan	4,53	32,9
27	Israel	4,51	38,6

Figura 7. Pantalla de ejemplo 4

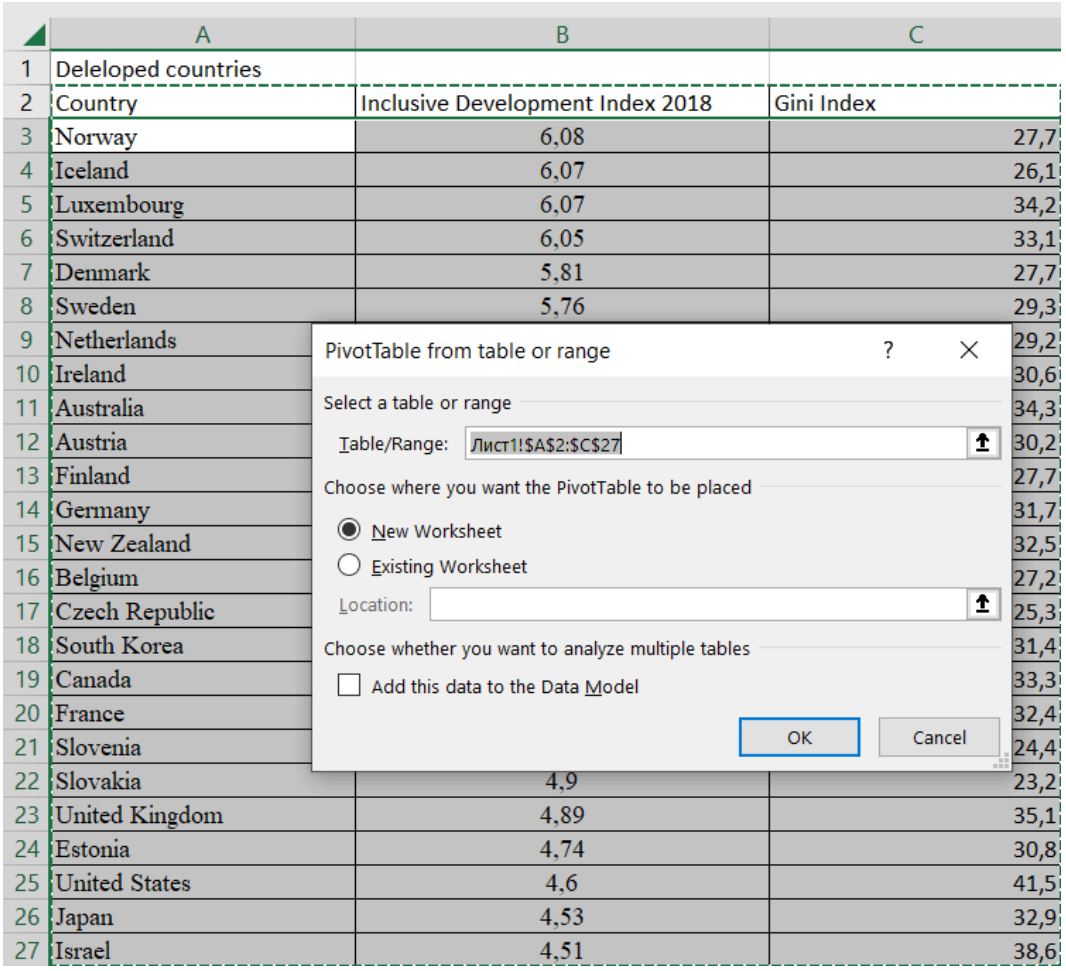
Una vez que tengamos nuestros datos en el formato adecuado, podemos hacer una tabla dinámica. Para hacer eso, resalta las columnas y selecciona la pestaña de insertar en la cinta (Figura 8).



	A	B	C
1	Deleloped countries		
2	Country	Inclusive Development Index 2018	Gini Index
3	Norway	6,08	27,7
4	Iceland	6,07	26,1
5	Luxembourg	6,07	34,2
6	Switzerland	6,05	33,1
7	Denmark	5,81	27,7
8	Sweden	5,76	29,3
9	Netherlands	5,61	29,2
10	Ireland	5,44	30,6
11	Australia	5,35	34,3
12	Austria	5,36	30,2
13	Finland	5,33	27,7
14	Germany	5,27	31,7
15	New Zealand	5,25	32,5
16	Belgium	5,14	27,2
17	Czech Republic	5,09	25,3
18	South Korea	5,09	31,4
19	Canada	5,06	33,3
20	France	5,05	32,4
21	Slovenia	4,93	24,4
22	Slovakia	4,9	23,2
23	United Kingdom	4,89	35,1
24	Estonia	4,74	30,8
25	United States	4,6	41,5
26	Japan	4,53	32,9
27	Israel	4,51	38,6

Figura 8. Pantalla de ejemplo 5

Una vez hecho, verás una pantalla como la mostrada a continuación (Figura 9).

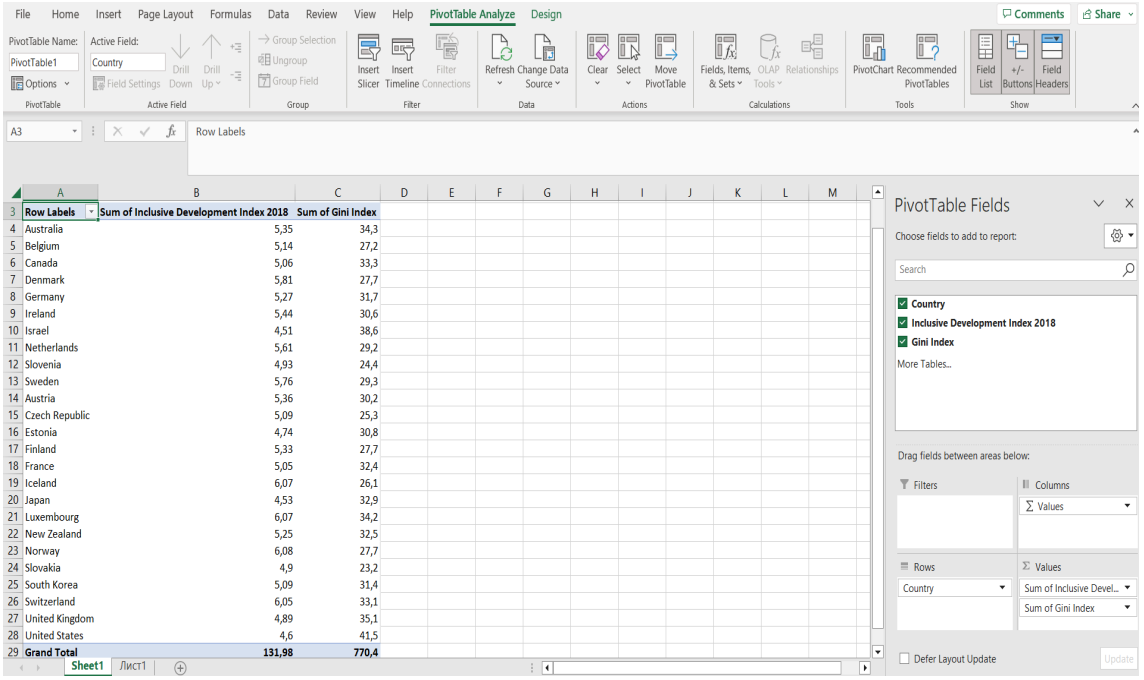


	A	B	C
1	Deleloped countries		
2	Country	Inclusive Development Index 2018	Gini Index
3	Norway	6,08	27,7
4	Iceland	6,07	26,1
5	Luxembourg	6,07	34,2
6	Switzerland	6,05	33,1
7	Denmark	5,81	27,7
8	Sweden	5,76	29,3
9	Netherlands		29,2
10	Ireland		30,6
11	Australia		34,3
12	Austria		30,2
13	Finland		27,7
14	Germany		31,7
15	New Zealand		32,5
16	Belgium		27,2
17	Czech Republic		25,3
18	South Korea		31,4
19	Canada		33,3
20	France		32,4
21	Slovenia		24,4
22	Slovakia	4,9	23,2
23	United Kingdom	4,89	35,1
24	Estonia	4,74	30,8
25	United States	4,6	41,5
26	Japan	4,53	32,9
27	Israel	4,51	38,6

Figura 9. Pantalla de ejemplo 6

Luego, Excel mostrará los datos que identificaste y preguntará si deseas crear una tabla dinámica con los datos seleccionados. Haz clic en OK. Luego verás una nueva hoja con una tabla dinámica en la hoja y una lista de campos de tabla dinámica a la derecha.

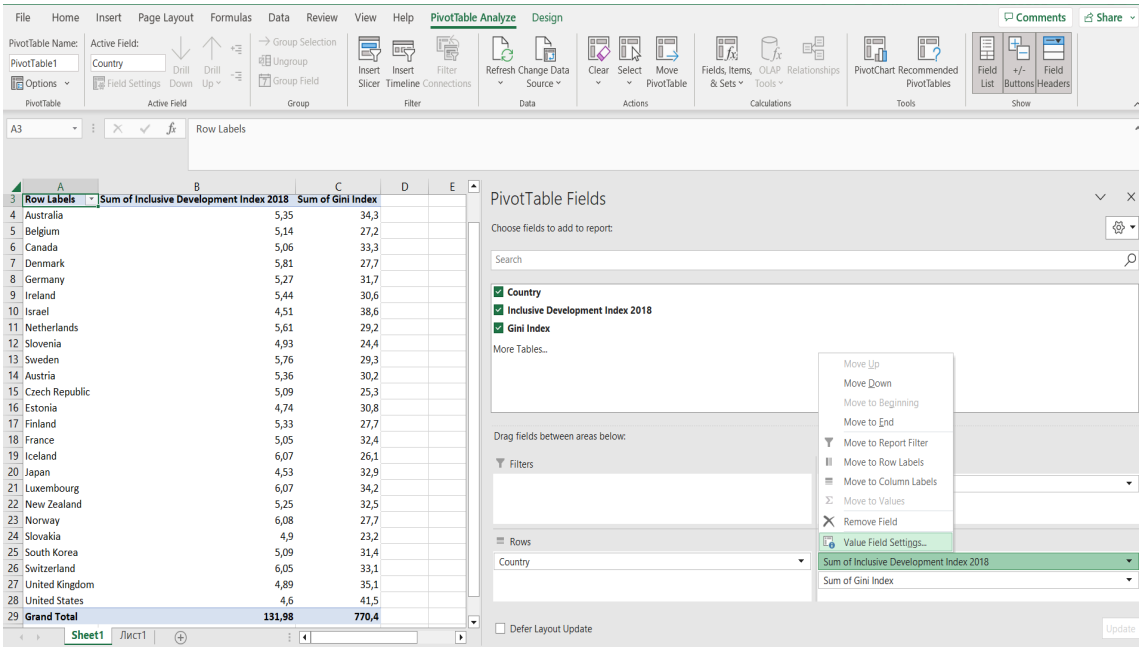
Ahora debemos decidir cómo queremos mostrar nuestros datos. Si revisas los datos proporcionados, tiene columnas que son descriptores de la fila de datos, a saber (país, IDI, Índice de Gini). Luego, los datos tienen columnas de estadísticas o información (Figura 10).



Row Labels	Sum of Inclusive Development Index 2018	Sum of Gini Index
Australia	5,35	34,3
Belgium	5,14	27,2
Canada	5,06	33,3
Denmark	5,81	27,7
Germany	5,27	31,7
Ireland	5,44	30,6
Israel	4,51	38,6
Netherlands	5,61	29,2
Slovenia	4,93	24,4
Sweden	5,76	29,3
Austria	5,36	30,2
Czech Republic	5,09	25,3
Estonia	4,74	30,8
Finland	5,33	27,7
France	5,05	32,4
Iceland	6,07	26,1
Japan	4,53	32,9
Luxembourg	6,07	34,2
New Zealand	5,25	32,5
Norway	6,08	27,7
Slovakia	4,9	23,2
South Korea	5,09	31,4
Switzerland	6,05	33,1
United Kingdom	4,89	35,1
United States	4,6	41,5
Grand Total	131,98	770,4

Figura 10. Pantalla de ejemplo 7

Para que nuestros cálculos tengan sentido, necesitamos cambiar Suma a Promedio para cada índice. Así que haz clic en la flecha desplegable en el lado derecho y selecciona "configuración del campo de valor" (Figuras 11-12).



Row Labels	Sum of Inclusive Development Index 2018	Sum of Gini Index
Australia	5,35	34,3
Belgium	5,14	27,2
Canada	5,06	33,3
Denmark	5,81	27,7
Germany	5,27	31,7
Ireland	5,44	30,6
Israel	4,51	38,6
Netherlands	5,61	29,2
Slovenia	4,93	24,4
Sweden	5,76	29,3
Austria	5,36	30,2
Czech Republic	5,09	25,3
Estonia	4,74	30,8
Finland	5,33	27,7
France	5,05	32,4
Iceland	6,07	26,1
Japan	4,53	32,9
Luxembourg	6,07	34,2
New Zealand	5,25	32,5
Norway	6,08	27,7
Slovakia	4,9	23,2
South Korea	5,09	31,4
Switzerland	6,05	33,1
United Kingdom	4,89	35,1
United States	4,6	41,5
Grand Total	131,98	770,4

Figura 11. Pantalla de ejemplo 8

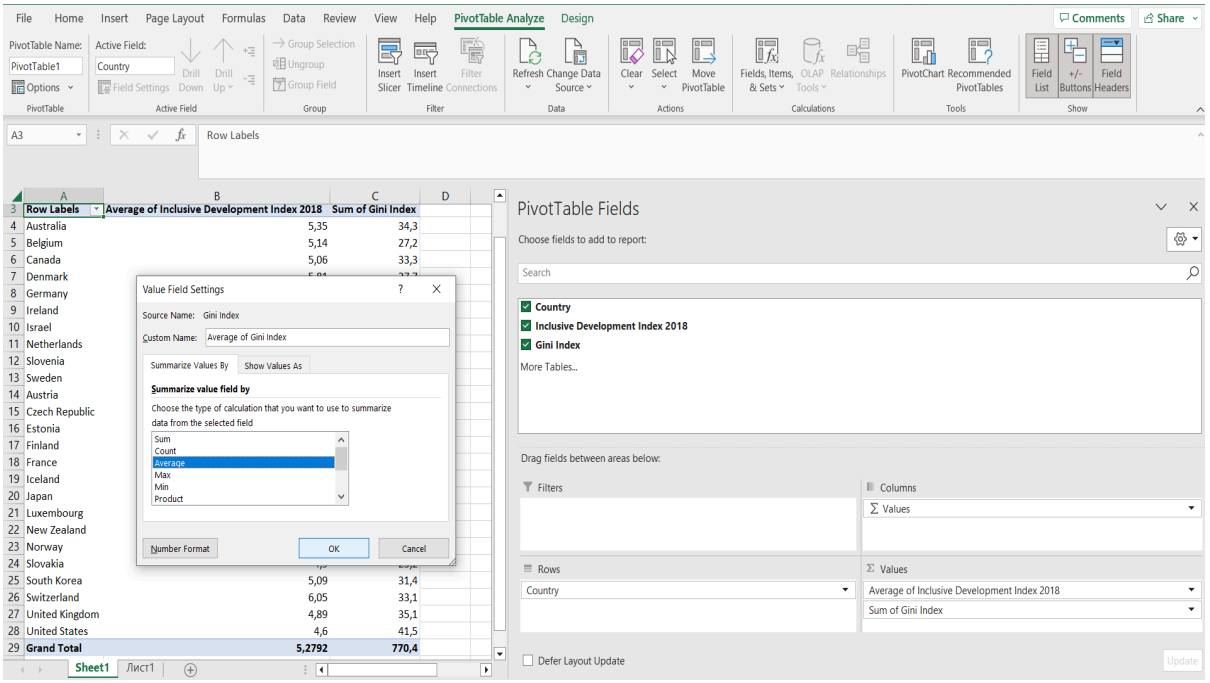


Figura 12. Pantalla de ejemplo 9

Desde aquí podemos graficar nuestros datos. Selecciona Opciones -> Gráfico dinámico y luego selecciona Columna -> Gráfico de columnas y presiona "OK" (Figuras 13-14).

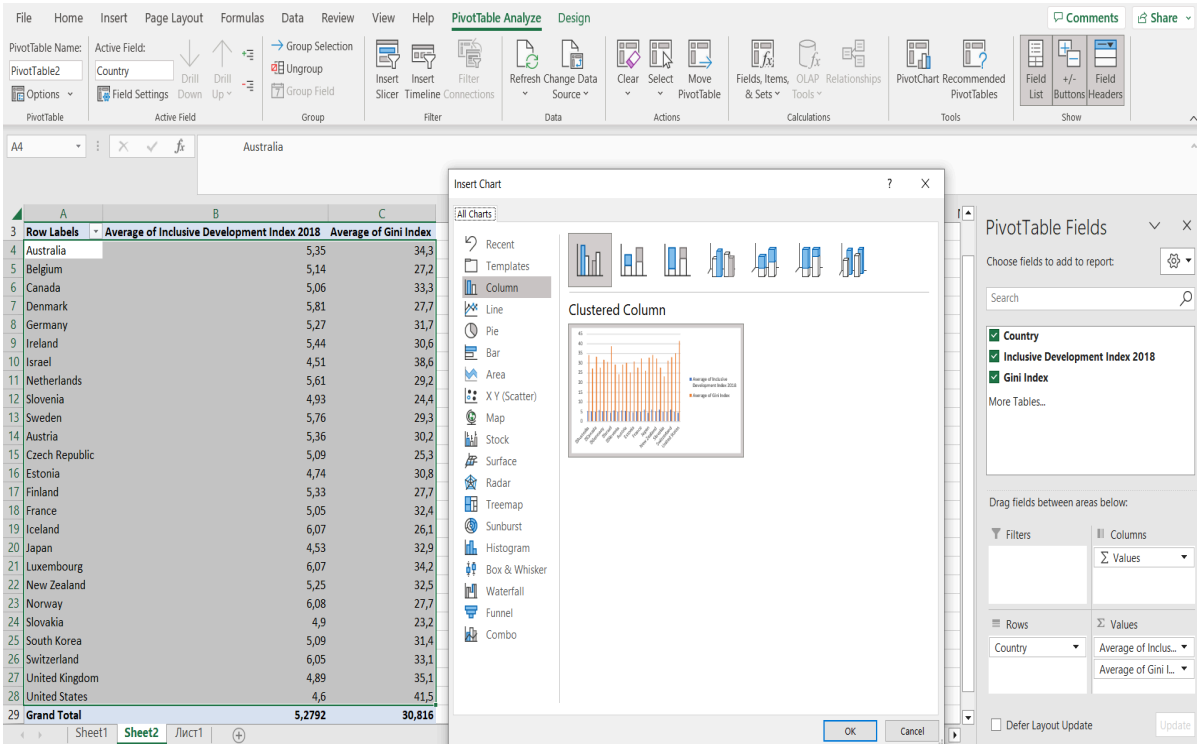
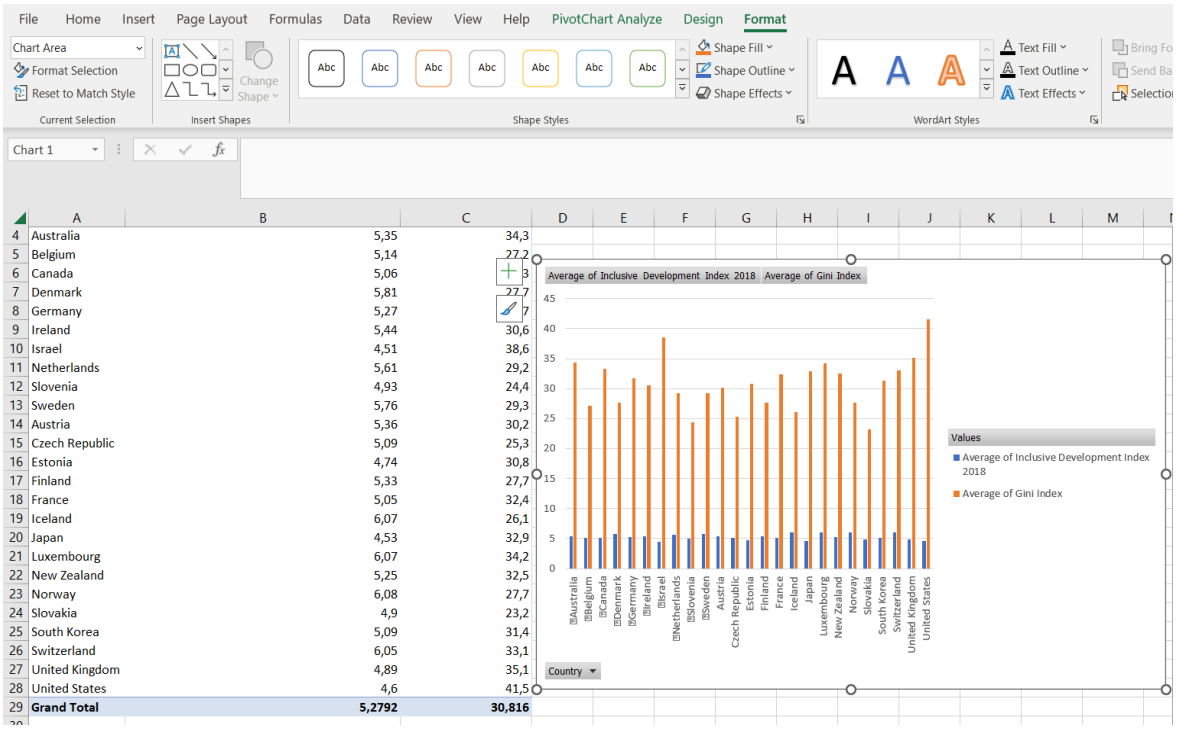


Figura 13. Pantalla de ejemplo 10



	A	B	C
4	Australia	5,35	34,3
5	Belgium	5,14	27,2
6	Canada	5,06	3
7	Denmark	5,81	27,7
8	Germany	5,27	7
9	Ireland	5,44	30,6
10	Israel	4,51	38,6
11	Netherlands	5,61	29,2
12	Slovenia	4,93	24,4
13	Sweden	5,76	29,3
14	Austria	5,36	30,2
15	Czech Republic	5,09	25,3
16	Estonia	4,74	30,8
17	Finland	5,33	27,7
18	France	5,05	32,4
19	Iceland	6,07	26,1
20	Japan	4,53	32,9
21	Luxembourg	6,07	34,2
22	New Zealand	5,25	32,5
23	Norway	6,08	27,7
24	Slovakia	4,9	23,2
25	South Korea	5,09	31,4
26	Switzerland	6,05	33,1
27	United Kingdom	4,89	35,1
28	United States	4,6	41,5
29	**Grand Total**	**5,2792**	**30,816**

Figura 14. Pantalla de ejemplo 11

También podemos agregar etiquetas de datos al gráfico final (Figura 15).

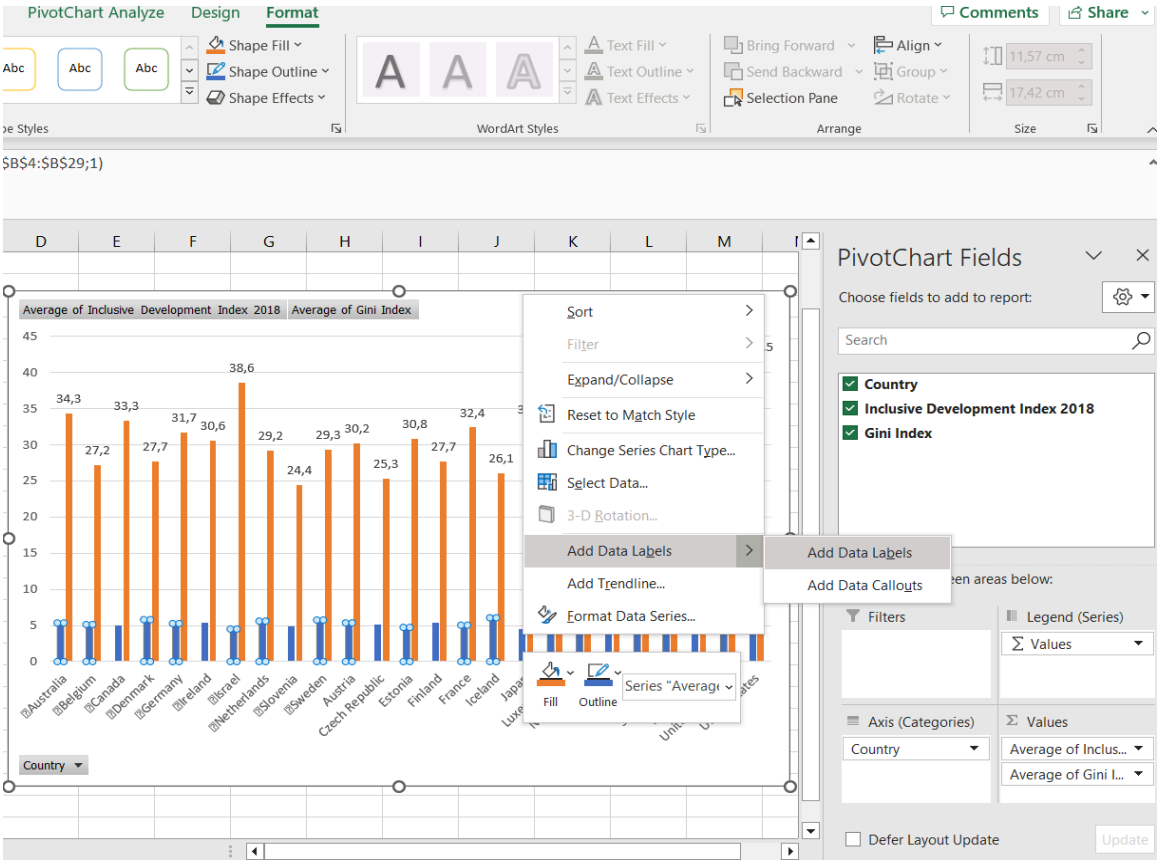


Figura 15. Pantalla de ejemplo 12

3. Medidas de frecuencia y tendencia central

La **medición de frecuencia** es un método estadístico utilizado para medir la cantidad de veces que ocurre un evento o resultado particular en cada muestra. La medición de frecuencia es el proceso de contar el número de ocurrencias de un valor particular o conjunto de valores en cada conjunto de datos. Se utiliza para medir la frecuencia relativa de un valor particular o conjunto de valores en cada conjunto de datos. La medición de frecuencia se utiliza en una variedad de campos, incluidos estadísticas, probabilidad y análisis de datos. La medición de frecuencia también puede utilizarse para medir la frecuencia de ciertos eventos u ocurrencias en cada momento. Por ejemplo, puede utilizarse para medir la frecuencia de terremotos en cada región durante un cierto período de tiempo (Medición de Frecuencia).

La medición de frecuencia es esencialmente una forma de determinar con qué frecuencia ocurre algo dentro de un período de tiempo dado. Esto podría ser cualquier cosa, desde el número de conteos por minuto, el número de ciclos por segundo en una medición de audio o electrónica, o el número de revoluciones de la rueda por unidad de tiempo en una medición de velocidad. Para medir la frecuencia, se necesitan tanto un temporizador como un contador, con el temporizador midiendo el tiempo de referencia y el contador contando el número de eventos dentro de ese tiempo (Figura 16).

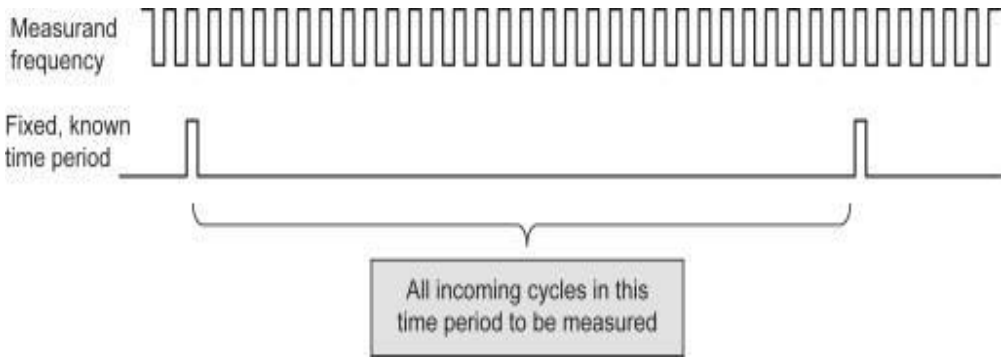


Figura 16. Explicación gráfica de la frecuencia (fuente abierta)

Las frecuencias de las variables se refieren a la cantidad de veces que un valor particular aparece en un conjunto de datos. Por ejemplo, si una variable tiene una frecuencia de 7, significa que el valor aparece siete veces en el conjunto de datos.

Las frecuencias acumuladas son útiles si se requiere información más detallada sobre un conjunto de datos. Pueden usarse para encontrar una mediana y un rango intercuartílico.

3.1. Calcular la media, la mediana y la moda a partir de la distribución de frecuencias

Si, por ejemplo, 37 personas trabajan en un supermercado, pueden recibir tasas diarias de $70, $110, $80, $90, $110, $70, $80, $100, $100, $90, $70, $70, $80, $70, $80, $80, $90, $80, $100, $100, $80, $80, $90, $80, $80, $100, $100, $100, $100, $90, $90, $90, $90, $90, $90, $90, $110.

La distribución de frecuencias puede demostrarse en una tabla donde los datos son un salario diario y la frecuencia muestra cuántas veces ocurre cada pago (Figura 17).

Data	Frequency	Cummulative frequency
70	5	5
80	10	15
90	11	26
100	8	34
110	3	37

Figura 17. Pantalla de ejemplo 13

La cantidad que se debe pagar por el supermercado a todo su personal por un día de trabajo es $3270. El nivel promedio de salario del personal por día es $884.

Media = Suma / N = [(705) + (8010) + (9011) + (1008) + (110*3)] / 37 = $884.

Mediana = Valor medio = $90. Moda = Más frecuente = $90.

Rango = Número más grande - Número más pequeño = 110 - 70 = $40 (Figura 18).

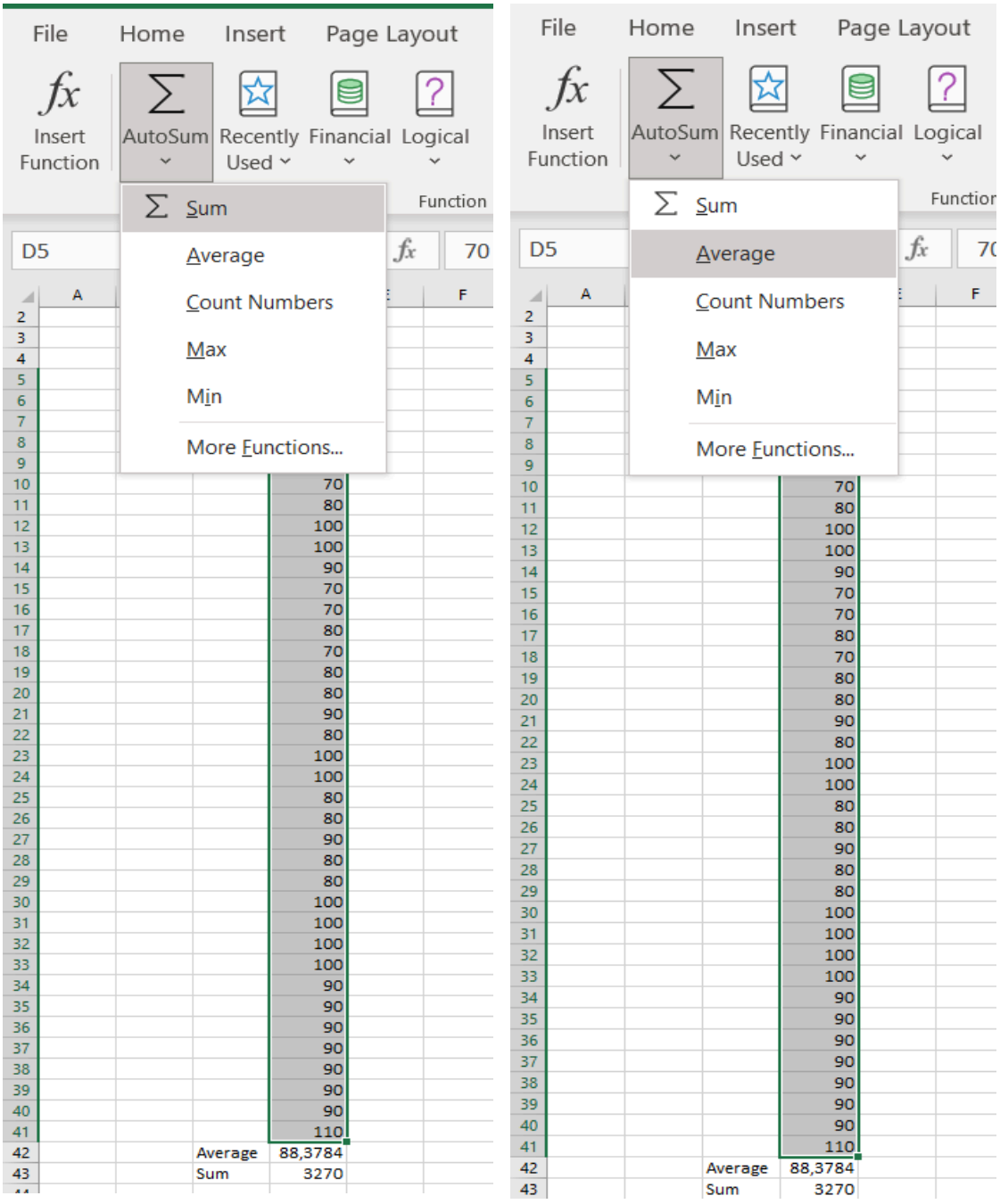


Figura 18. Pantalla de ejemplo 14

4. Medidas de dispersion

Las **medidas de dispersión** son medidas estadísticas que indican cuán disperso está un conjunto de datos. Proporcionan información sobre la

variabilidad de los datos y pueden usarse para comparar diferentes conjuntos de datos.

En estadística, la dispersión es el grado en que una distribución se estira o se comprime. Los métricos comunes de dispersión estadística son la varianza, la desviación estándar y el rango intercuartílico. Por ejemplo, cuando la varianza de los datos en un conjunto es alta, los datos están ampliamente dispersos. Por el contrario, cuando la varianza es baja, los datos en el conjunto están agrupados. La dispersión se contrasta con la ubicación o la tendencia central (Manikandan 2011).

La dispersión es la distribución de los puntos de datos alrededor de la media. Un ejemplo de dispersión puede ser el rango de alturas de las personas en un aula. La altura media de la clase puede ser de 5 pies, pero el rango de alturas puede ser de 4 a 6 pies.

Los dos polígonos de frecuencia dibujados en el gráfico a continuación muestran muestras que tienen la misma media, pero los datos en uno están mucho más dispersos que en el otro (Figura 19).

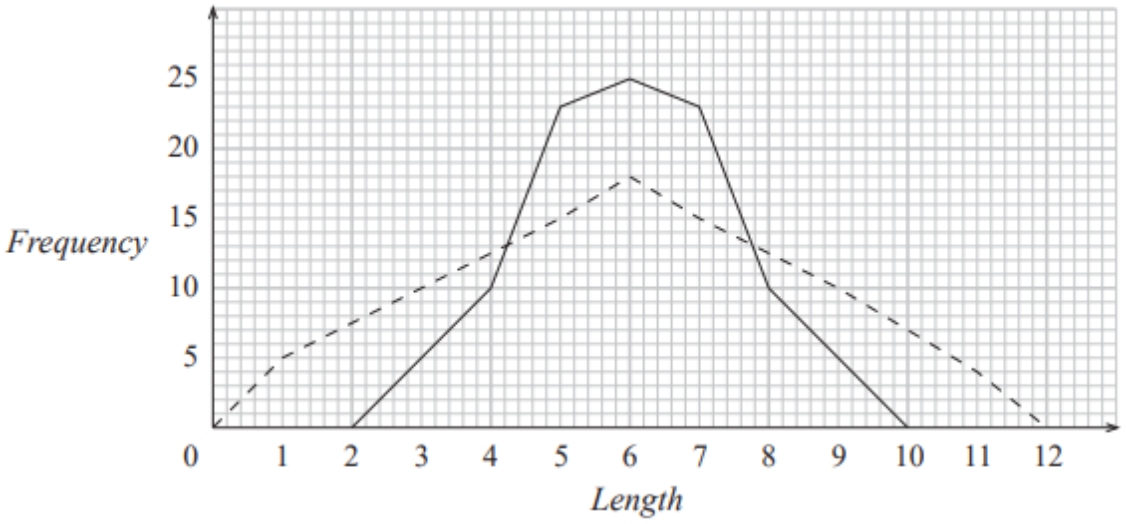


Figura 19. Ejemplos de frecuencia (fuente abierta)

El rango (valor más alto - valor más bajo) proporciona una medida simple de cuán dispersos están los datos.

La desviación estándar (s.d.) es una medida mucho más útil y se da por la fórmula (1):

$$\text{s.d.} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad (1)$$

donde x_i representa cada punto de datos (x_1, x_2, ..., x_n), $\bar{x}$ es la media, n es el numero de valores.

Entonces, la desviación estándar es la medida de la dispersión de los datos. Cuanto mayor sea su valor, más dispersos estarán los datos. Esto se ilustra con los dos polígonos de frecuencia mostrados anteriormente. Aunque ambos conjuntos de datos tienen la misma media, los datos representados por el polígono de frecuencia "punteado" tendrán una desviación estándar mayor que los otros.

4.1. Demostración: medidas de la tendencia central

La tendencia central en Excel se puede calcular con media, mediana y moda.. La media puede calcularse utilizando MEDIA= promedio(D5:D41) = 88.37 (Figura 20).

La mediana puede calcularse utilizando MEDIANA = mediana(D5:D41) = 90 (Figura 21).

La moda puede calcularse utilizando MODA = moda(D5:D41) = 90 (Figura 22).

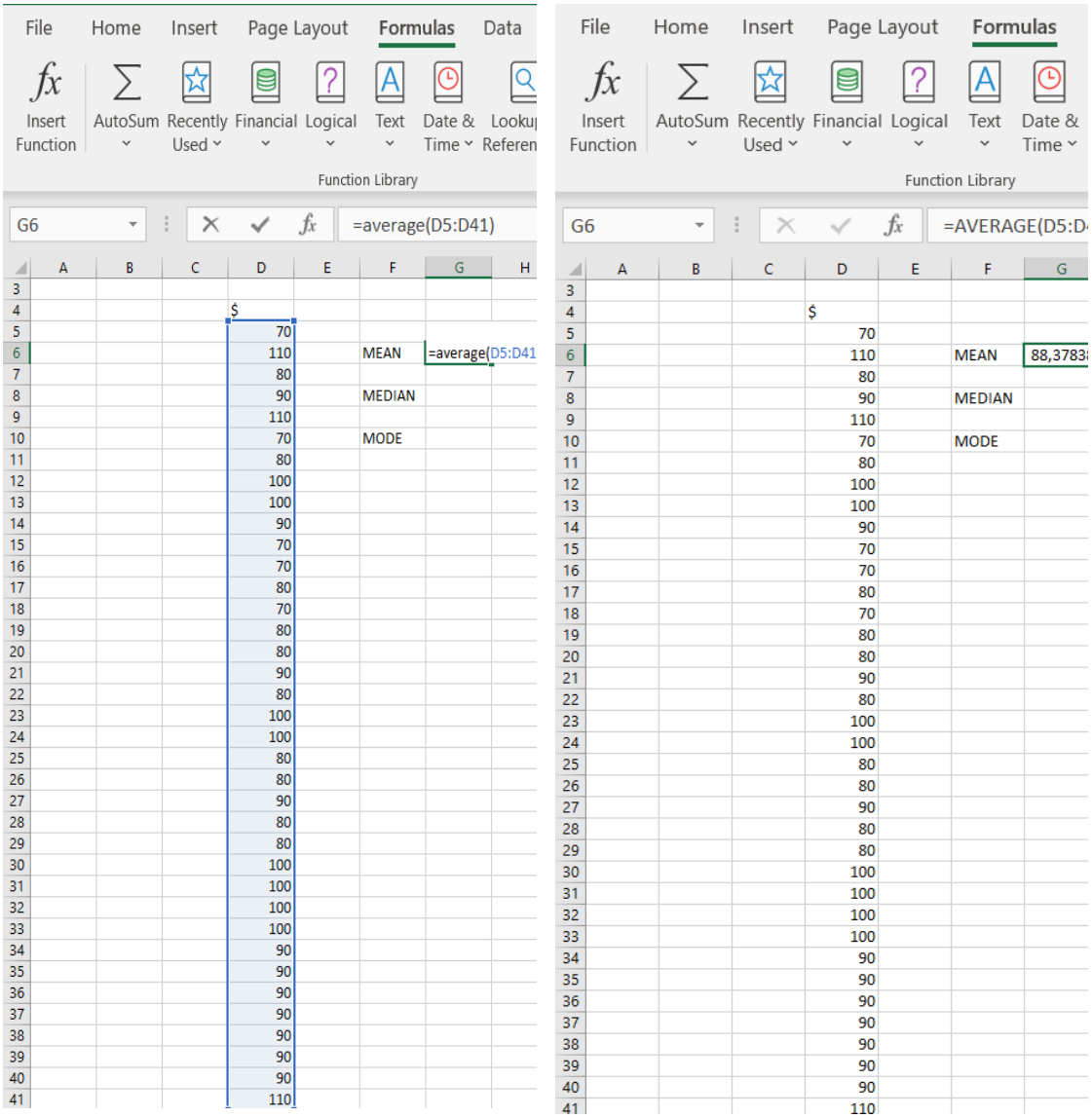


Figura 20. Pantalla de ejemplo 15

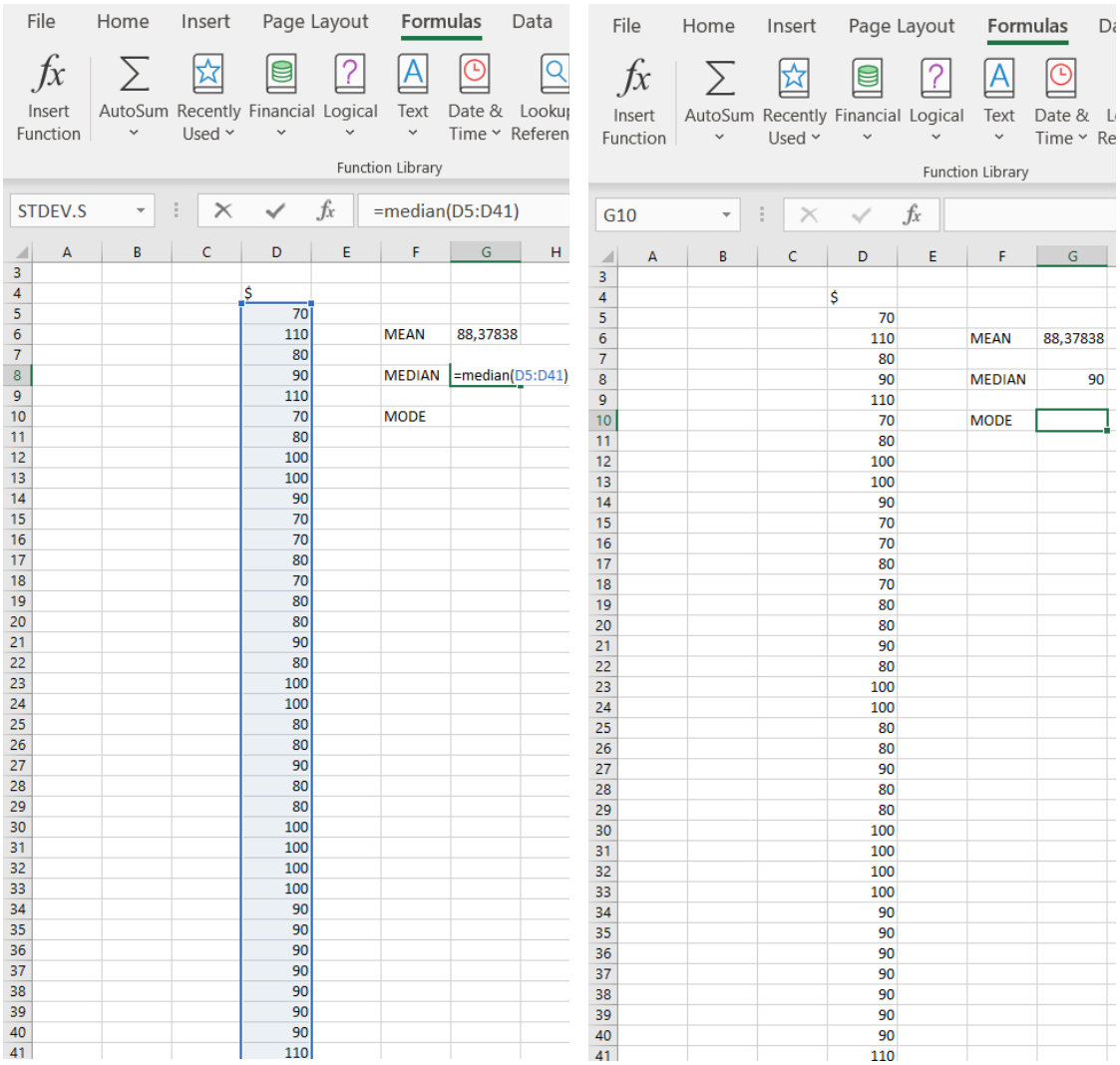


Figura 21. Pantalla de ejemplo 16

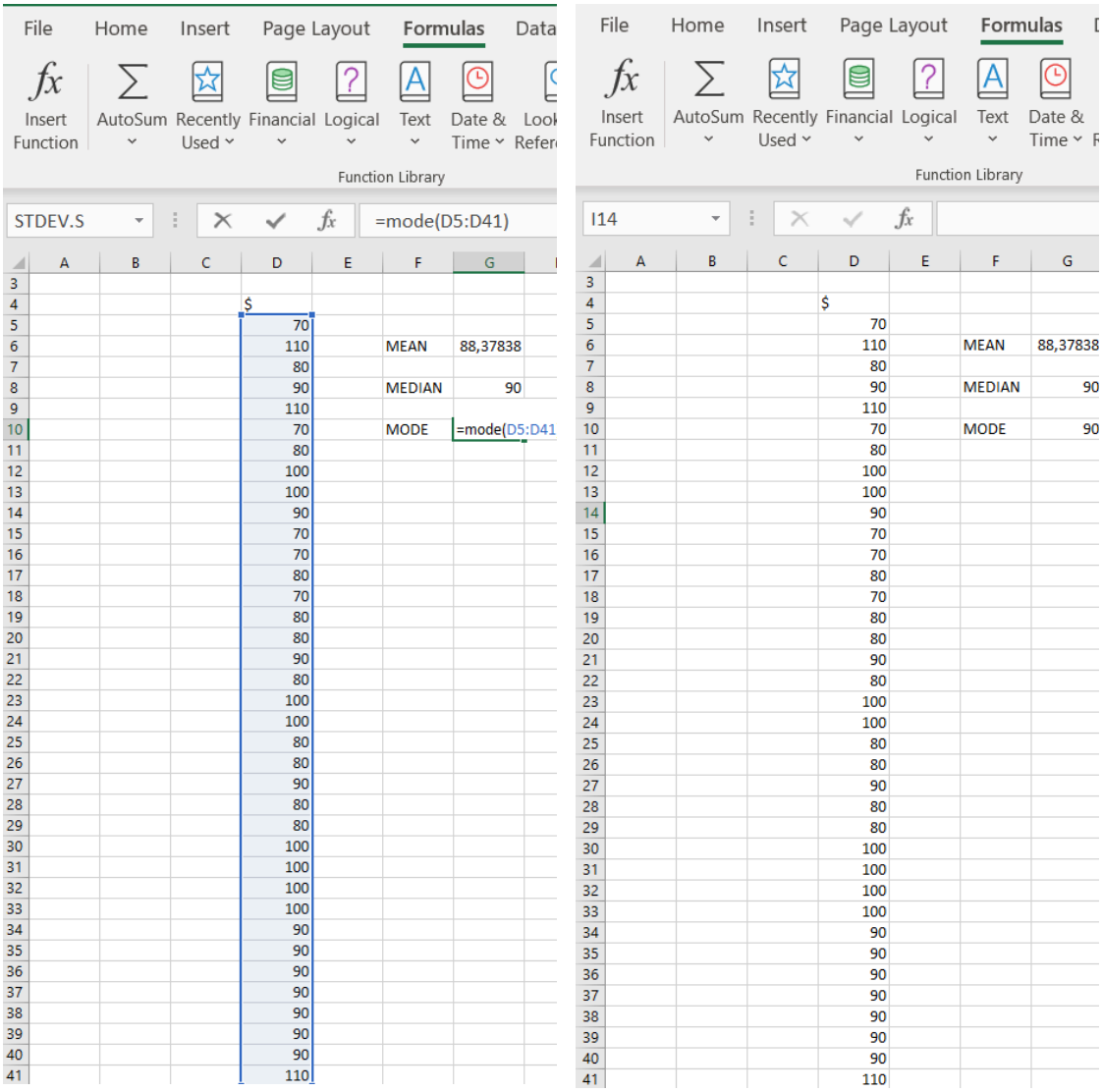


Figura 22. Pantalla de ejemplo 17

4.2. Demostración: medidas de la dispersión

La varianza (dispersión) en Excel puede calcularse con la herramienta de fórmula (más funciones) - VAR.S (varianza de la muestra) - Figuras 23-24.

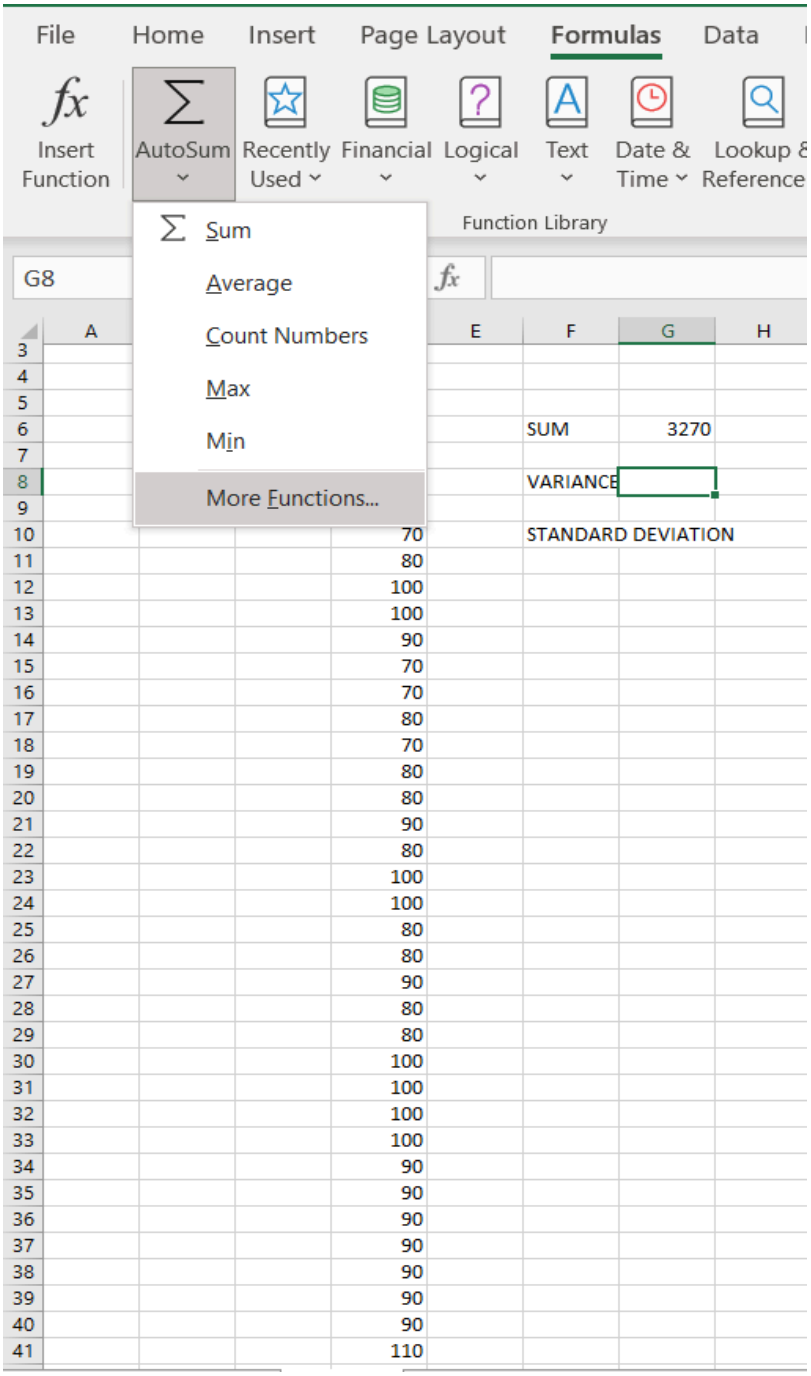


Figura 23. Pantalla de ejemplo 18

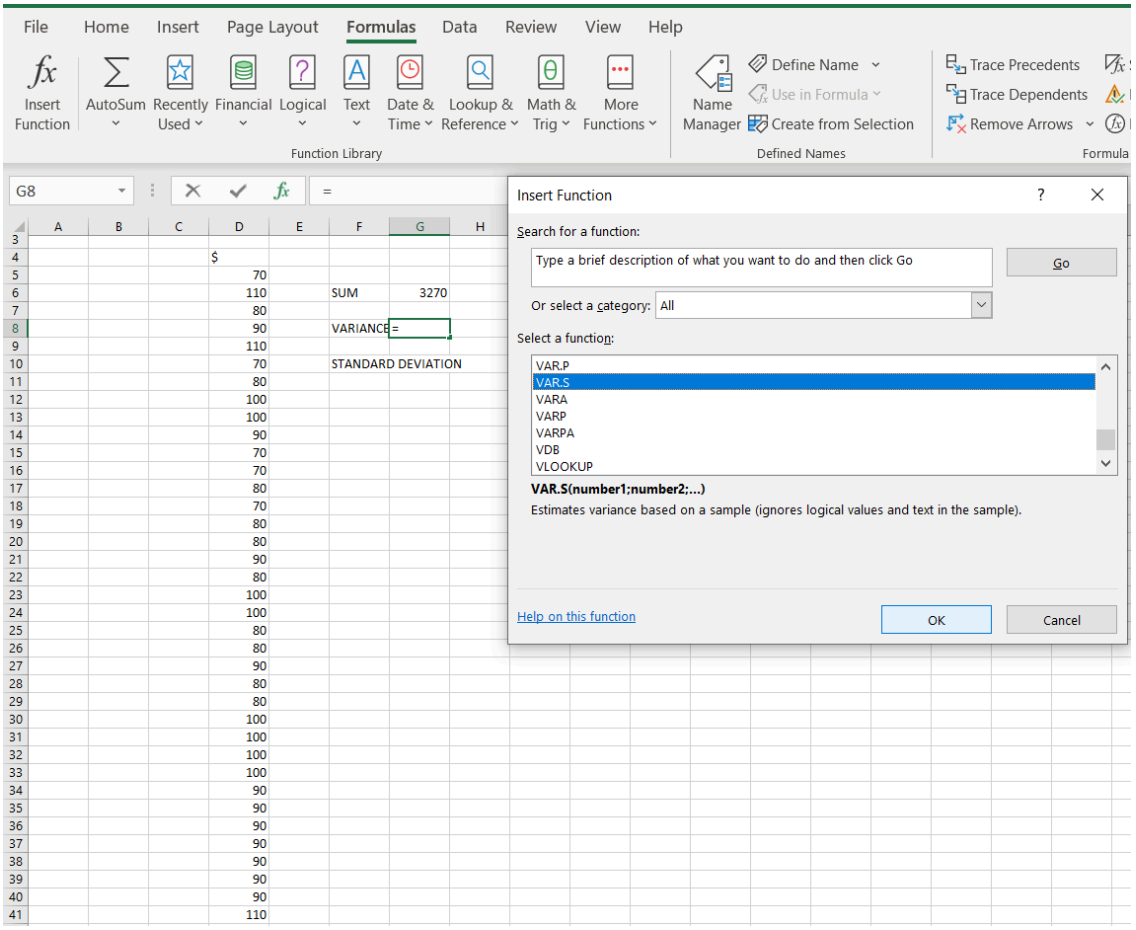


Figura 24. Pantalla de ejemplo 19

Necesitamos seleccionar los valores en la columna a ser probados (D5

) y presionar el botón OK. En nuestro caso, la varianza será 136.18 (Figura 25).

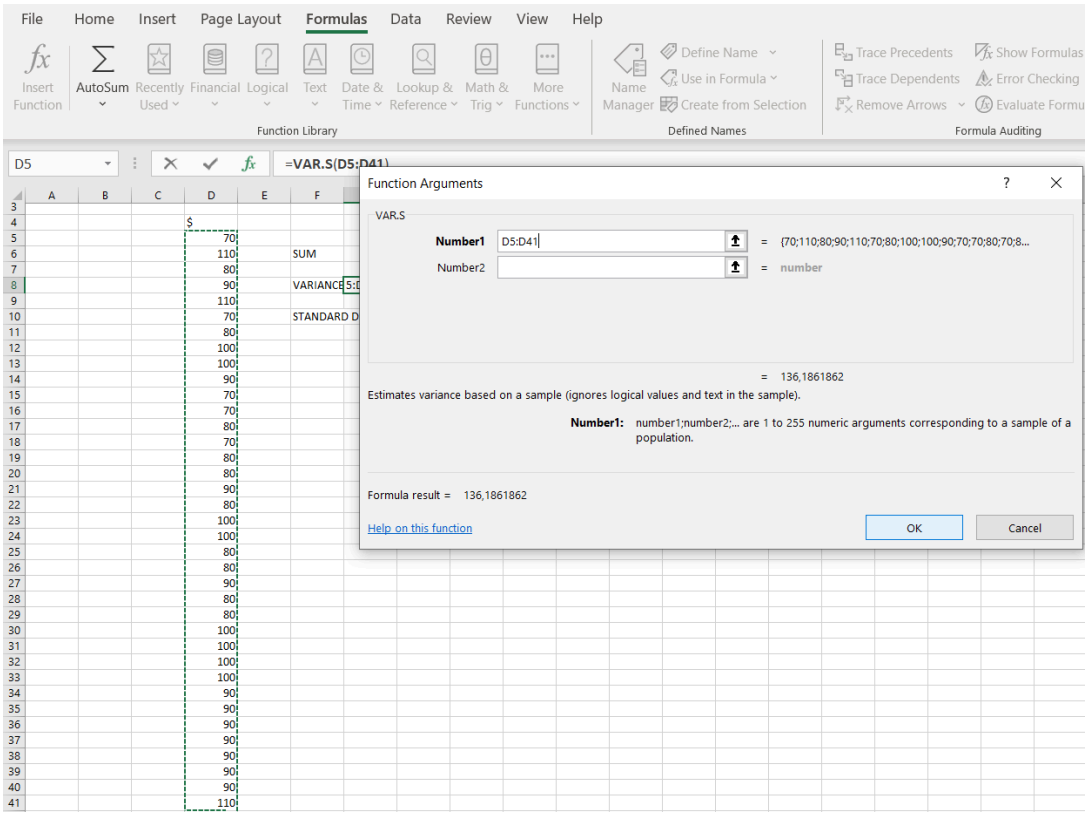


Figura 25. Pantalla de ejemplo 20

De manera similar a la varianza, podemos encontrar la desviación estándar en Excel (Figura 26).

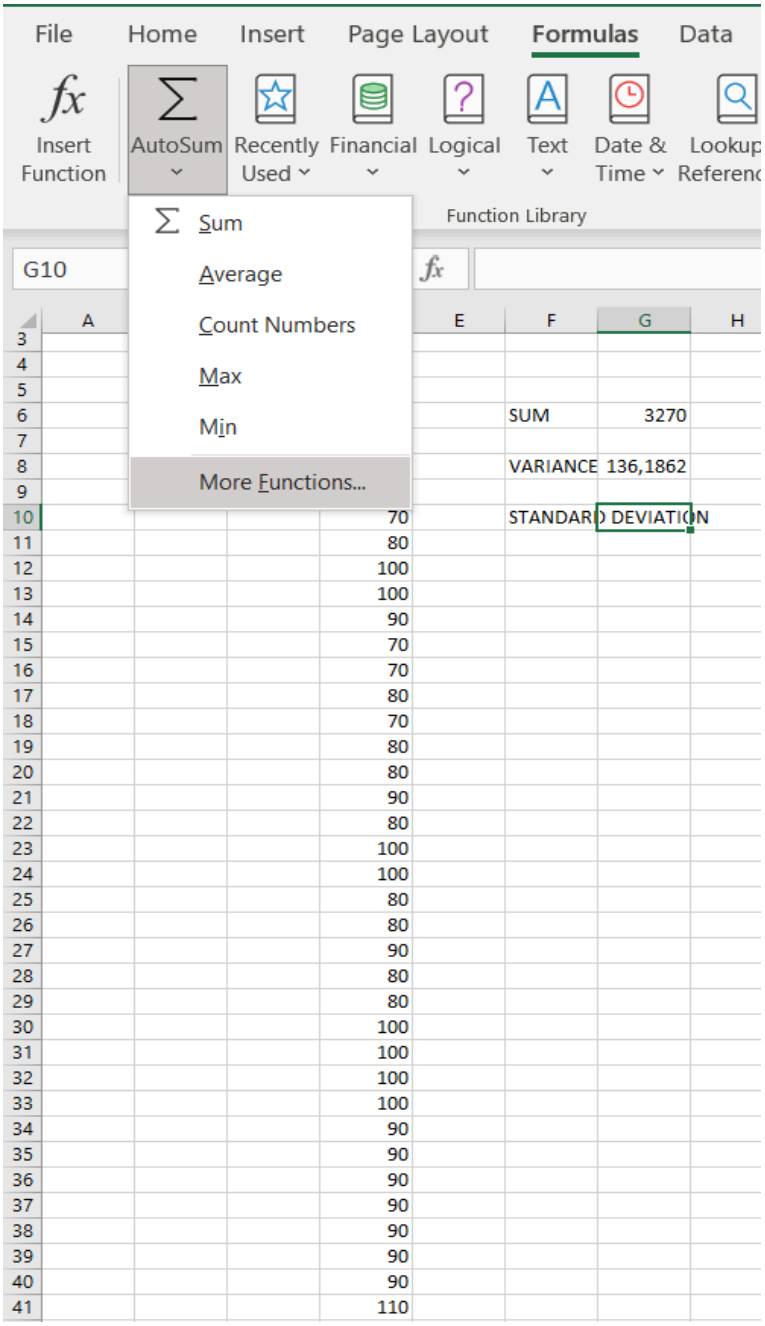


Figura 26. Pantalla de ejemplo 21

Después de hacer clic en el botón de fórmula, vaya a más funciones y seleccione STDEV.S (desviación estándar de la muestra) - Figura 27.

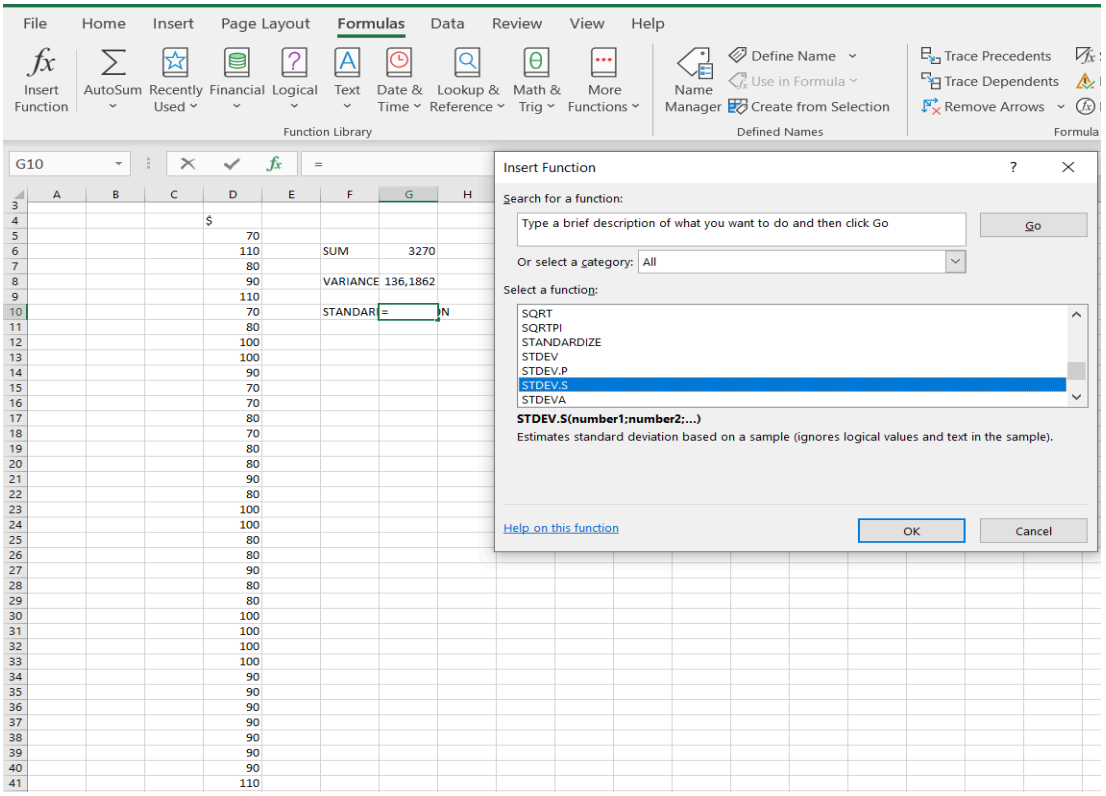


Figura 27. Pantalla de ejemplo 22

Seleccionamos el área de datos (D5:D41) y después de hacer clic en el botón OK, obtenemos el resultado – 11.66 (Figura 28).

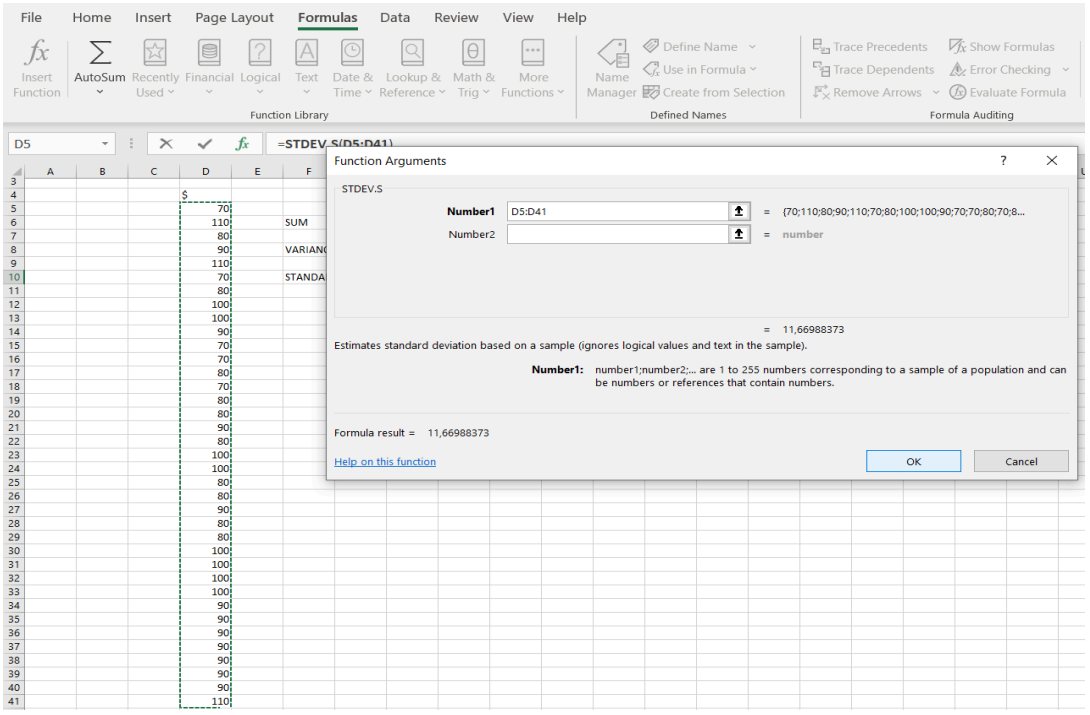


Figura 28. Pantalla de ejemplo 23

5. Probabilidad y la distribución normal gaussiana

La probabilidad es una medida de la probabilidad de que ocurra un evento. Se expresa como un número entre 0 y 1, donde 0 indica que el evento es imposible y 1 indica que el evento es seguro. La probabilidad se utiliza para cuantificar la incertidumbre asociada con un evento o conjunto de eventos dado (Figura 29).

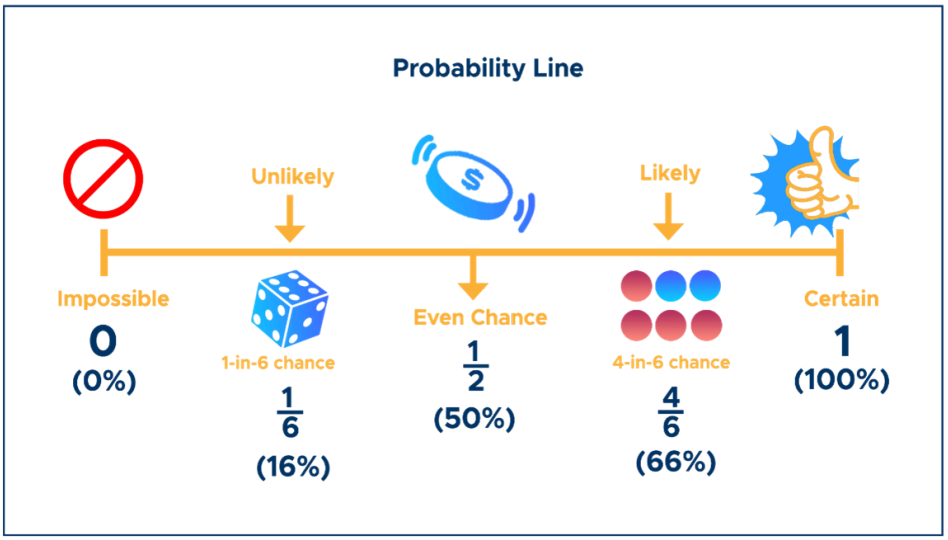


Figura 29. Línea de probabilidad (fuente abierta)

La probabilidad de que una mascota hable inglés es 0 o imposible. Lanzar un 6 en un dado tiene una probabilidad de 1/6 o 16%. Lanzar una moneda al aire y que salga cara tiene una probabilidad de 1/2 o 50%. Sacar 4 canicas rojas de un frasco que contiene 6 canicas tiene una probabilidad de 4/6 o 66%. Sacar cualquier canica de un frasco que contiene 6 canicas es seguro o

100%. La probabilidad puede expresarse como una fracción, decimal o porcentaje.

Por ejemplo, cuando se lanza un dado, hay seis resultados potenciales que pueden ocurrir: 1, 2, 3, 4, 5 o 6. La probabilidad de que aparezca cualquiera de estos números es 1/6 o 16% (Figura 30).

Figura 30. Probabilidad del dado

La probabilidad de que ocurra un evento (A) puede calcularse utilizando la fórmula (2):

$$P(A) = f / N, \quad (2)$$

donde f – es el número de formas en que puede ocurrir el evento (frecuencia), N – es el número total de resultados posibles.

Las probabilidades y la probabilidad están relacionadas, pero las probabilidades de que ocurra un evento solo pueden determinarse una vez que se ha calculado la probabilidad del evento (Sullivan 2016).

La **distribución normal (distribución gaussiana)** es un tipo de distribución de probabilidad que es simétrica alrededor de la media. Esto significa que los datos cercanos a la media son más probables de ocurrir que los datos alejados de la media. La distribución normal a menudo se representa como una curva en forma de campana (Figura 31).

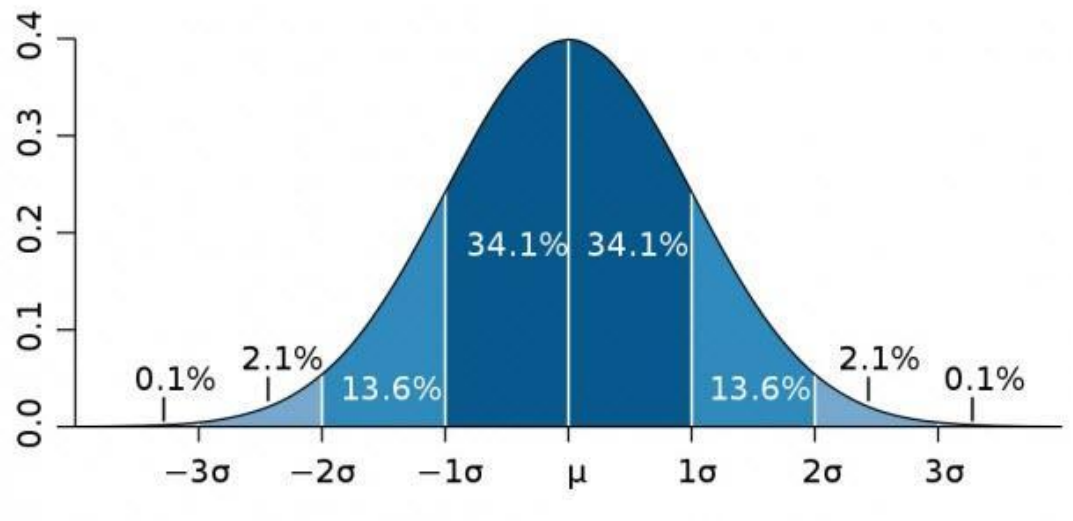


Figura 31. Distribución gaussiana (fuente abierta)

Una distribución de probabilidad es una forma de mostrar los diferentes valores que puede tomar una variable aleatoria y cuán probable es cada valor. El rango de valores está entre los valores mínimos y máximos posibles. Los valores exactos dependen de la media (promedio), desviación estándar, sesgo y curtosis de la distribución. Las distribuciones de probabilidad se utilizan para medir la probabilidad de ciertos resultados. La más común es la distribución normal, también conocida como la "curva en forma de campana". Otras distribuciones también se utilizan. La función de densidad de probabilidad es el proceso utilizado para determinar la distribución de probabilidad de un fenómeno (Sullivan 2016). Las funciones de distribución acumulativa se crean sumando las probabilidades de ocurrencias. Académicos, analistas financieros y gestores de fondos utilizan distribuciones de probabilidad para predecir los rendimientos esperados de una acción, etc.

En una distribución normal, la media es cero y la desviación estándar es 1. Tiene un sesgo cero y una curtosis de 3. Las distribuciones normales son simétricas, pero no todas las distribuciones simétricas son normales. Muchos fenómenos naturales tienden a aproximarse a la distribución normal. En finanzas, la mayoría de las distribuciones de precios no son, sin embargo, perfectamente normales.

5.1. Juego "El problema de Monty Hall"

El problema de Monty Hall es un famoso rompecabezas de probabilidad nombrado en honor al presentador de un programa de juegos, Monty Hall. Se basa en un programa de juegos en el que a un concursante se le presentan tres puertas, detrás de una de las cuales hay un premio. Al concursante se le pide que elija una de las puertas, y luego el presentador, Monty Hall, abre una de las otras puertas para revelar una opción perdedora (Figura 32). Al concursante se le pregunta entonces si le gustaría cambiar su elección a la puerta que queda sin abrir.

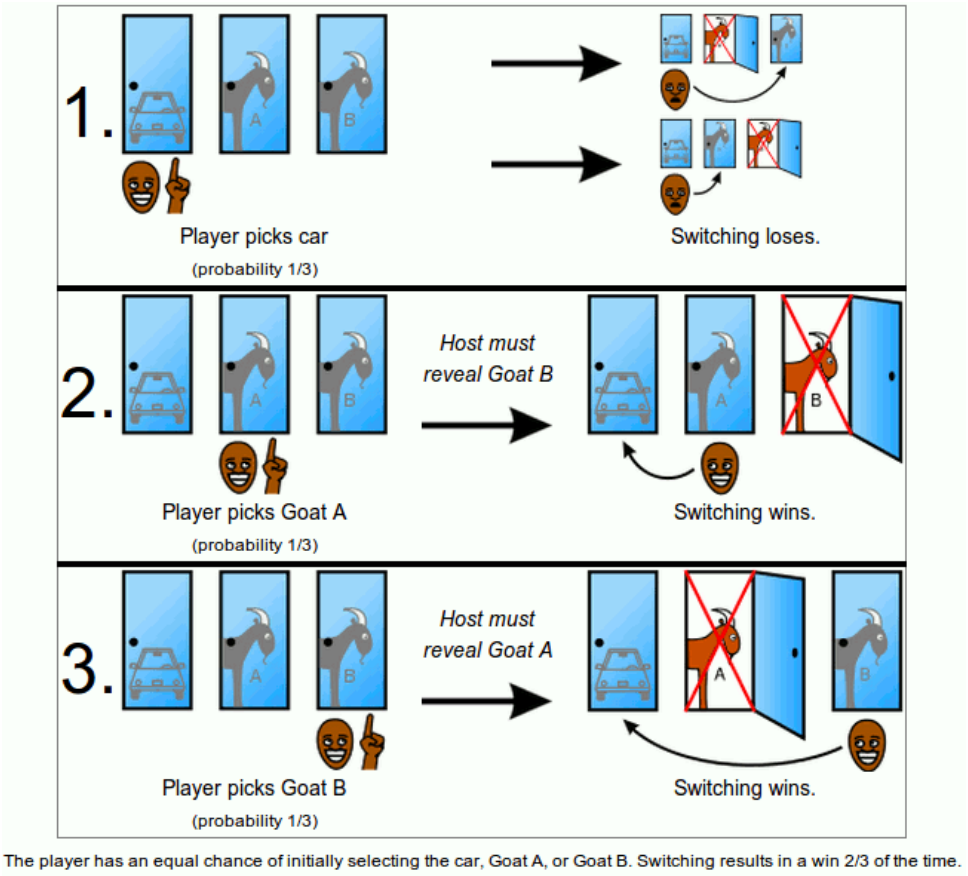


Figura 32. Problema de Monty Hall (fuente abierta)

Para probar las probabilidades de ganar el juego, intenta jugarlo 50 veces utilizando una estrategia de "elegir y mantener", donde eliges una puerta y la mantienes sin importar lo que ocurra (<LINK> https://montyhall.io/). Descubrirás que tu tasa de éxito se estabilizará alrededor de 1/3. Luego,

reinicia y juega 20 veces utilizando un enfoque de "elegir y cambiar", donde eliges una puerta y cambias a la otra cuando Monty descubre una cabra.

No importa lo que haga Monty, mis probabilidades de ganar con mi elección original permanecen en 1 de 3. La otra puerta debe tener las 2 de 3 probabilidades restantes de ganar, lo que explica por qué las probabilidades "mejoran" del otro lado.

5.2. Demostración: Probabilidad

La probabilidad es la posibilidad de que ocurra un evento, expresada como una relación de resultados favorables al número total de resultados posibles. Excel tiene una función PROB que puede usarse para calcular la probabilidad de un evento.

Para obtener resultados de probabilidad precisos, debemos organizar los datos antes de calcular. La tabla a continuación muestra las ventas de productos y sus respectivas probabilidades. Todas las probabilidades deben sumar 100, de lo contrario, la función PROB devolverá un error #NUM! (Figura 33).

Product sales	Probability
20	0.23
50	0.50
80	0.18
110	0.09

Figura 33. Pantalla de ejemplo

Queremos calcular la probabilidad de que las ventas de un producto estén entre 50 y 80. Para realizar el cálculo, ingresamos esta fórmula en la celda J32=PROB(I25:I28,J25:J28,J30,J31), donde I25:I28 es el rango que contiene los valores de las ventas de productosi, J25:J28 contiene las probabilidades para cada cantidad, J30 es el límite inferior de 50, mientras que J31 es el límite superior de 80 (Figuras 34-35).

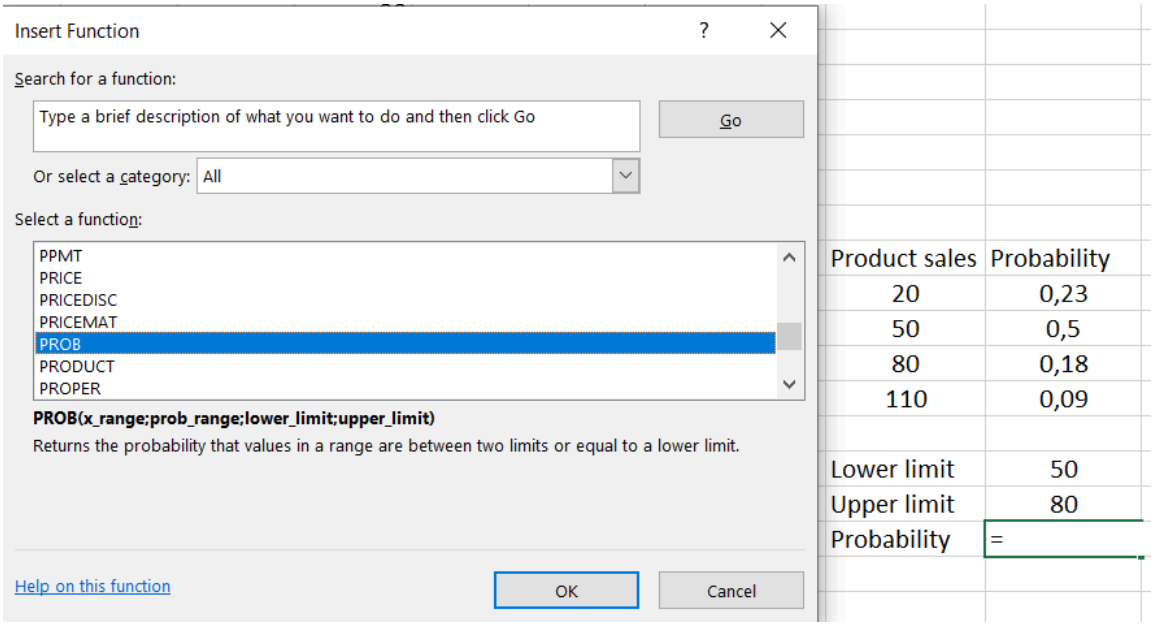


Figura 34. Pantalla de ejemplo

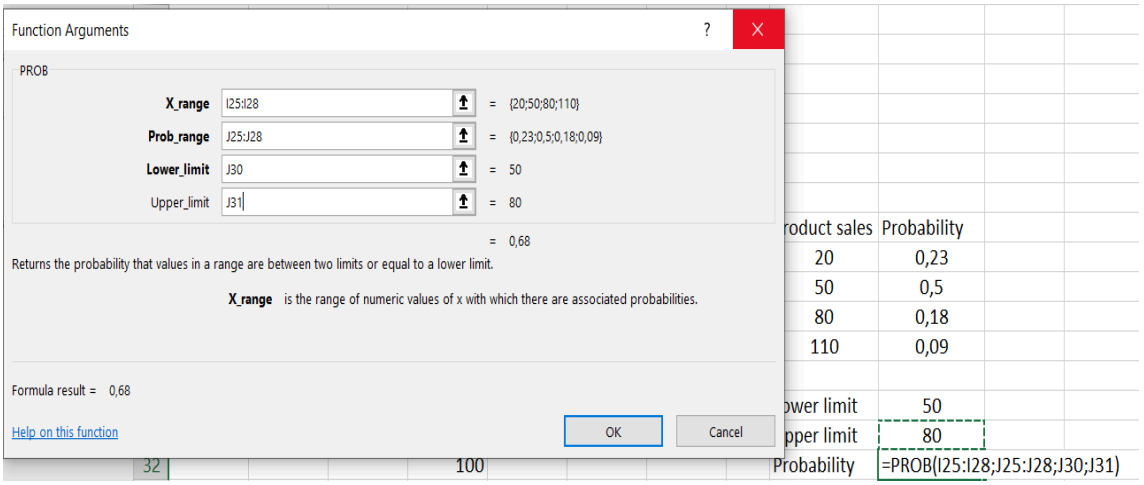


Figura 35. Pantalla de ejemplo

Como resultado, la probabilidad en la celda J32 es 0.68 o 68%, que es la probabilidad de que las ventas de productos estén entre 50 y 80 (Figura 36).

Product sales	Probability
20	0,23
50	0,5
80	0,18
110	0,09
Lower limit	50
Upper limit	80
Probability	0,68

Figura 36. Pantalla de ejemplo 27

5.3. Demostración: Distribución Normal

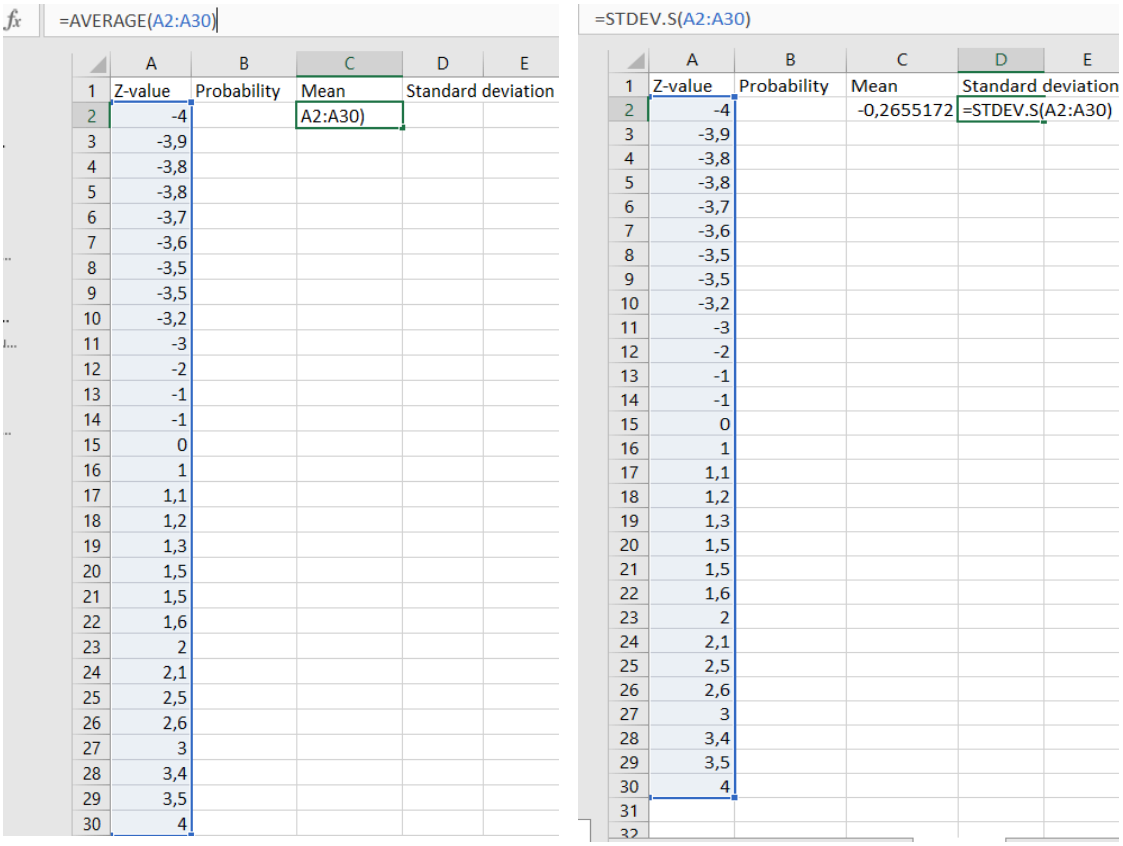
La función NORMDIST es una función estadística de Excel que calcula la función de densidad de probabilidad normal o la función de distribución normal acumulativa para un conjunto dado de parámetros, como una media y una desviación estándar indicadas.

La función NORMDIST utiliza los siguientes argumentos (fórmula 3):

=NORMDIST(x,mean,standard_dev,cumulative) (3)

donde *x* (argumento requerido) – es el valor para el cual deseamos calcular la distribución; *mean* (argumento requerido) – es la media aritmética de la distribución; *standard_dev* (argumento requerido) – es la desviación estándar de la distribución; *cumulative* (argumento requerido) – es un valor logico (especifica el tipo de distribución a utilizar: TRUE (función de distribución normal acumulativa) o FALSE (función de densidad de probabilidad normal).

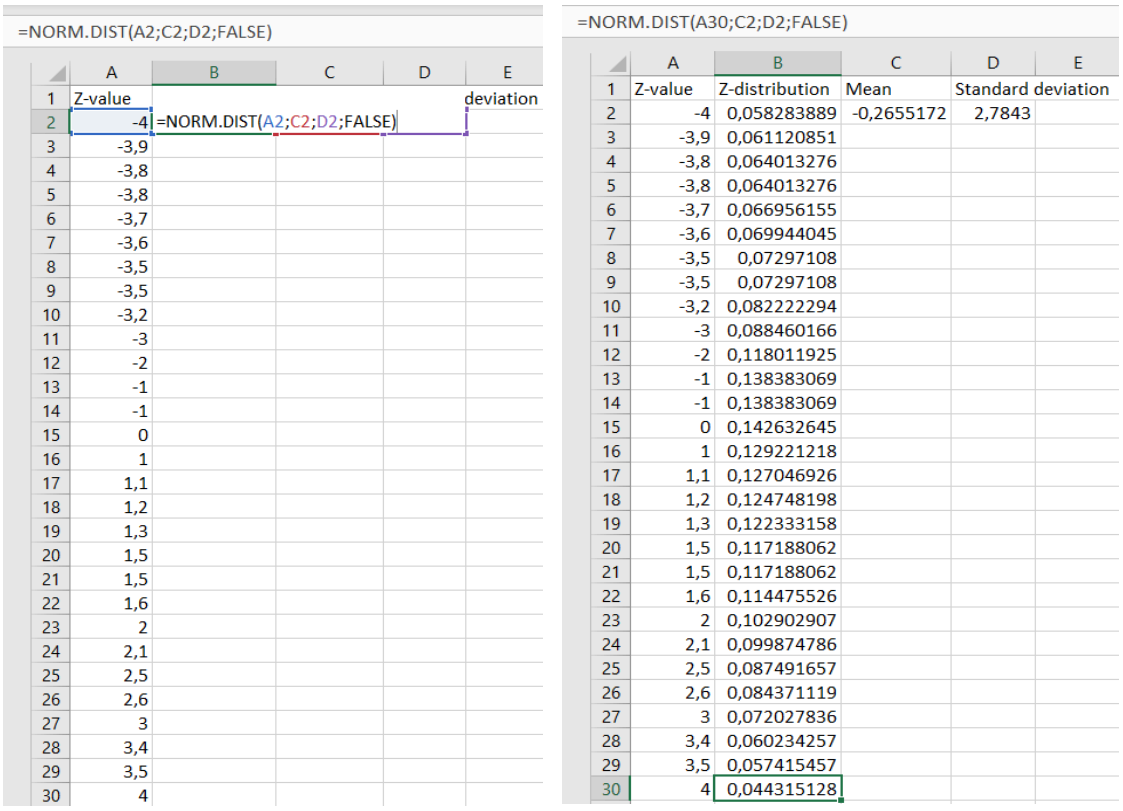
Todos los pasos de cálculo se muestran en las figuras 37-40.



	A	B	C	D	E
1	Z-value	Probability	Mean	Standard deviation	
2	-4		A2:A30)		
3	-3,9				
4	-3,8				
5	-3,8				
6	-3,7				
7	-3,6				
8	-3,5				
9	-3,5				
10	-3,2				
11	-3				
12	-2				
13	-1				
14	-1				
15	0				
16	1				
17	1,1				
18	1,2				
19	1,3				
20	1,5				
21	1,5				
22	1,6				
23	2				
24	2,1				
25	2,5				
26	2,6				
27	3				
28	3,4				
29	3,5				
30	4				

	A	B	C	D	E
1	Z-value	Probability	Mean	Standard deviation	
2	-4		-0,2655172	=STDEV.S(A2:A30)	
3	-3,9				
4	-3,8				
5	-3,8				
6	-3,7				
7	-3,6				
8	-3,5				
9	-3,5				
10	-3,2				
11	-3				
12	-2				
13	-1				
14	-1				
15	0				
16	1				
17	1,1				
18	1,2				
19	1,3				
20	1,5				
21	1,5				
22	1,6				
23	2				
24	2,1				
25	2,5				
26	2,6				
27	3				
28	3,4				
29	3,5				
30	4				
31					

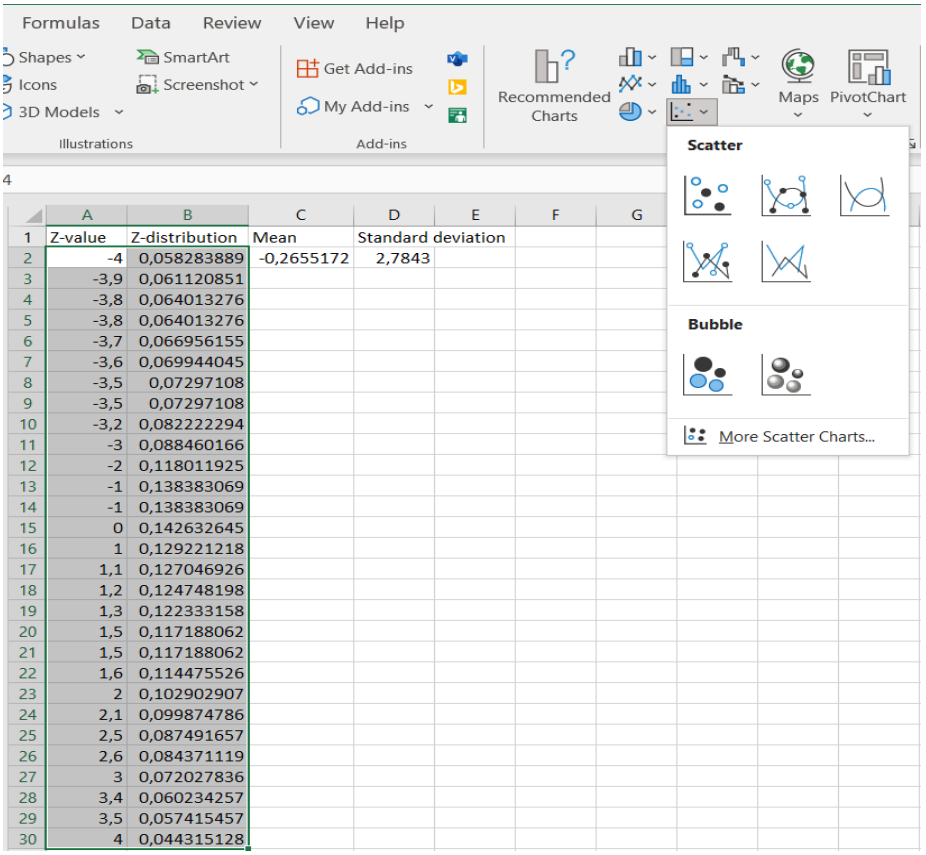
Figura 37. Pantalla de ejemplo 28



	A	B	C	D	E
1	Z-value				deviation
2	-4	=NORM.DIST(A2;C2;D2;FALSE)			
3	-3,9				
4	-3,8				
5	-3,8				
6	-3,7				
7	-3,6				
8	-3,5				
9	-3,5				
10	-3,2				
11	-3				
12	-2				
13	-1				
14	-1				
15	0				
16	1				
17	1,1				
18	1,2				
19	1,3				
20	1,5				
21	1,5				
22	1,6				
23	2				
24	2,1				
25	2,5				
26	2,6				
27	3				
28	3,4				
29	3,5				
30	4				

	A	B	C	D	E
1	Z-value	Z-distribution	Mean	Standard deviation	
2	-4	0,058283889	-0,2655172	2,7843	
3	-3,9	0,061120851			
4	-3,8	0,064013276			
5	-3,8	0,064013276			
6	-3,7	0,066956155			
7	-3,6	0,069944045			
8	-3,5	0,07297108			
9	-3,5	0,07297108			
10	-3,2	0,082222294			
11	-3	0,088460166			
12	-2	0,118011925			
13	-1	0,138383069			
14	-1	0,138383069			
15	0	0,142632645			
16	1	0,129221218			
17	1,1	0,127046926			
18	1,2	0,124748198			
19	1,3	0,122333158			
20	1,5	0,117188062			
21	1,5	0,117188062			
22	1,6	0,114475526			
23	2	0,102902907			
24	2,1	0,099874786			
25	2,5	0,087491657			
26	2,6	0,084371119			
27	3	0,072027836			
28	3,4	0,060234257			
29	3,5	0,057415457			
30	4	0,044315128			

Figura 38. Pantalla de ejemplo 29



	A	B	C	D	E	F	G
1	Z-value	Z-distribution	Mean	Standard deviation			
2	-4	0,058283889	-0,2655172	2,7843			
3	-3,9	0,061120851					
4	-3,8	0,064013276					
5	-3,8	0,064013276					
6	-3,7	0,066956155					
7	-3,6	0,069944045					
8	-3,5	0,07297108					
9	-3,5	0,07297108					
10	-3,2	0,082222294					
11	-3	0,088460166					
12	-2	0,118011925					
13	-1	0,138383069					
14	-1	0,138383069					
15	0	0,142632645					
16	1	0,129221218					
17	1,1	0,127046926					
18	1,2	0,124748198					
19	1,3	0,122333158					
20	1,5	0,117188062					
21	1,5	0,117188062					
22	1,6	0,114475526					
23	2	0,102902907					
24	2,1	0,099874786					
25	2,5	0,087491657					
26	2,6	0,084371119					
27	3	0,072027836					
28	3,4	0,060234257					
29	3,5	0,057415457					
30	4	0,044315128					

Figura 39. Pantalla de ejemplo 30

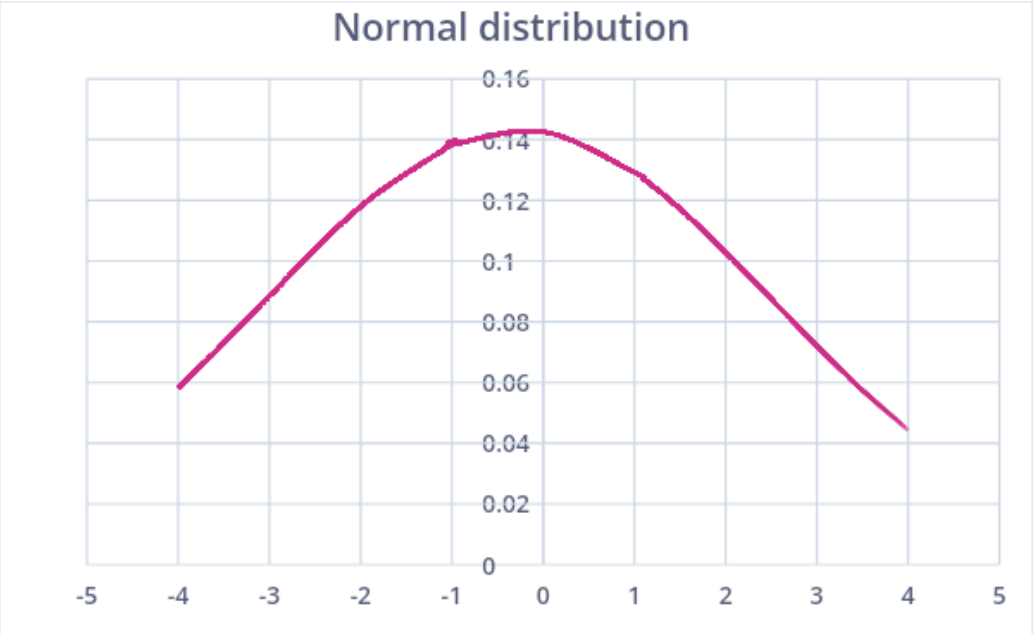


Figura 40. Pantalla de ejemplo 31

6. Muestreo y experimentos controlados aleatorizados

Para que los economistas, financieros y gerentes tomen decisiones informadas, se necesitan datos. De dónde provienen los datos y cómo se analizan por personas y máquinas depende de muchos factores, como lo que estamos tratando de hacer con los resultados, cuán precisas deben ser las conclusiones, el tamaño del presupuesto para la investigación, etc. Hay todo un espectro de experimentos que los gerentes pueden realizar, desde experimentos rápidos e informales hasta estudios piloto, experimentos de campo y estudios de laboratorio. Uno de los experimentos más estructurados es el experimento controlado aleatorizado.

En el experimento, la variable de interés para el investigador se llama variable dependiente (puede haber múltiples variables dependientes), sin embargo, hay muchas variables independientes: factores que afectan directa o indirectamente a la variable dependiente. Generalmente, durante el experimento, el investigador intenta identificar y considerar una o más variables independientes, sin embargo, muchos factores pueden convertirse en un obstáculo para el experimento.

La información necesaria para que un investigador analice y tome decisiones a niveles macro y micro de gestión puede obtenerse de diversas fuentes de datos (por ejemplo, https://data.worldbank.org, https://prosperity.com, https://transparency.org, https://statista.com, https://ec.europa.eu). La tarea principal aquí es conocer claramente qué tomaremos como variable dependiente y qué variables independientes necesitamos para un análisis, pruebas y formulación de hipótesis correctos para su verificación. Todas las decisiones de gestión deben basarse en un análisis de datos de fuentes abiertas (PIB per cápita, operaciones de exportación-importación, tasa de desempleo, tasa de inflación, índices de desarrollo socioeconómico de diferentes países, etc.) o en los resultados de sus propias encuestas si se necesitan datos específicos para la

investigación. El uso de métodos cuantitativos y cualitativos de investigación, así como su combinación, ayuda a obtener resultados relevantes y a tomar decisiones correctas en la vida empresarial y social.

7. Distribuciones de muestreo y el teorema del límite central

Una distribución de muestreo es una distribución de probabilidad de una estadística que se calcula a partir de muchas muestras tomadas de una población particular. Es la distribución de frecuencias de los diversos posibles resultados de una estadística de una población. Una población es el conjunto completo de elementos de los cuales se toma una muestra estadística y puede referirse a un grupo de personas, objetos, eventos, visitas al hospital o mediciones. En otras palabras, es una observación colectiva de sujetos que comparten una característica común.

Muchos de los datos recopilados y utilizados por académicos, estadísticos, investigadores, publicistas, analistas, etc., son muestras, no grupos completos. Una muestra es un grupo más pequeño tomado de una población más grande (Barone, Brock, Schmitt 2023).

El **teorema del límite central** (TLC) establece que cuando se toma un tamaño de muestra suficientemente grande de una población con un nivel finito de varianza, la media de la muestra estará cerca de la media de la población y la muestra seguirá una distribución normal. Esto se debe a la ley de los grandes números, que establece que a medida que aumenta el tamaño de la muestra, la varianza de la muestra se acercará a la varianza de la población. Abraham de Moivre introdujo por primera vez la idea del teorema del límite central en 1733, pero no fue reconocido oficialmente hasta que George Pólya, un matemático húngaro, le dio el nombre en 1930.

El teorema del límite central se compone de varias características clave (Barone, Brock, Schmitt 2023):

1. El muestreo es sucesivo (algunas unidades de muestra son comunes con unidades de muestra seleccionadas en ocasiones anteriores).
2. El muestreo es aleatorio (todas las muestras deben seleccionarse al azar para que tengan la misma posibilidad estadística de ser seleccionadas).
3. Las muestras deben ser independientes (las selecciones o resultados de una muestra no deben influir en las futuras muestras o en otros resultados de muestra).
4. Las muestras deben ser limitadas (la muestra no debe ser más del 10% de una población si el muestreo se realiza sin reemplazo).
5. El tamaño de la muestra está aumentando (el teorema del límite central es relevante a medida que se seleccionan más muestras).

Generalmente, se piensa que un tamaño de muestra de 30-50 es suficiente para que el teorema del límite central sea aplicable, lo que resulta en que las medias de las muestras estén normalmente distribuidas. A medida que se toman más muestras, el gráfico de los resultados se parecerá cada vez más a una distribución normal. Incluso con tamaños de muestra más pequeños, como 8 o 5, el teorema del límite central aún puede aproximarse.

Ilustrar gráficamente el TLC puede hacerse a través de un experimento que involucra el lanzamiento de un dado. A medida que aumenta el número de lanzamientos, la forma de la distribución de las medias se parecerá cada vez más a la de una distribución normal (Figura 41).

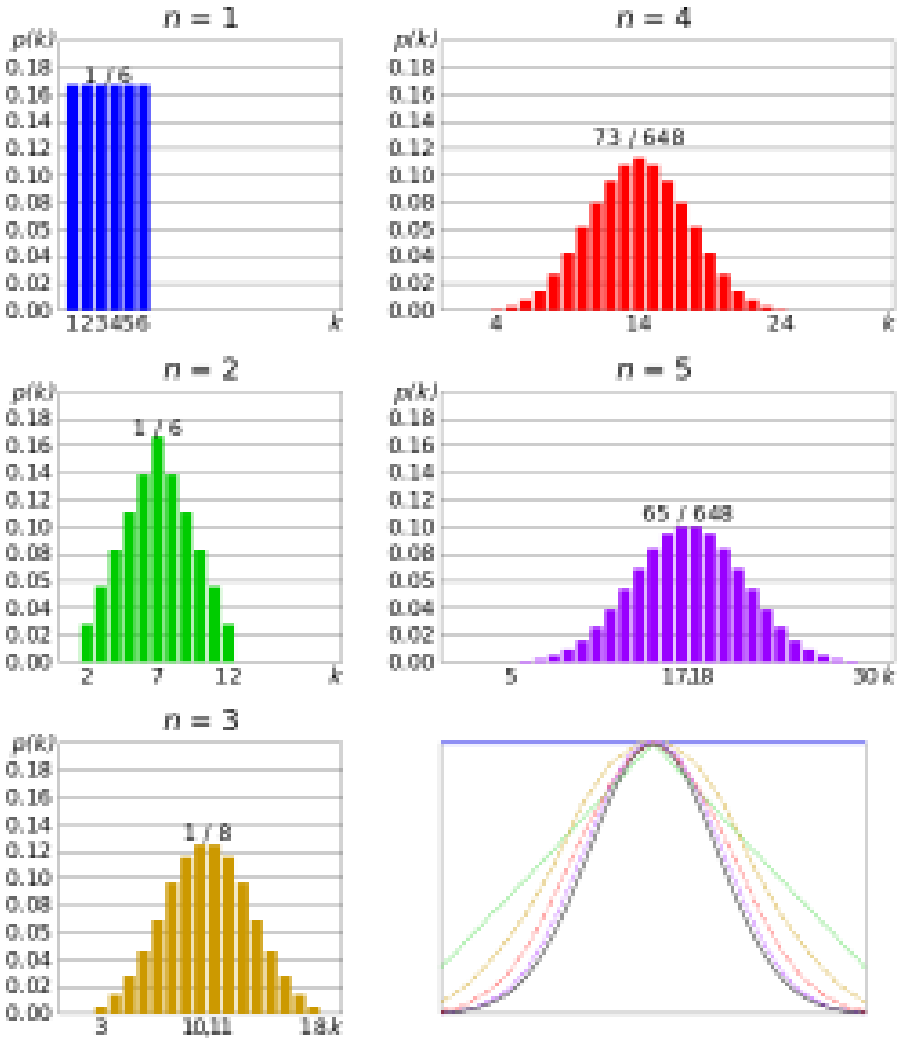


Figura 41. Ilustración del TLC (fuente abierta)

A z-score (puntaje estándar) es una medida de cuántas desviaciones estándar por debajo o por encima de la media de la población se encuentra un puntaje bruto. Para utilizar un z-score, necesitamos conocer la media y la desviación estándar de la población.

La fórmula básica del z-score para una muestra es (4):

$$z = (x - \mu) / (\sigma / \sqrt{n}), \quad (4)$$

donde x – es el puntaje de la prueba (número asociado con "mayor que"test score , μ – mesia (average), σ – desviación estándar, n – es el tamaño de la muestra.

The CLT puede utilizarse para analizar los rendimientos de una sola acción o de un índice más grande. Por ejemplo, si un inversor quiere analizar el rendimiento de un índice con 1000 acciones, puede estudiar una muestra aleatoria de acciones para estimar el rendimiento total del índice. Para garantizar la precisión, se deben elegir al menos 30-50 acciones de diferentes sectores y cambiar regularmente para evitar sesgos.

Por ejemplo, hay 250 perros en una exposición canina que pesan un promedio de 12 libras con una desviación estándar de 8 libras (Figura 42). Si se eligen al azar 4 perros, cuál es la probabilidad de que tengan un peso promedio mayor de 8 libras y menor de 25 libras?

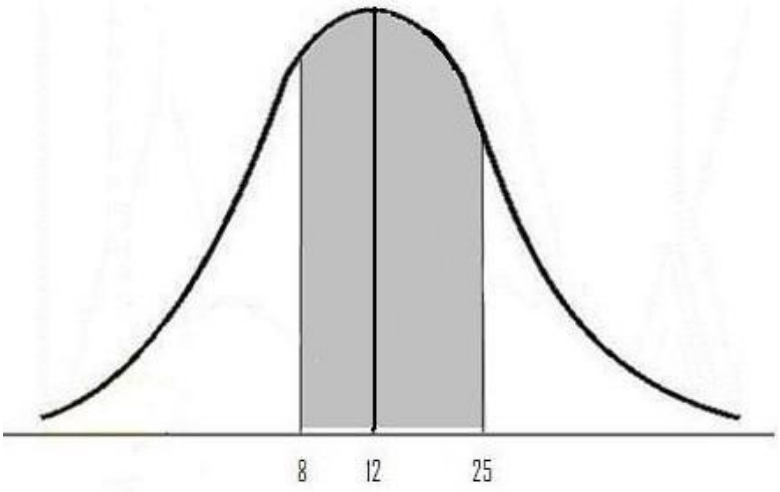


Figura 42. Ilustración del ejemplo (fuente abierta)

En la z-table (https://www.z-table.com/?utm_content=cmp-true), el valor z de 3.25 corresponde a 0.9994 y el valor z de -1 corresponde a 0.1587.

0.9994 - 0.1587 = 0.8407 o 84.07%.

8. Regresión

La regresión lineal es un método de predicción de una respuesta escalar única basado en una o más variables explicativas. Si solo hay una variable

explicativa, se conoce como regresión lineal simple, y si hay múltiples variables explicativas, se llama regresión lineal múltiple. Esto es diferente de la regresión lineal multivariante, que se utiliza para predecir múltiples variables dependientes que están correlacionadas.

La regresión lineal es una forma de análisis que modela la relación entre una variable de respuesta y una o más variables explicativas ajustando una función predictora lineal con parámetros desconocidos que se estiman a partir de los datos. Este tipo de modelo se conoce como modelo lineal y típicamente asume que la media de la respuesta dada por los valores de las variables explicativas es una función afín de esos valores. En algunos casos, se puede utilizar la mediana u otro cuantil en lugar de la media. La regresión lineal, similar a otros análisis de regresión, se concentra en la probabilidad de la respuesta basada en los valores de los predictores en lugar de la probabilidad colectiva de todas las variables, que es el área del análisis multivariante.

El análisis de regresión es un conjunto de técnicas estadísticas utilizadas para determinar la conexión entre una variable dependiente ("resultado" o "variable de respuesta") y una o más variables independientes ("predictores" o "características"). El tipo más común de análisis de regresión es la regresión lineal, que encuentra la línea (o una combinación lineal más compleja) que mejor se ajusta a los datos basándose en un criterio matemático particular.

Técnicas más comunes de análisis de regresión (Henrique 2019):

1. Regresión Lineal. La técnica de modelado más ampliamente utilizada es la regresión lineal, que asume una conexión lineal entre una variable dependiente (Y) y una variable independiente (X). Utiliza una línea de regresión, también conocida como línea de mejor ajuste (Regresión lineal).

La conexión lineal se define por la fórmula (5):

$$Y = c + m*X + e, \quad (5)$$

donde c – es la intersección, m – es la pendiente de la línea, e – término de error.

El modelo de regresión lineal puede ser simple (con solo una variable dependiente y una variable independiente) o complejo (con múltiples variables dependientes e independientes).

Una correlación lineal entre dos variables puede ser positiva o negativa, puede no haber relación entre las variables o la correlación es no lineal.

La fuerza de la relación entre variables también puede variar: desde perfectamente positiva ($R^2 = 1$), perfectamente negativa ($R^2 = -1$) hasta débilmente positiva ($R^2 = 0.4$), o débilmente negativa ($R^2 = -0.4$) – Figure 43.

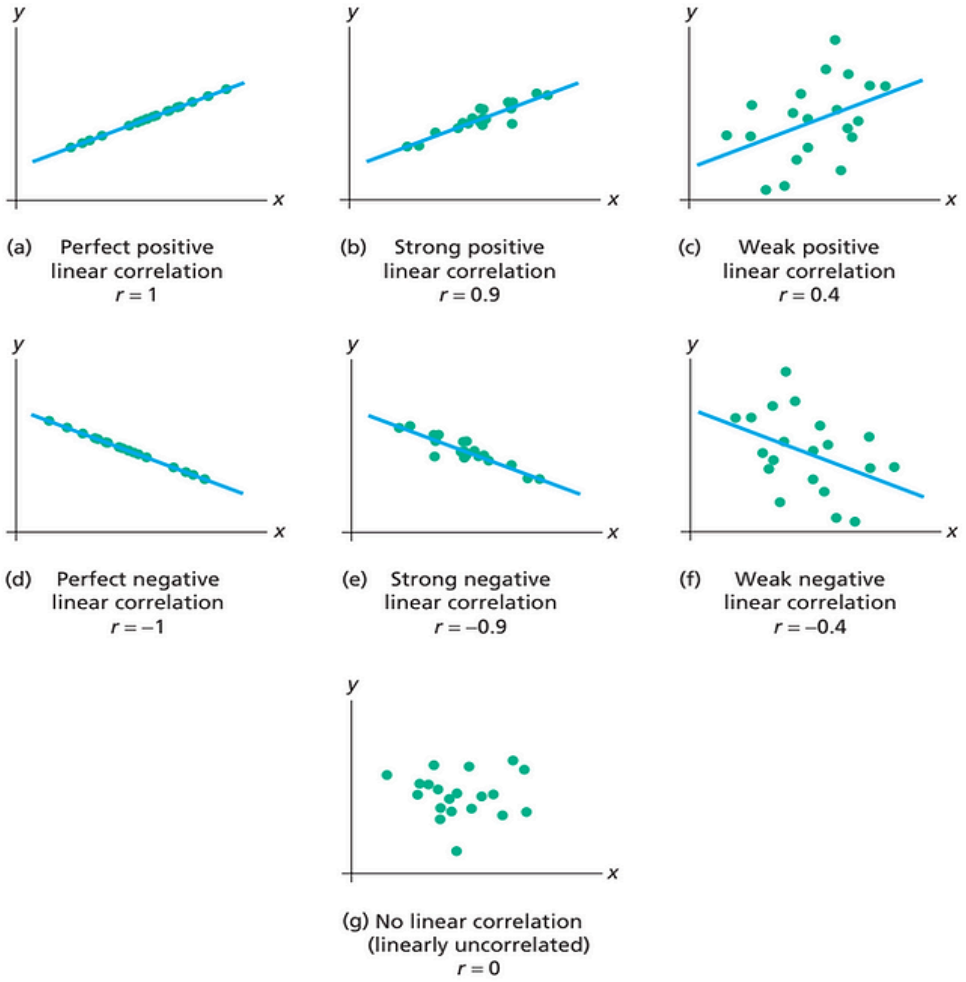


Figura 43. Ejemplos de correlación (fuente abierta)

Ejemplos de regresión lineal: la relación entre los niveles de contaminación y el aumento de las temperaturas; la relación entre la tasa de desempleo y el PIB per cápita; la relación entre la altura, el peso y la cantidad de ejercicio

Co-funded by the European Union

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

como variables independientes y la presión arterial como la variable dependiente.

2. Regresión Logística. La regresión logística es una técnica utilizada cuando la variable dependiente es discreta, lo que significa que solo puede tomar uno de dos valores. Esta técnica se utiliza para calcular la probabilidad de eventos mutuamente excluyentes, como aprobar / reprobar, verdadero / falso, 0 / 1, etc. La relación entre las variables independientes y dependientes se representa mediante una curva sigmoidea, y la probabilidad de que ocurra el evento es un valor entre 0 y 1.

Ejemplo de regresión logística: al examinar las características de los visitantes, como los sitios de donde provienen, el número de visitas a su sitio y la actividad en su sitio (variables independientes), es posible calcular la probabilidad de que un visitante acepte una oferta en su sitio web (variable dependiente). Esto puede ayudar a tomar decisiones más informadas sobre si promover la oferta en su sitio (Figura 44).

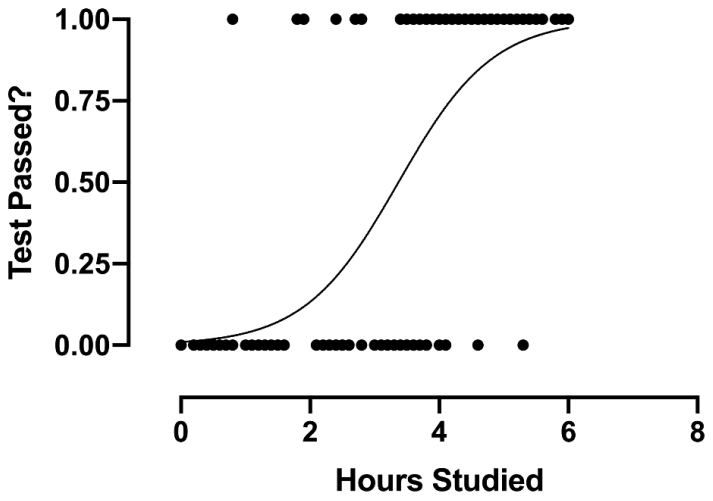


Figura 44. Ejemplo de visitas al sitio (fuente abierta)

3. Regresión Polinómica. El análisis de regresión polinómica es un método de representación de la relación no lineal entre variables dependientes e independientes. Es una variación del modelo de regresión lineal múltiple, pero la línea de ajuste óptima es curva en lugar de recta. Hay varios tipos de regresión polinómica, incluyendo lineal, cuadrática, cúbica y polinomios de

orden superior. Cada tipo de regresión polinómica se utiliza para modelar diferentes tipos de relaciones entre las variables dependientes e independientes.

Ejemplo de regresión polinómica: la relación entre la edad de una persona y su altura, donde la ecuación tomaría la forma de una ecuación cuadrática (Figura 45).

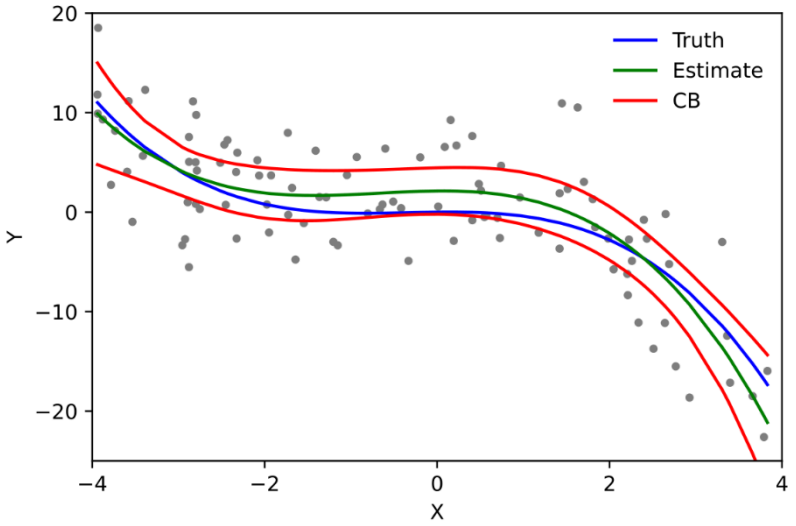


Figura 45. Ejemplo de conexión entre edad y altura (fuente abierta)

4. Regresión Ridge. Cuando los datos muestran multicolinealidad, la técnica de regresión ridge se utiliza para tratar la alta correlación entre las variables independientes. Aunque las estimaciones de mínimos cuadrados son precisas en esta situación, sus grandes varianzas pueden causar que el valor observado sea diferente del valor real. La regresión ridge reduce los errores estándar ajustando las estimaciones de regresión estimadas.

5. Regresión Lasso. La técnica lasso (Least Absolute Shrinkage and Selection Operator), al igual que la regresión ridge, penaliza el tamaño del coeficiente de regresión. Además, la regresión lasso utiliza la selección de variables, lo que resulta en la reducción de los valores de los coeficientes a cero.

Ejemplo de regresión ridge o lasso: un modelo que se utiliza para predecir el precio de una casa basado en su tamaño, ubicación y otros factores. El

modelo puede identificar qué características son más importantes para predecir el precio de la casa y luego usar esas características para hacer la predicción (Figura 46).

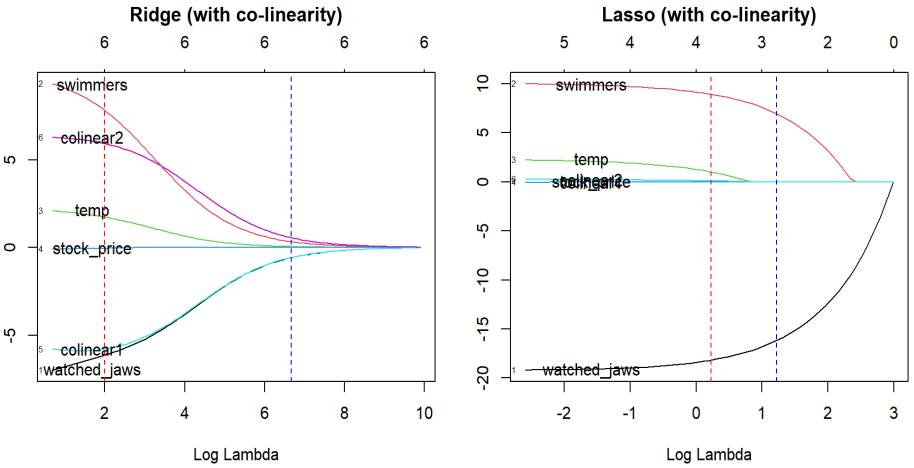


Figura 46. Ejemplo de predicción de precios (fuente abierta)

6. Regresión de Cuantiles. La regresión de cuantiles es un tipo de regresión lineal que se utiliza cuando los datos no cumplen con los requisitos de la regresión lineal o cuando hay valores atípicos presentes. Es una técnica común en estadísticas y econometría.

Ejemplo de regresión de cuantiles: predecir el salario mediano de un grupo de personas basado en su nivel educativo, edad y experiencia (Figura 47).

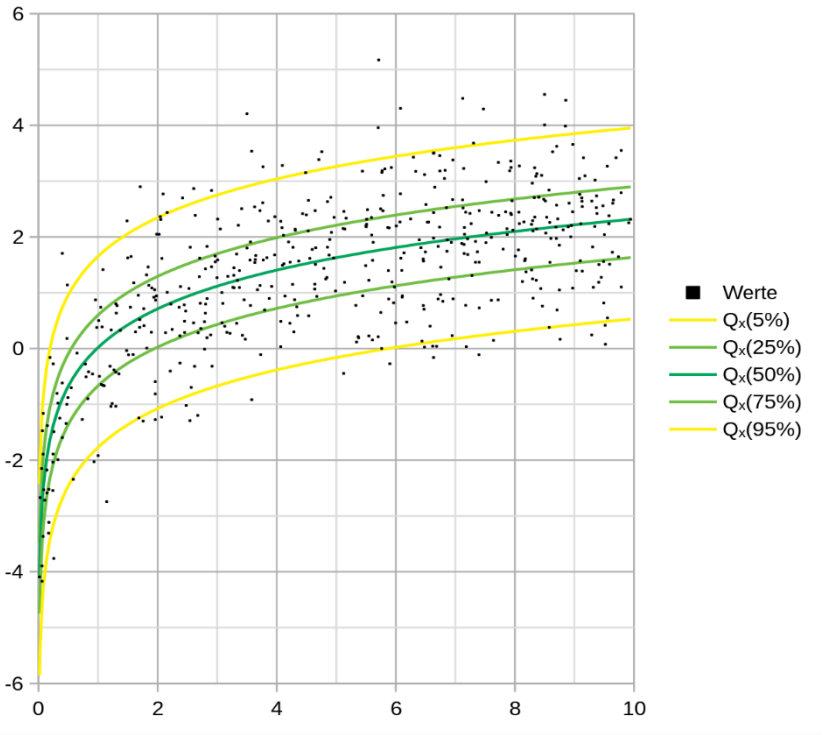


Figura 47. Ejemplo de predicción del salario mediano (fuente abierta)

7. Regresión de Red Elástica. La regresión de red elástica es una combinación de las regresiones ridge y lasso, que es beneficiosa cuando se trabaja con datos que tienen fuertes correlaciones. Regulariza los modelos de regresión utilizando las penalizaciones de ambas regresiones.

Ejemplo de regresión de red elástica: un modelo que predice el éxito de una campaña de marketing basado en la demografía del cliente, el rendimiento de campañas anteriores y otros factores (Figura 48). El modelo utilizaría las técnicas de regresión ridge y lasso para regularizar el modelo y reducir el riesgo de sobreajuste.

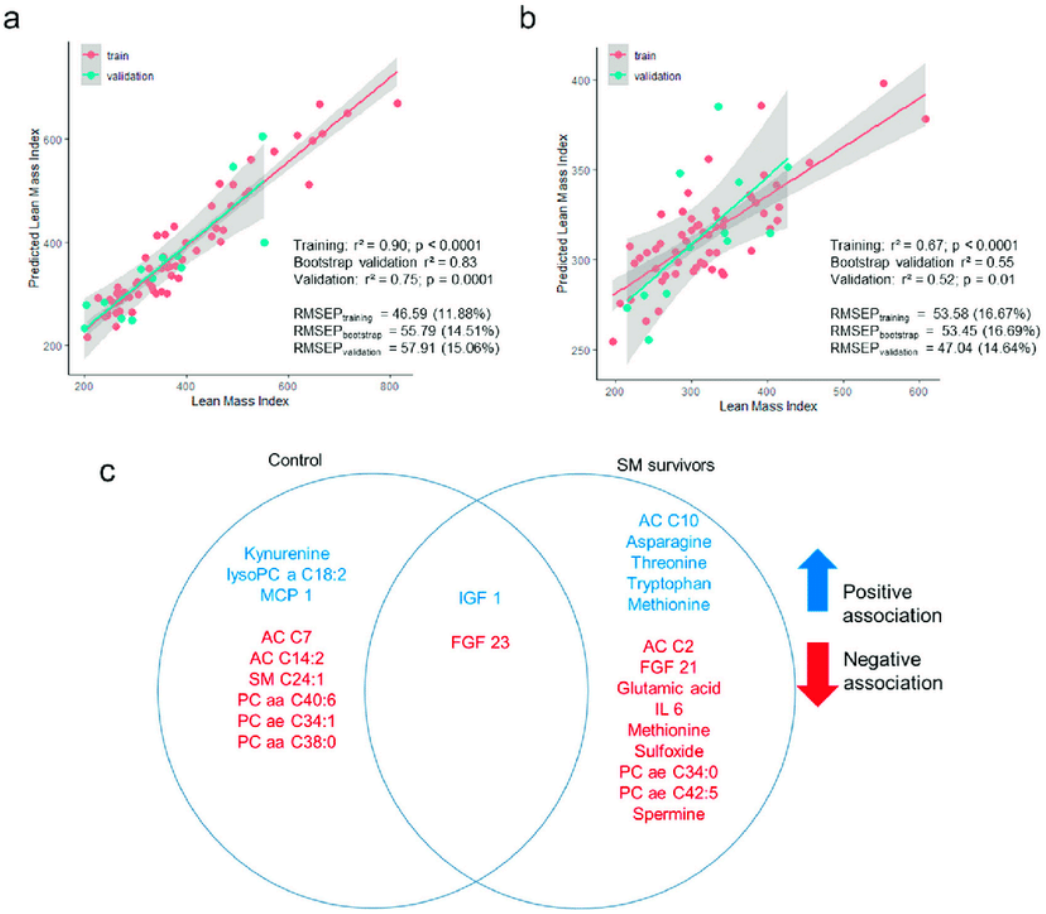
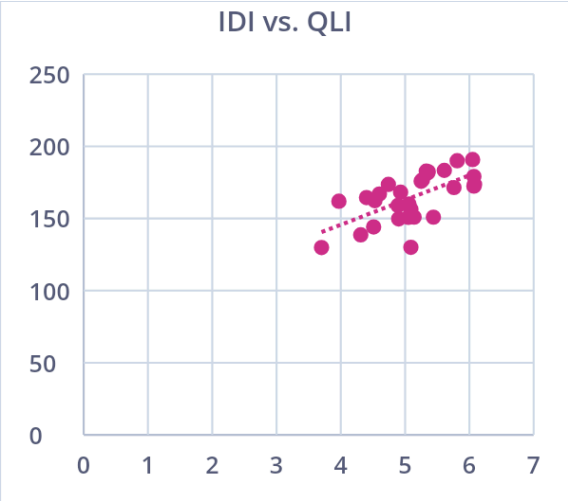


Figura 48. Ejemplo de predicción de campaña de marketing (fuente abierta)

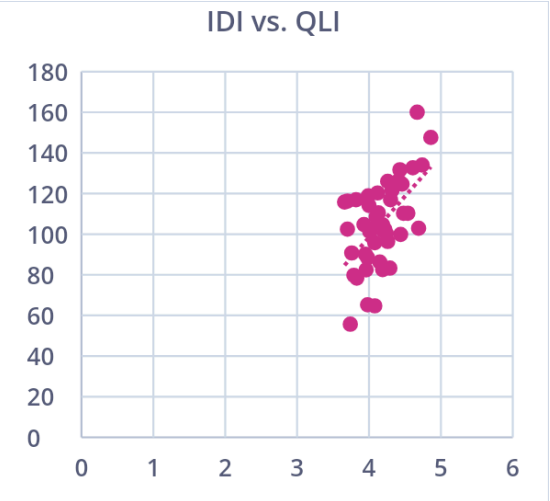
El **análisis de regresión** es una técnica de aprendizaje automático que puede utilizarse para identificar la relación entre dos variables, una dependiente y una independiente, y para medir el impacto de la variable independiente en la variable dependiente. Ofrece dos beneficios principales: determinar la conexión entre las dos variables y medir la fuerza de la influencia de la variable independiente en la variable dependiente. Se pueden utilizar varios modelos de regresión, como la regresión ecológica, la regresión por pasos, la regresión jackknife y la regresión robusta, para obtener más precisión a partir de los datos.

En la investigación empresarial y económica, el método más utilizado es la regresión lineal, que demuestra bastante bien la dependencia del cambio de una cantidad en el cambio de otra. Por ejemplo, según los datos de las figuras a continuación, hay una correlación directa entre el Índice de

Desarrollo Inclusivo (IDI) y el Índice de Calidad de Vida (QLI), lo que significa que un nivel más alto de inclusión económica correlaciona con un nivel más alto de satisfacción con la vida en todo el mundo. Sin embargo, debe tenerse en cuenta que esta correlación es más significativa en los países desarrollados que en los países en desarrollo. Esto puede explicarse por el entorno institucional establecido y las orientaciones de valores de las personas que viven en países altamente desarrollados (Figura 49).



The relationship between IDI and QLI in developed countries



The relationship between IDI and QLI in developing countries

Figura 49. Análisis de correlación

En Excel, para construir estos gráficos de regresión, es necesario seleccionar dos columnas de datos, seleccionar un gráfico de dispersión en la pestaña "Insertar" y construir la dependencia. Cuanto mayor sea el conjunto de datos, más relevantes serán los resultados del análisis (Figuras 50-52).

	A	B	C	D	E
1	Developed countries				
2	Country	Inclusive Development Index 2018	Gini Index	Ranking of happiness 2019-2021	Quality of Life Index b
3	Norway	6,08	27,7	7,365	173,57
4	Iceland	6,07	26,1	7,557	179,1
5	Luxembourg	6,07	34,2	7,404	172,45
6	Switzerland	6,05	33,1	7,512	190,82
7	Denmark	5,81	27,7	7,636	190,01
8	Sweden	5,76	29,3	7,384	171,4
9	Netherlands	5,61	29,2	7,415	183,31
10	Ireland	5,44	30,6	7,041	150,89
11	Australia	5,35	34,3	7,162	181,52
12	Austria	5,36	30,2	7,163	182,37
13	Finland	5,33	27,7	7,821	182,79
14	Germany	5,27	31,7	7,034	176,76
15	New Zealand	5,25	32,5	7,2	175,77
16	Belgium	5,14	27,2	6,805	150,89
17	Czech Republic	5,09	25,3	6,92	156,33
18	South Korea	5,09	31,4	5,935	130,02
19	Canada	5,06	33,3	7,025	159,99
20	France	5,05	32,4	6,687	150,73
21	Slovenia	4,93	24,4	6,63	168,2
22	Slovakia	4,9	23,2	6,391	149,68
23	United Kingdom	4,89	35,1	6,943	158,99
24	Estonia	4,74	30,8	6,341	173,56
25	United States	4,6	41,5	6,977	166,98
26	Japan	4,53	32,9	6,039	162,32
27	Israel	4,51	38,6	7,364	144,14
28	Spain	4,4	34,3	6,476	164,48
29	Italy	4,31	35,2	6,467	138,63

Figura 50. Pantalla de ejemplo 32

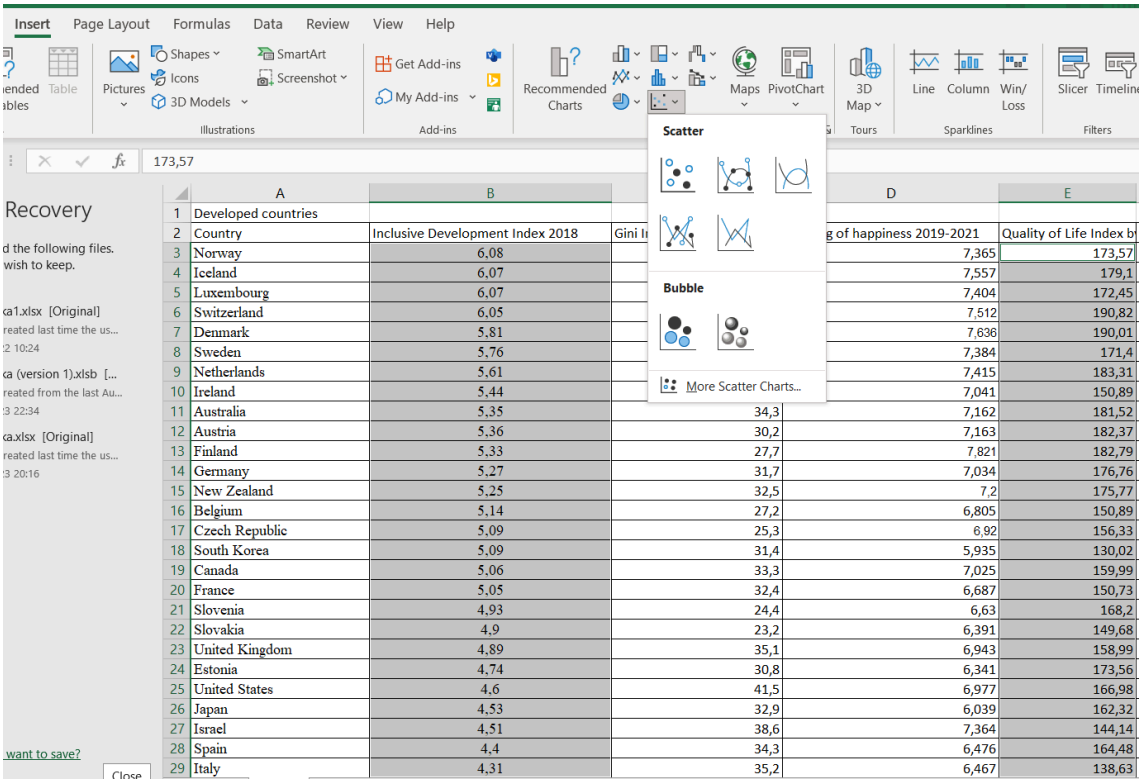


Figura 51. Pantalla de ejemplo 33

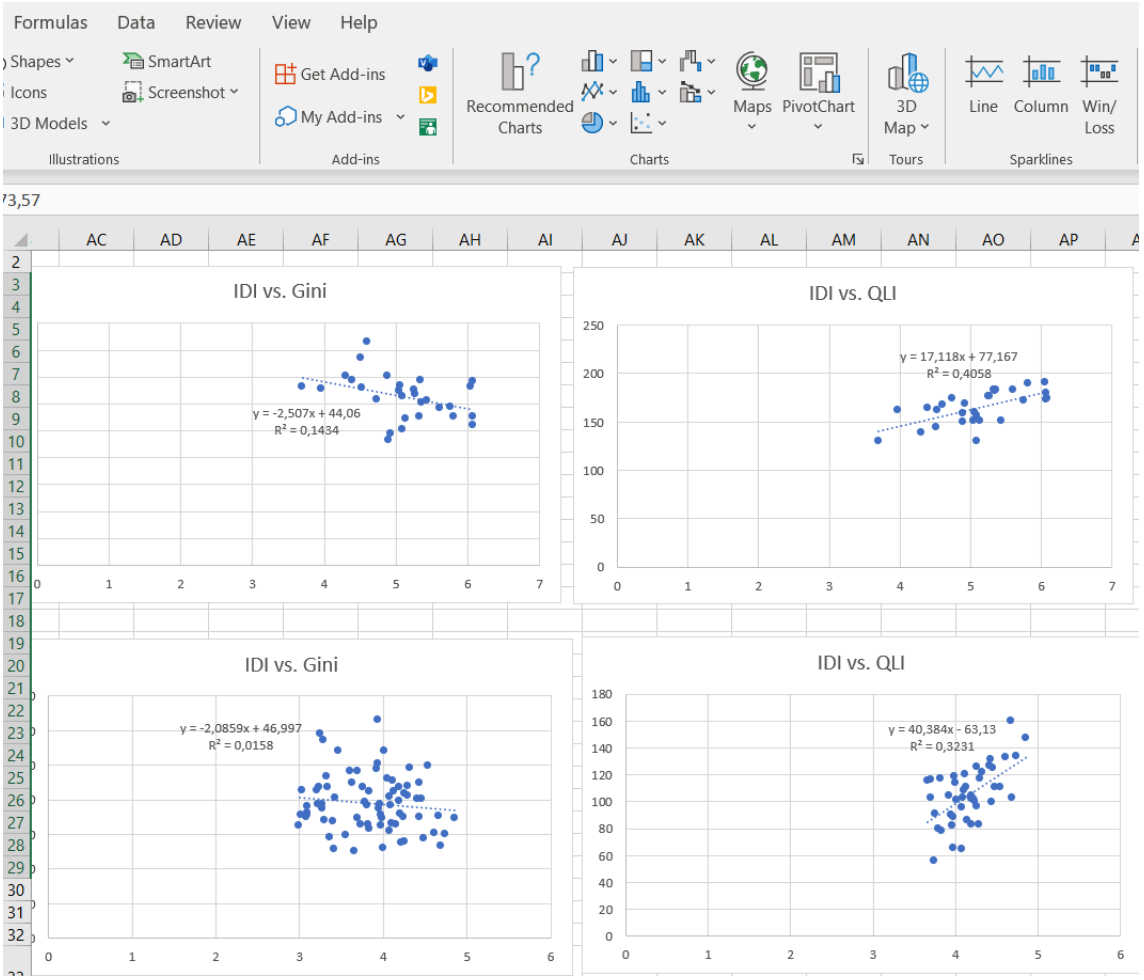


Figura 52. Pantalla de ejemplo 34

9. Common Tests of Significance (Pruebas comunes de significancia)

Las pruebas de significancia estadística se emplean para calcular la probabilidad de que la correlación observada en los datos sea simplemente una coincidencia; es decir, que las variables no estén conectadas en la población general. Estas pruebas pueden utilizarse para eliminar hipótesis que son improbables de ser verdaderas. Además, proporcionan una medida estándar que es fácilmente comprendida por muchas personas y

pueden utilizarse para comparar los resultados de un proyecto de investigación con otro (Kindness, Kvilhaug 2023).

Es beneficioso utilizar tanto pruebas de significancia estadística como medidas de asociación al analizar datos. Las primeras pueden determinar la probabilidad de que exista una relación, mientras que las segundas pueden medir la intensidad y dirección de la relación. Utilizar ambos métodos juntos puede proporcionar los resultados más completos.

Una prueba de significancia es un proceso formal utilizado para comparar datos observados con una hipótesis que se está evaluando por su precisión. La hipótesis es una declaración sobre un parámetro, como la proporción de la población (p) o la media de la población (μ). Los resultados de la prueba se expresan en términos de una probabilidad, que indica cuán bien coinciden los datos y la hipótesis (Kindness, Kvilhaug 2023).

Cuando se hipotetiza una relación entre dos variables, deben hacerse dos preguntas: ¿cuál es la probabilidad de que exista la relación y, si es así, cuán fuerte es? Para responder a estas preguntas, se utilizan dos tipos de herramientas: pruebas de significancia estadística y medidas de asociación.

Las pruebas de significancia estadística se emplean para determinar la probabilidad de que la relación observada entre dos variables sea simplemente una coincidencia. Si se tomaran múltiples muestras de la misma población, ¿se observaría la misma relación en cada muestra? Si se realizara un censo de la población, ¿se vería la misma relación en la población de la que se tomó la muestra? ¿O es el resultado simplemente una cuestión de azar?

Las pruebas de significancia estadística nos permiten calcular la probabilidad de que la relación que hemos observado sea simplemente el resultado del azar. También nos ayudan a determinar la probabilidad de cometer un error si asumimos que existe una relación. Aunque es imposible estar completamente seguro de que dos variables estén relacionadas, la teoría de la probabilidad y la curva normal pueden utilizarse para estimar la probabilidad de estar equivocado si asumimos que existe una relación. Utilizando la teoría de la probabilidad y la curva

normal, podemos calcular la probabilidad de estar incorrectos si asumimos que la relación que hemos encontrado es precisa. Si la probabilidad de estar equivocados es baja, entonces podemos concluir que nuestra observación de la relación es un hallazgo estadísticamente significativo.

La significancia estadística indica que es muy probable que exista una conexión entre dos variables. Sin embargo, la significancia estadística no es lo mismo que la significancia práctica. Es posible tener un resultado estadísticamente significativo, pero las implicaciones de ese resultado pueden no tener ningún uso práctico. Por lo tanto, es esencial que los investigadores consideren tanto la significancia estadística como la práctica de cualquier hallazgo de investigación. Podemos descubrir que hay una correlación estadísticamente significativa entre la edad de una persona y su satisfacción con los servicios de recreación de la ciudad, con los ciudadanos mayores siendo un 3% menos satisfechos que los ciudadanos más jóvenes. Sin embargo, ¿es esta diferencia del 3% lo suficientemente significativa como para ser motivo de preocupación?

Los pasos en las pruebas de significancia estadística se descubren en la tabla a continuación e incluyen algunos elementos (Kindness, Kvilhaug 2023):

1) Declarar la hipótesis de investigación;

2) Declarar la hipótesis nula;

3) Seleccionar un nivel de probabilidad de error (nivel alfa);

4) Seleccionar y calcular la prueba de significancia estadística;

5) Interpretar los resultados.

№	Pasos	Significado	Ejemplos
1.	Declarar la hipótesis de investigación	Una hipótesis de investigación describe la relación prevista entre dos variables, que puede expresarse de manera general o con detalles sobre la dirección y el tamaño de la relación	General: la duración del programa de capacitación laboral está relacionada con la tasa de colocación laboral de los capacitados. Dirección: cuanto más largo sea el programa de capacitación, mayor será

Co-funded by the European Union

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

			la tasa de colocación laboral de los capacitados
2.	Declarar la hipótesis nula	La hipótesis nula típicamente sugiere que no hay conexión entre las dos variables. La hipótesis nula puede sugerir que no hay una relación significativa entre las variables propuestas en la hipótesis de investigación	Los programas de capacitación más largos colocarán la misma cantidad o menos capacitados en empleos que los programas más cortos
3.	Seleccionar un nivel de probabilidad de error (nivel alfa)	En cualquier proyecto de investigación, existe la posibilidad de que el investigador interprete incorrectamente la relación entre dos variables. Estos errores pueden clasificarse en dos tipos: Error Tipo I, que ocurre cuando el investigador cree erróneamente que existe una relación cuando no la hay, y Error Tipo II, que ocurre cuando el investigador cree erróneamente que no existe una relación cuando sí la hay	Hipótesis de investigación: el nuevo medicamento es mejor para tratar ataques cardíacos que el medicamento antiguo. Hipótesis nula: el nuevo medicamento no es mejor para tratar ataques cardíacos que el medicamento antiguo
4.	Seleccionar y calcular la prueba de significancia estadística	Chi Cuadrado se emplea como una prueba de significancia estadística para datos nominales y ordinales	Se puede hipotetizar que existe una relación entre el tipo de programa de capacitación asistido y el éxito de colocación laboral de los capacitados
5.	Interpretar los resultados	Es esencial tener en cuenta que las pruebas de significancia estadística no proporcionan información sobre la intensidad de la conexión entre dos variables, la dirección de la asociación, la probabilidad de un error Tipo I, la fiabilidad y precisión de la investigación, o evidencia absoluta y definitiva de una relación	Las pruebas de significancia estadística se utilizan para proporcionar una medida ampliamente comprendida de un proyecto de investigación, que luego puede compararse con los resultados de otros estudios. Esto permite la comunicación de

			información clave sobre el proyecto

Figura 53. Significancia estadística (fuente abierta)

10. Remuestreo

El remuestreo en estadística es un conjunto de técnicas utilizadas para estimar la precisión de las estadísticas de muestra, como medianas, diferencias y percentiles, utilizando subconjuntos de datos disponibles o extrayendo aleatoriamente con reemplazo de un conjunto de puntos de datos. Además, el remuestreo puede utilizarse para realizar pruebas de permutación y validar modelos mediante subconjuntos aleatorios.

El remuestreo es un conjunto de métodos utilizados en estadísticas para obtener más información sobre una muestra. Esto puede involucrar repetir el proceso de muestreo o evaluar su precisión. Al utilizar estas técnicas adicionales, el remuestreo puede a menudo mejorar la precisión general y estimar cualquier incertidumbre dentro de una población. El muestreo es el acto de seleccionar ciertos grupos de una población para recopilar datos. El remuestreo generalmente implica llevar a cabo procedimientos de prueba similares con tamaños de muestra de ese grupo. Esto puede significar probar la misma muestra nuevamente o seleccionar nuevas muestras.

El remuestreo implica técnicas como el bootstrap y las pruebas de permutación. En cuanto al muestreo, hay cuatro enfoques principales: muestreo aleatorio simple (donde cada persona o pieza de datos tiene la misma posibilidad de ser elegida), muestreo sistemático (donde se asignan números o valores a las personas y se utilizan intervalos para dividir el grupo), muestreo estratificado (donde la población se divide en subgrupos según ciertas cualidades) y muestreo por conglomerados (donde se seleccionan grupos al azar, lo que a menudo lleva a resultados variados) - Figura 54.

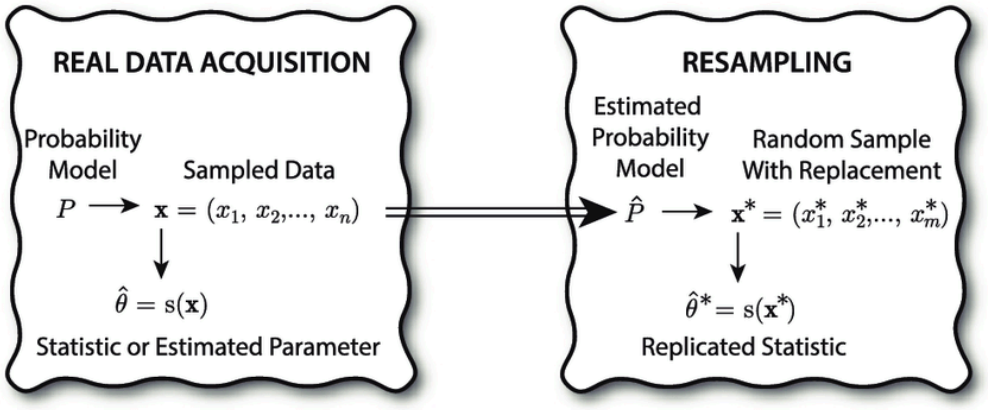


Figura 54. Explicación del remuestreo (fuente abierta)

Hay varios tipos de remuestreo (Remuestreo – estadísticas):

1. **Bootstrap**. La técnica bootstrap implica tomar múltiples muestras de la misma población y repetir las mismas observaciones. Por ejemplo, si seleccionas 10 personas de un grupo de 100 para observar una hipótesis, podrías hacer esto múltiples veces, seleccionando cada vez a 10 personas diferentes. Esto ayuda a reducir posibles errores estadísticos, ya que puedes calcular la media o mediana entre las muestras para obtener resultados más precisos. Este método, a menudo denominado el método de plug-in, se utiliza ampliamente en campos como la física y los algoritmos genéticos.

2. **Jackknife**. El jackknife es un método de remuestreo que se utiliza para detectar varianza o sesgo en una muestra. Implica sacar una observación de un grupo para formar un subconjunto. Este proceso se repite con cada observación siendo removida una a la vez para determinar si hay valores atípicos en la muestra. Por ejemplo, si hay 10 observaciones numeradas del uno al diez, una puede ser removida y los resultados observados. Después de eso, se pueden quitar dos números y verificar los números restantes hasta 10 para ver si alguno de ellos destaca.

3. **Validación cruzada.** La validación cruzada es una técnica a menudo utilizada por los estadísticos para modelos estadísticos predictivos. Este método implica reservar una porción de los datos dentro de una muestra como el conjunto de validación, mientras que los datos restantes se utilizan

como el conjunto de entrenamiento. El conjunto de entrenamiento se prueba para hacer predicciones sobre el conjunto de validación y se mide la precisión de las predicciones tomando la media de los resultados de cada validación cruzada.

4. **Prueba de permutación.** Las pruebas de permutación implican ejecutar repetidamente una prueba exacta con una hipótesis nula para crear una muestra de una población y observar los mismos resultados. Este método de prueba puede determinar la intercambiabilidad de diferentes observaciones o la probabilidad de intercambiar etiquetas dentro de un conjunto.

11. Comparaciones múltiples

Cuando se realizan múltiples inferencias estadísticas al mismo tiempo o cuando se elige un subconjunto de parámetros basados en valores observados, esto se conoce como el problema de las comparaciones múltiples (multiplicidad o pruebas múltiples) en estadísticas. A medida que se realizan más inferencias, la probabilidad de hacer inferencias incorrectas aumenta. Para contrarrestar esto, se han creado varios métodos estadísticos, generalmente estableciendo un nivel de significancia más alto para comparaciones individuales para equilibrar el número de inferencias que se están realizando (Problema de Comparaciones Múltiples).

En la década de 1950, los estadísticos J. Tukey y H. Scheffé dieron más atención al problema de las comparaciones múltiples. En 1996, se llevó a cabo en Israel la primera conferencia internacional sobre procedimientos de comparación múltiple.

Realizaremos un experimento con una sola moneda lanzándola 10 veces. Suponemos que la moneda es justa, lo que significa que tiene un 50% de posibilidades de mostrar cara en un solo lanzamiento. Declararemos que la moneda es injusta si sale 9 caras y 1 cruz, pero es poco probable que esto ocurra. Por lo tanto, es probable que encontremos que la moneda es justa.

Realizaremos 5 pruebas lanzando 5 monedas 10 veces cada una. Si observamos 9 caras y 1 cruz en cualquiera de las monedas, podemos asumir que la moneda es injusta, aunque todas las 5 monedas sean justas. La probabilidad de que esto ocurra para una sola moneda es baja, pero después de realizar 5 pruebas, es más probable que al menos una de las monedas muestre 9 caras y 1 cruz debido al azar.

El ejemplo del lanzamiento de la moneda demuestra que cuantas más pruebas realizamos, mayor es la probabilidad de detectar una correlación o efecto que puede no existir realmente; esto se conoce como el problema de las comparaciones múltiples (Figura 55). Además, aumenta la probabilidad de creer erróneamente que un efecto está presente cuando es debido al azar.

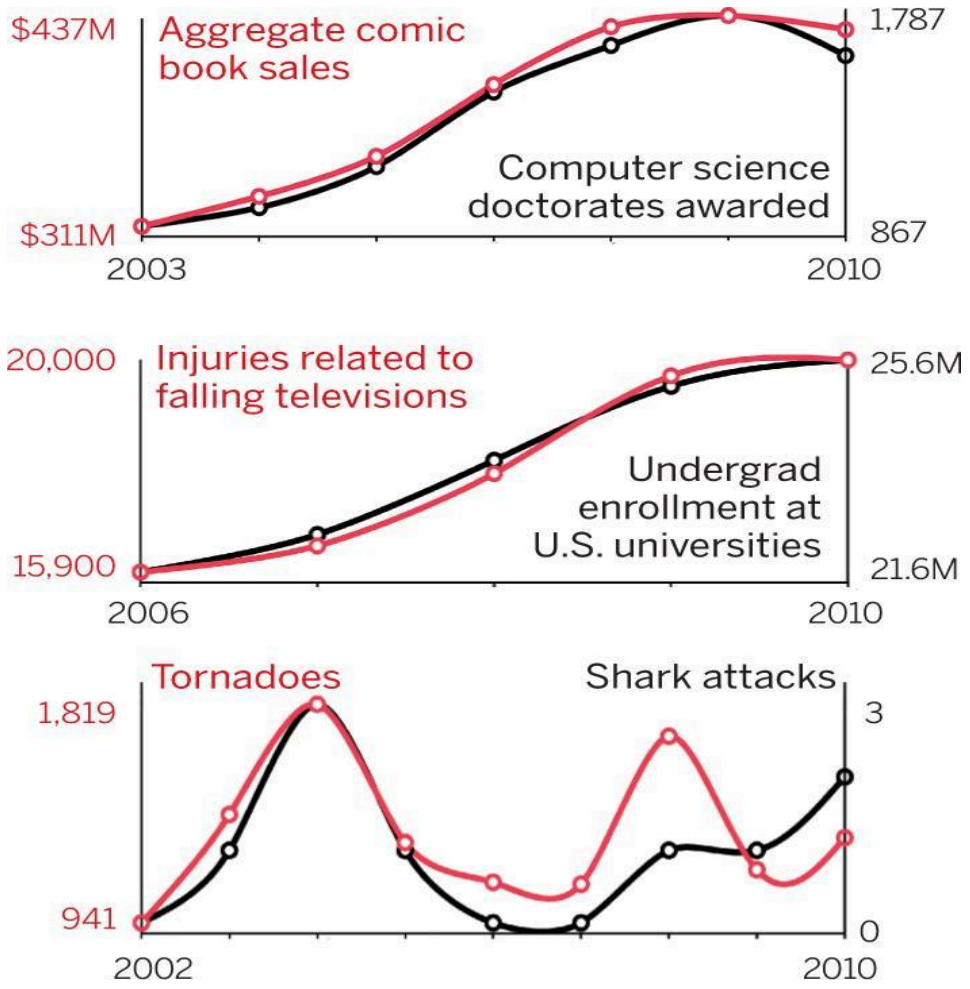


Figura 55. Problema de comparaciones múltiples (fuente abierta)

12. Skewness y Kurtosis

En muchos análisis estadísticos, es esencial identificar el valor promedio y la dispersión de un conjunto de datos. Además, el sesgo y la curtosis pueden utilizarse para describir aún más los datos.

El sesgo es una medida de la asimetría de una distribución. El sesgo es una indicación de cuán asimétrica es una distribución o un conjunto de datos (Sharma 2023). Si es simétrica, parecerá la misma cuando se vea desde el punto central a ambos lados. Una distribución es asimétrica cuando su lado izquierdo y derecho no son imágenes especulares.

Una distribución normal tiene un sesgo de cero y cualquier dato simétrico debería tener un sesgo cercano a cero. Los datos que están sesgados hacia la izquierda tienen un sesgo negativo, mientras que los datos que están sesgados hacia la derecha tienen un sesgo positivo. Esto significa que la cola izquierda es más larga que la cola derecha. Si los datos tienen múltiples picos, esto puede influir en la dirección del sesgo, ya que una distribución sesgada hacia la derecha tiene una cola derecha más larga en comparación con la cola izquierda (Figura 56).

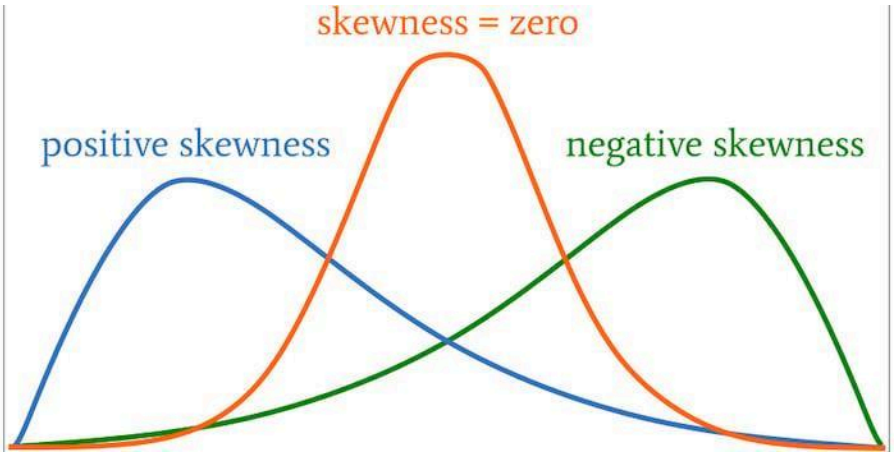


Figura 56. Ejemplos de sesgo (fuente abierta)

Por qué es importante conocer el sesgo de los datos (Gawali 2023; Sharma 2023)?

En primer lugar, para crear mejores modelos lineales, es beneficioso comprender el sesgo de los datos, ya que los modelos lineales asumen que la distribución de la variable independiente y la variable objetivo son similares.

En segundo lugar, examinemos la distribución de la potencia de los autos. Es evidente que la distribución está sesgada positivamente. Si queremos usar esto como una característica para un modelo que predice el mpg de un auto, podemos hacerlo. Los datos que tenemos están sesgados en una dirección positiva, lo que significa que hay más autos con menor potencia que con mayor potencia (Figura 57). Por lo tanto, cuando usamos estos datos para entrenar nuestro modelo, será más preciso al predecir el mpg de autos con menor potencia que aquellos con mayor potencia.

En tercer lugar, el sesgo de nuestra distribución indica que los valores atípicos se encuentran principalmente en el lado derecho del gráfico. El sesgo es positivo, lo que significa que la cola derecha del gráfico es más larga que la izquierda.

La curtosis es una forma de determinar si los datos están concentrados alrededor de la media más que una distribución normal (Sharma 2023). Los datos con alta curtosis generalmente tienen más valores atípicos, mientras que los datos con baja curtosis tienen menos valores atípicos. Una distribución uniforme sería el ejemplo más extremo de esto.

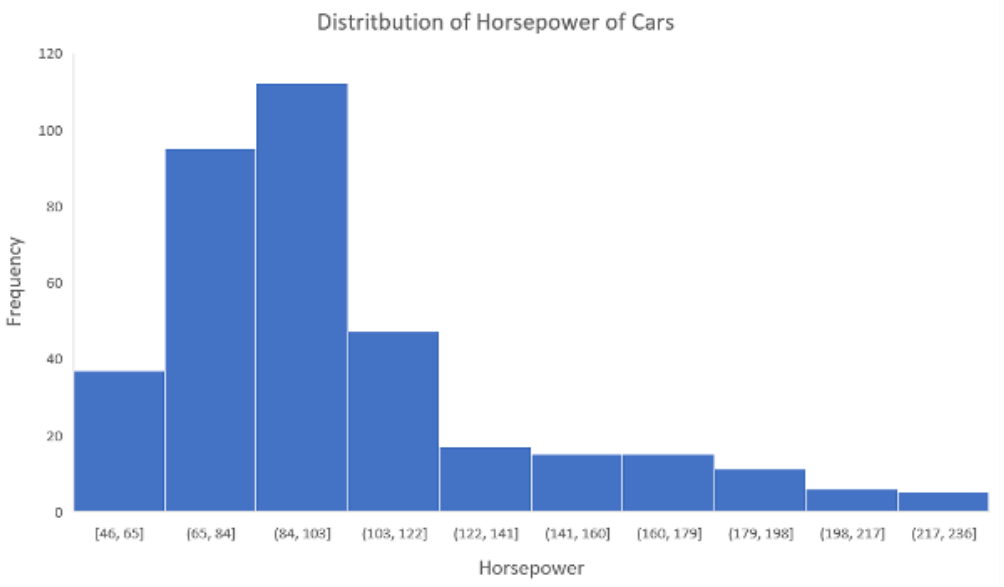


Figura 57. Distribución de la potencia de los autos (fuente abierta)

La distribución normal estándar tiene una curtosis de cero, lo cual se logra utilizando la curtosis en exceso (Gawali 2023). Además, una distribución con curtosis positiva se dice que tiene una forma "de cola pesada", mientras que una distribución con curtosis negativa se dice que tiene una forma "de cola ligera".

Los datos pueden tener un pico más plano, como si estuviera comprimido, lo que se conoce como Curtosis Negativa (Platykurtic). Si los datos tienen un pico más pronunciado, como si estuviera estirado, se llama Curtosis Positiva (Leptokurtic).

El valor de la curtosis para una distribución simétrica se espera que sea 3. Si la curtosis es mayor que 3, se conoce como Curtosis Positiva y el rango de valores para esto es de 1 al infinito. Por otro lado, si la curtosis es menor que 3, se conoce como Curtosis Negativa y el rango de valores para esto es de -2 al infinito. Cuanto mayor sea el valor de la curtosis, más pronunciado será el pico.

La mayoría de los programas de software estadístico de propósito general proporcionan acceso a los coeficientes de sesgo y curtosis.

12.1. Demostración: Sesgo y Curtosis

Excel proporciona las funciones SKEW y KURT para calcular el sesgo y la curtosis de S, es decir, si R es un rango en Excel que contiene los elementos de datos en S, entonces SKEW(R) = el sesgo de S y KURT(R) = la curtosis de S.

Supongamos que S = {2, 5, -1, 3, 4, 5, 0, 2}. El sesgo de S = -0.43, es decir, SKEW(R) = -0.43, donde R es un rango en una hoja de Excel que contiene los datos en S. Dado que este valor es negativo, la curva que representa la distribución está sesgada hacia la izquierda (es decir, la parte más gruesa de la curva está a la derecha). También SKEW.P(R) = -0.34. La curtosis de S = -0.94, es decir, KURT(R) = -0.94, donde R es un rango en una hoja de Excel que contiene los datos en S. La curtosis de la población es -1.114 (Figura 58).

	A	B	C	D	E	F	G
1	**Skewness and Kurtosis**						
2							
3				SKEW	-0.42705		=SKEW(B4:B11)
4		2		KURT	-0.93979		=KURT(B4:B11)
5		5					
6		-1		n	8		=COUNT(B4:B11)
7		3					
8		4		SKEWP	-0.3424		=SKEW.P(B4:B11)
9		5			-0.3424		=SKEWP(B4:B11)
10		0			-0.3424		=E3*(E6-2)/SQRT(E6*(E6-1))
11		2					
12				KURTP	-1.11419		=KURTP(B4:B11)
13					-1.11419		=(KURT(B4:B11)*(E6-2)*(E6-3)/(E6-1)-6)/(E6+1)

Figura 58. Pantalla de ejemplo 35

También es posible calcular SKEW y KURT utilizando el Análisis de datos en Excel (Figuras 59-60).

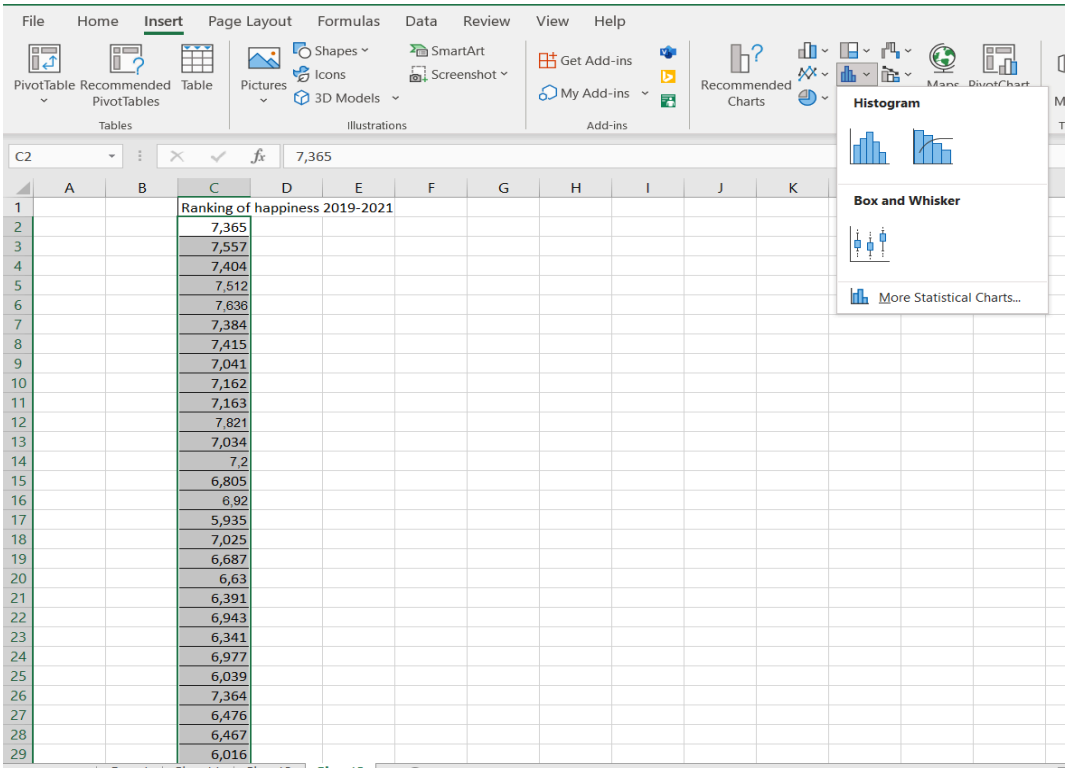


Figura 59. Pantalla de ejemplo 36

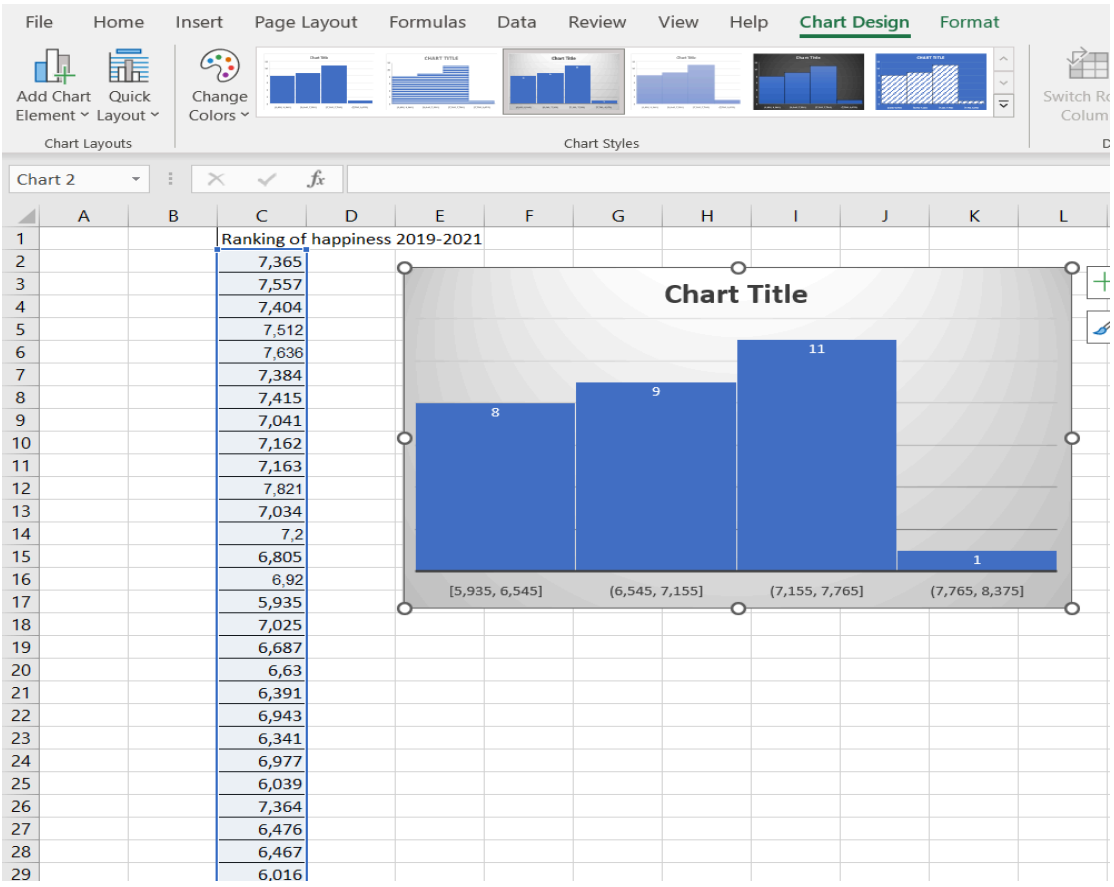


Figura 60. Pantalla de ejemplo 37

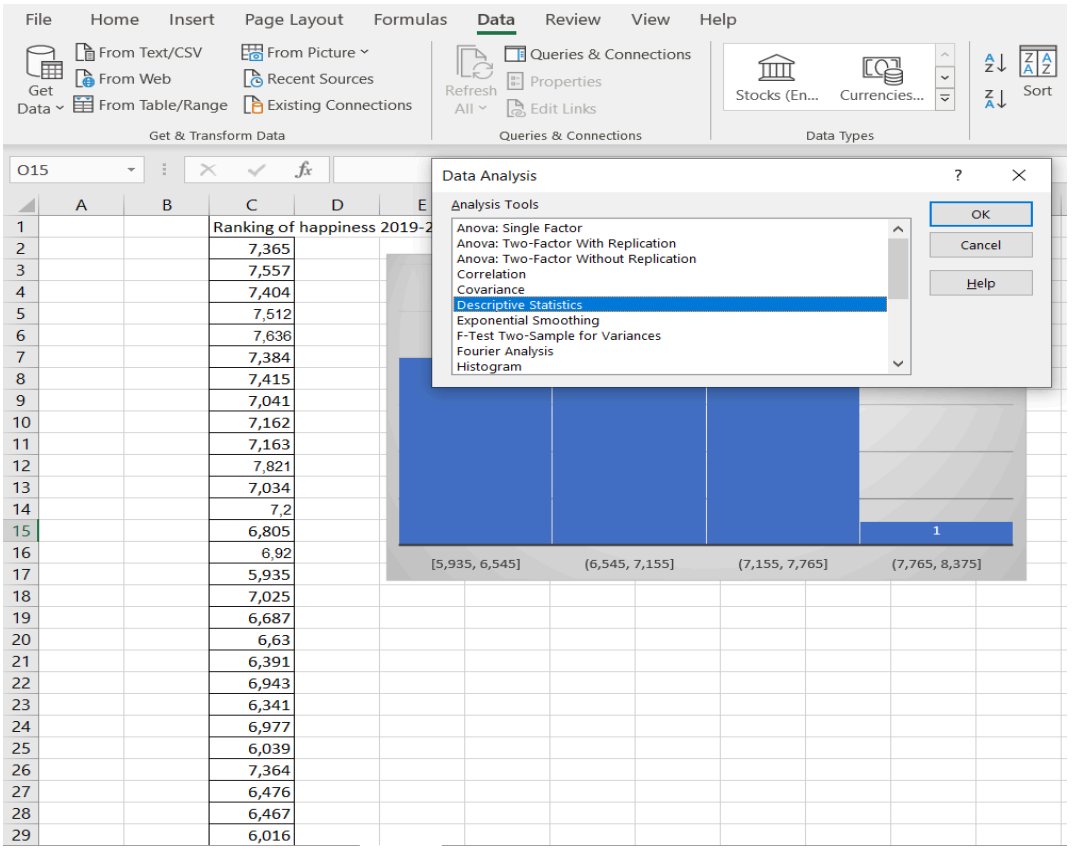


Figura 61. Pantalla de ejemplo 38

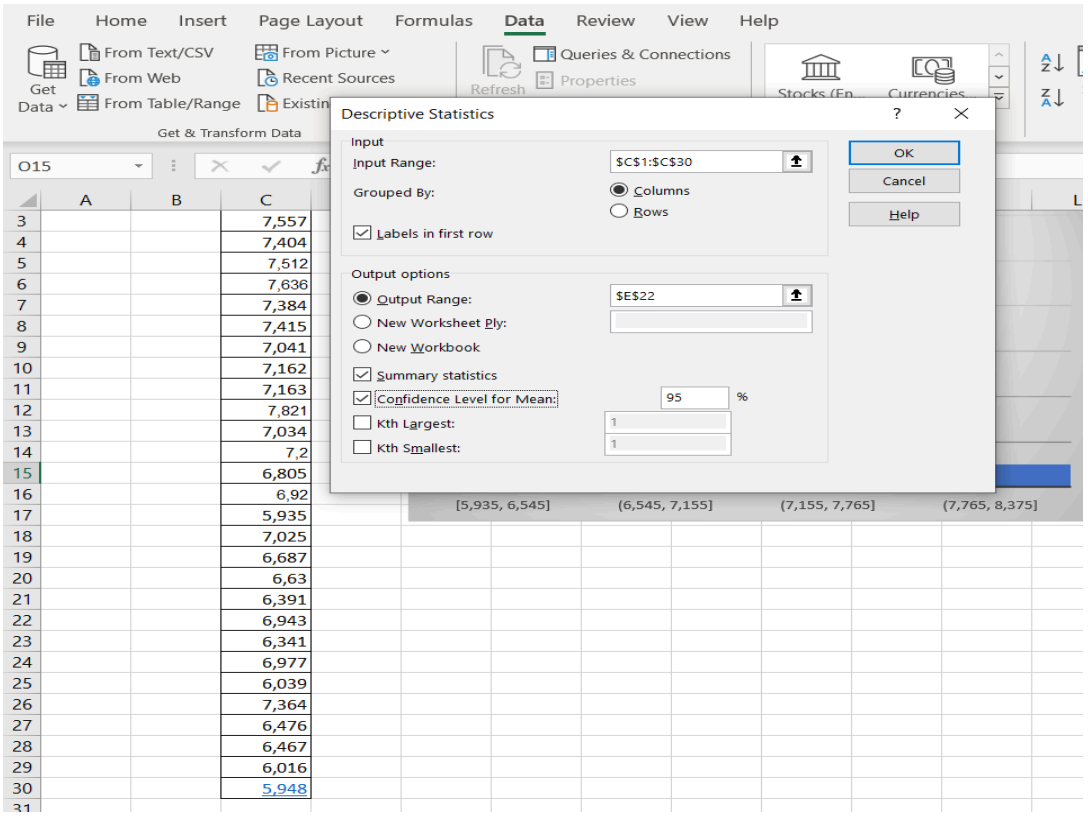


Figura 62. Pantalla de ejemplo 39

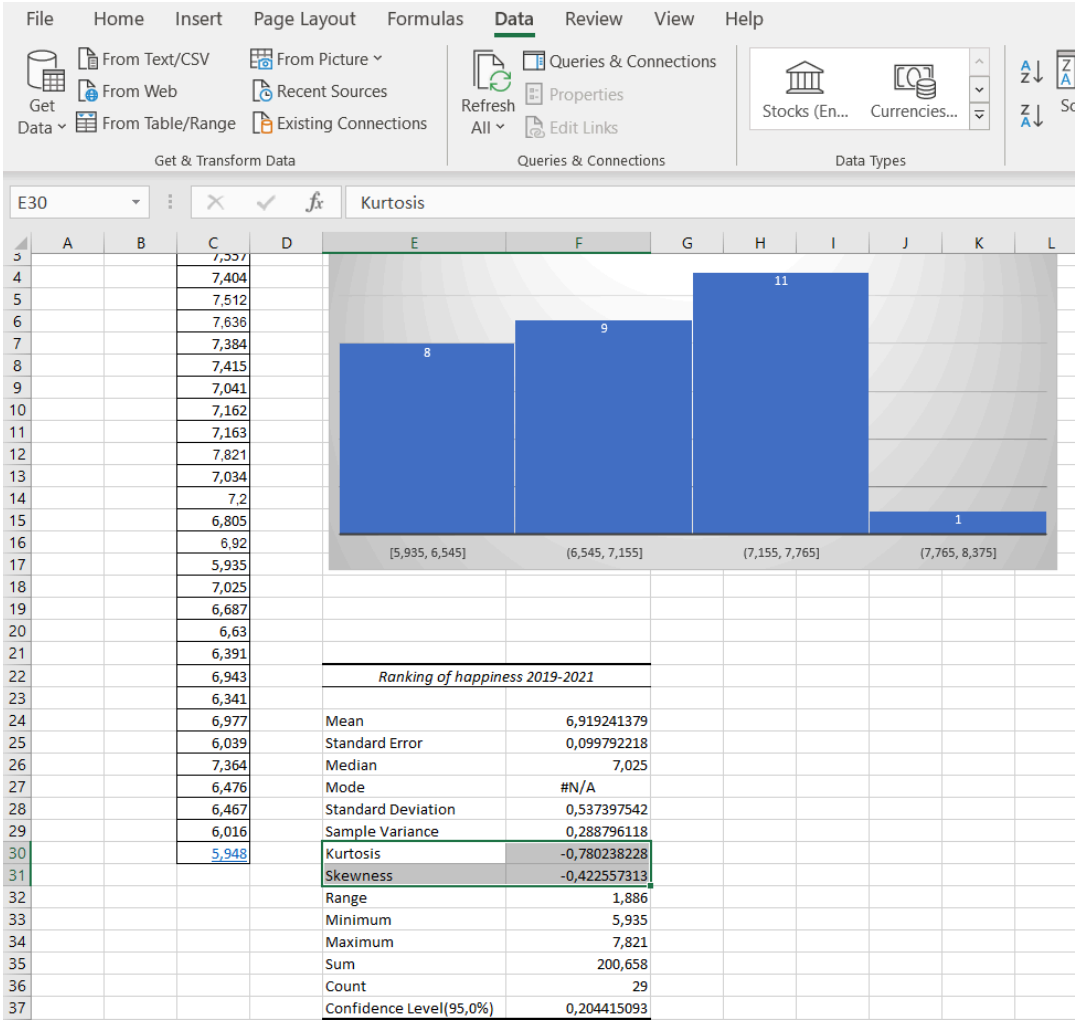


Figura 63. Pantalla de ejemplo 40

Referencias

- Barone, A., Brock, T., Schmitt, K.R. (2023). Sample Distribution: Definition, How It's Used, and Example. https://www.investopedia.com/terms/s/sampling-distribution.asp
- Descriptive Statistics. https://cleartax.in/glossary/descriptive-statistics
- Exploratory Data Analysis. https://www.itl.nist.gov/div898/handbook/eda/eda.htm

- Frequency Measurement. https://www.sciencedirect.com/topics/computer-science/frequency-measurement
- Gawali, S. (2023). Skewness and Kurtosis: Quick Guide. https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics
- Henrique, A. (2019). An Introduction to Linear Regression. https://medium.com/@alexandre.hsd/an-introduction-to-linear-regression-13527642f49
- Kindness, D., Kvilhaug, S. (2023). Statistical Significance Definition, Types, and How It's Calculated. https://www.investopedia.com/terms/s/statistical-significance.asp
- Linear regression. https://en.wikipedia.org/wiki/Linear_regression
- Manikandan, S. (2011). Measures of dispersion. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198538
- Measuring the Impacts of Business on Well-Being and Sustainability. https://www.oecd.org/statistics/Measuring-impacts-of-business-on-well-being.pdf
- Multiple Comparisons Problem. https://handwiki.org/wiki/Multiple_comparisons_problem
- Resampling – statistics. https://www.chemeurope.com/en/encyclopedia/Resampling_%28statistics%29.html
- Sharma, A. (2023). Understanding Skewness in Data and Its Impact on Data Analysis. https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics
- Sullivan, L. (2016). The Role of Probability. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability_print.html