Universidad de la Ciudad de Aguascalientes

SEMINARIO DE TESIS

Clase 11

Mentes que transforman el mundo

Universidad de la Ciudad de Aguascalientes

Análisis de datos cuantitativos

Mentes que transforman el mundo

Revisión del ejercicio anterior

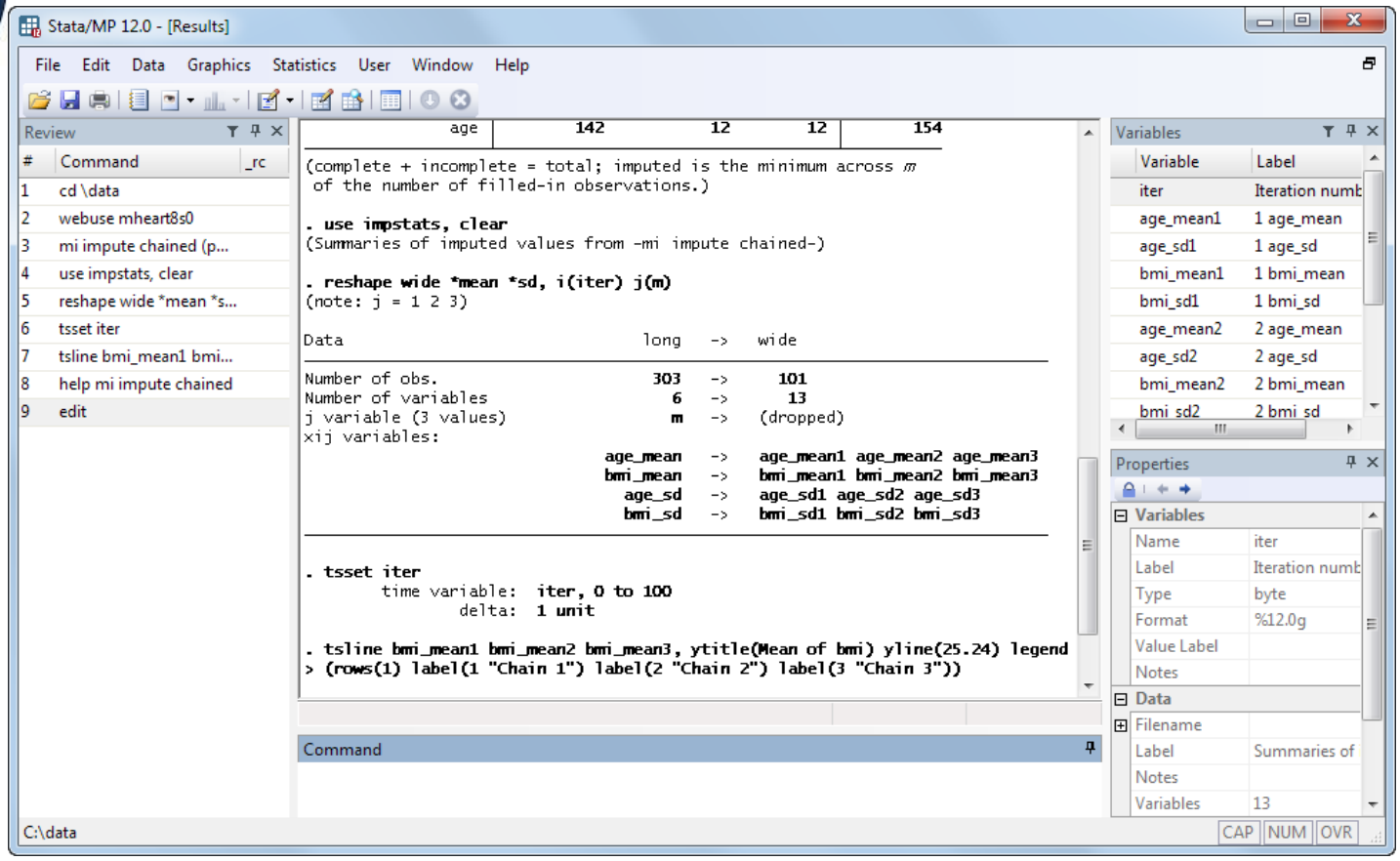
$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum S_i^2}{S_T^2}\right]$$

α:	Coeficiente de confiabilidad del cuestionario →	**0.75**
k:	Número de ítems del instrumento →	20
$\sum_{i=1}^{k} S_i^2$:	Sumatoria de las varianzas de los ítems. →	7.914
S_t^2:	Varianza total del instrumento. →	27.690

RANGO	CONFIABILIDAD
0.53 a menos	Confiabilidad nula
0.54 a 0.59	Confiabilidad baja
0.60 a 0.65	Confiable
0.66 a 0.71	Muy confiable
0.72 a 0.99	Excelente confiabilidad
1	Confiabilidad perfecta

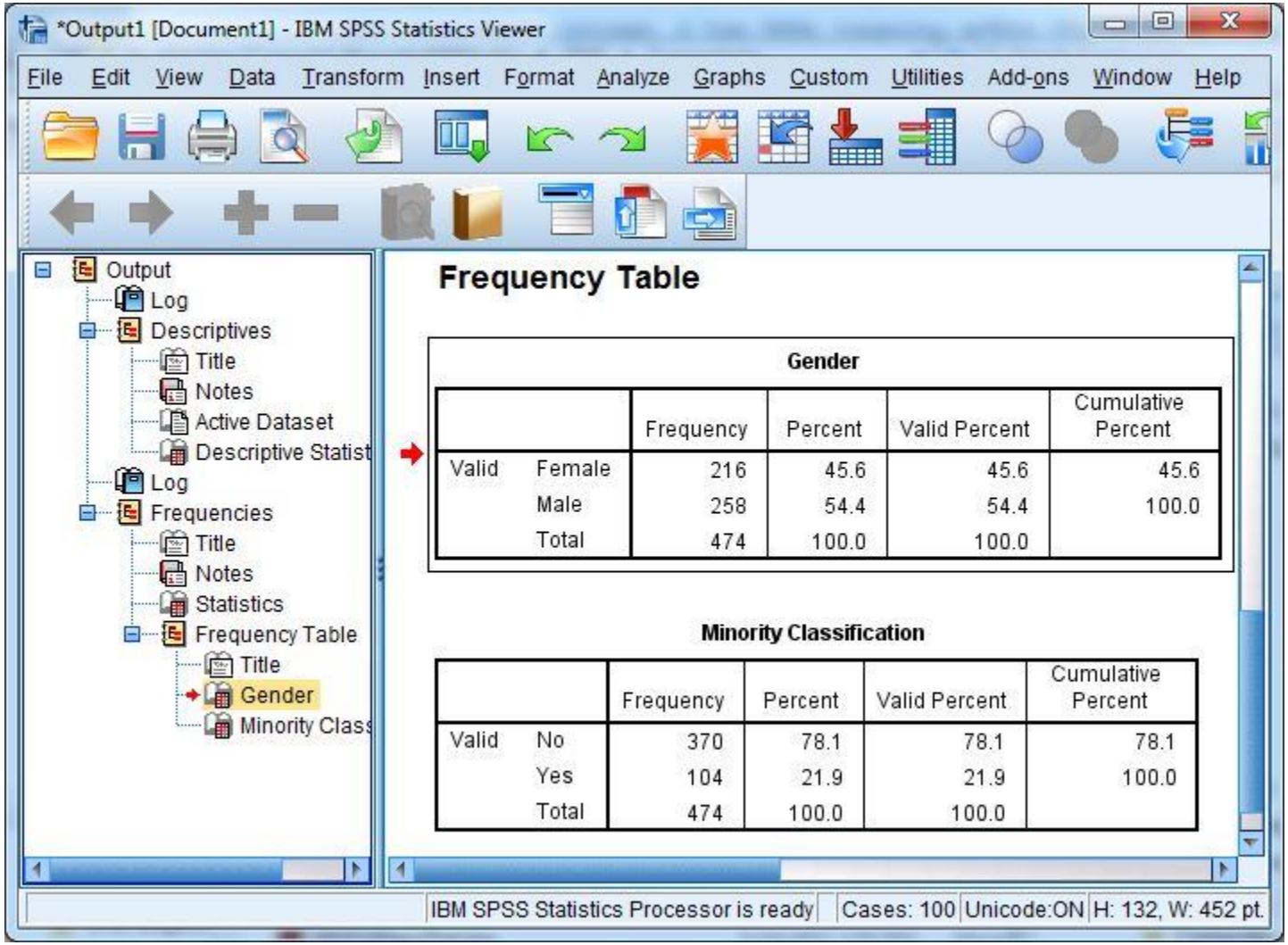
Como hacer análisis de datos

Programa: STATA



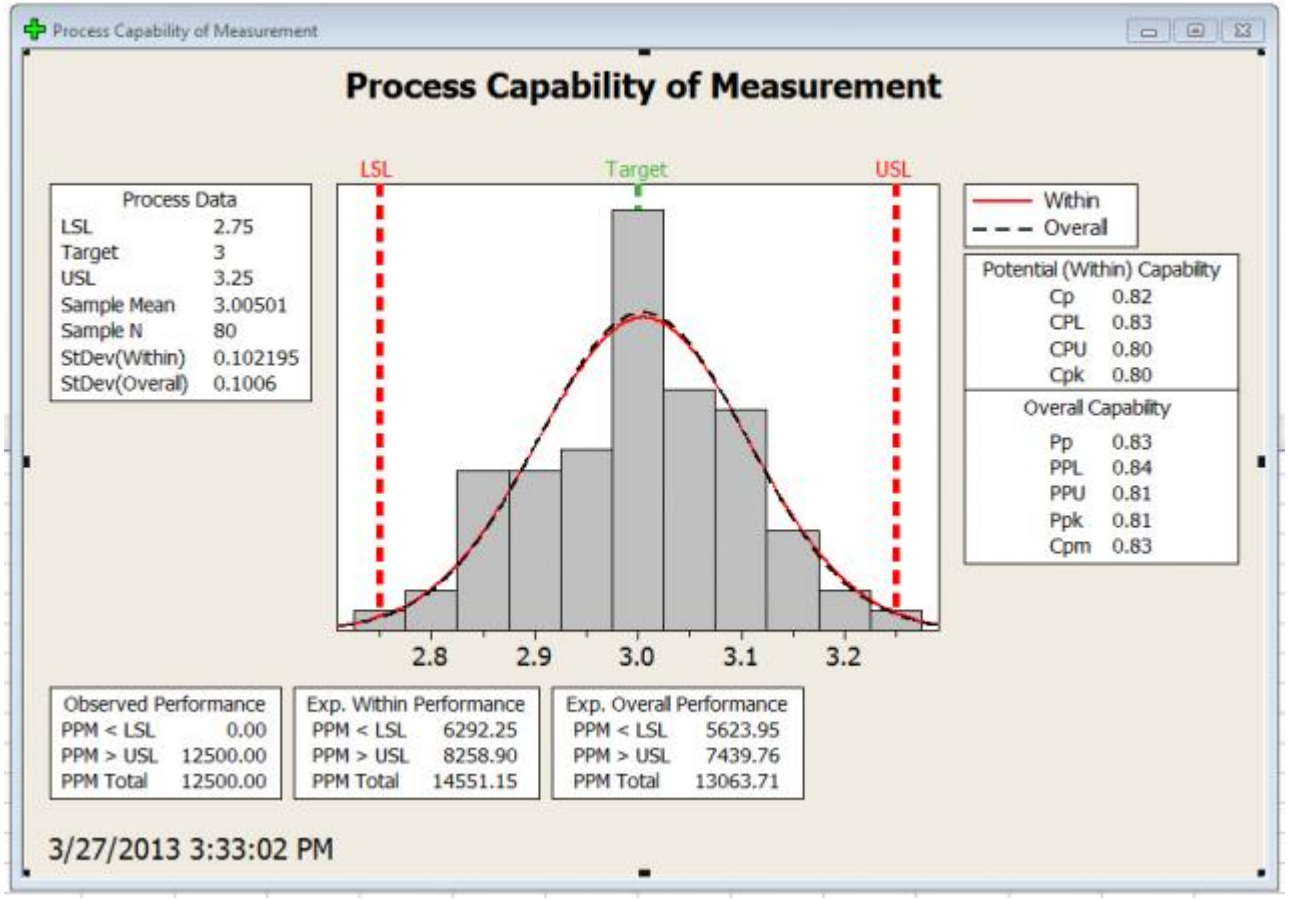
Como hacer análisis de datos

Programa: SPSS



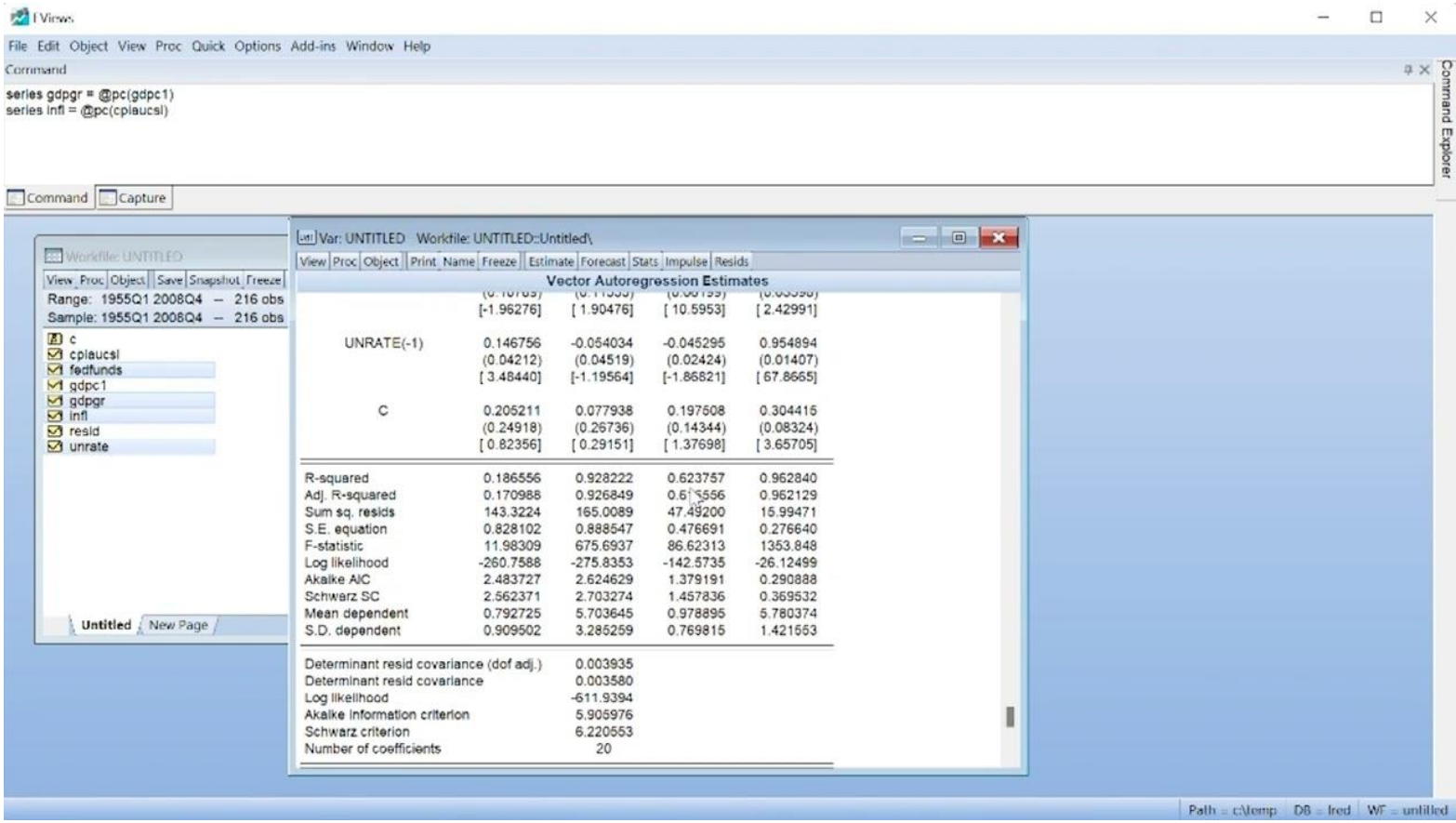
Como hacer análisis de datos

Programa: Minitab



Como hacer análisis de datos

Programa: Eviews



Como hacer análisis de datos

Stata
•Econometría avanzada y análisis de datos de panel.
•Series temporales y modelos estadísticos complejos.
•Programación y automatización con scripts personalizados.
•Gestión y manipulación de grandes bases de datos.

SPSS
•Análisis estadístico amigable para usuarios sin experiencia en programación.
•Estadísticas descriptivas e inferenciales fáciles de realizar.
•Análisis de encuestas y manejo de datos de encuestas.
•Interfaz intuitiva y fácil de usar.

Minitab
•Control de calidad y herramientas de Six Sigma.
•Análisis estadístico básico y descriptivo.
•Diseño de experimentos (DOE).
•Interfaz simple y amigable para análisis rápido.

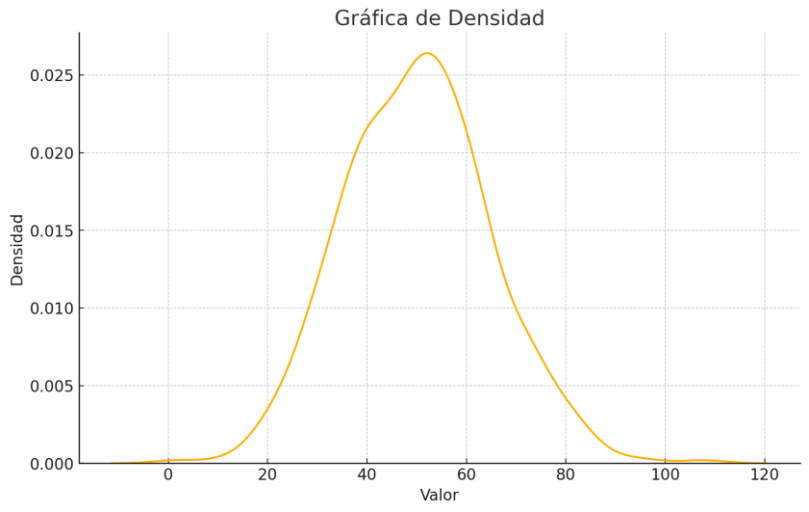
Eviews
•Series temporales y econometría avanzada.
•Análisis de datos de panel y modelos de cointegración.
•Modelado y pronóstico económico y financiero.
•Interfaz gráfica con opciones de simulación de escenarios.

Distribución de frecuencias

Una **distribución de frecuencias** es una representación tabular o gráfica que muestra cómo se distribuyen los valores de un conjunto de datos en diferentes categorías o intervalos. Su propósito es resumir un conjunto de datos y mostrar cuántas veces (la **frecuencia**) se repite cada valor o grupo de valores.

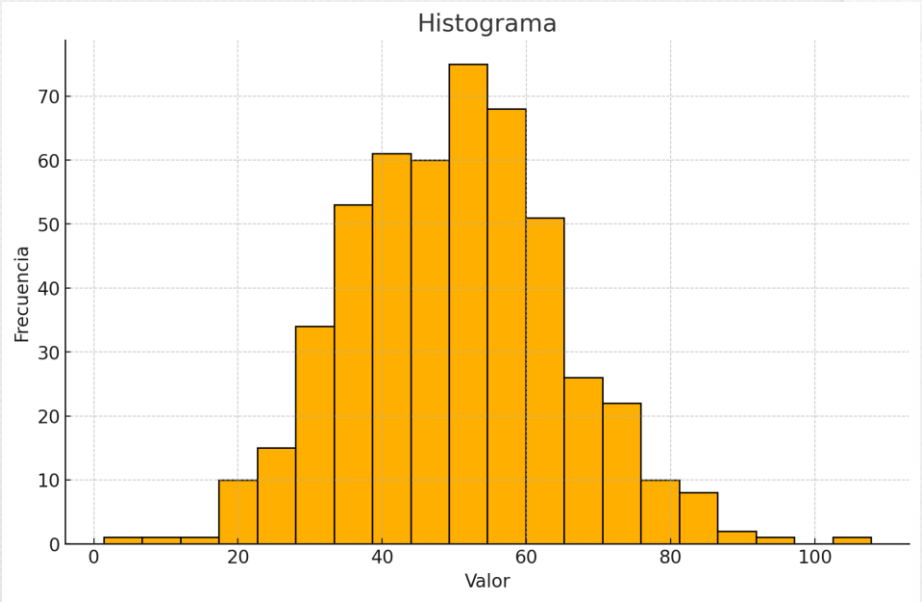
Calificación	Frecuencia absoluta	Frecuencia relativa (%)
70	4	20%
75	3	15%
80	5	25%
85	4	20%
90	3	15%
95	1	5%

Gráficas



Histograma
Gráfica de densidad
Gráfica de tallo y hoja
Gráfica Circular

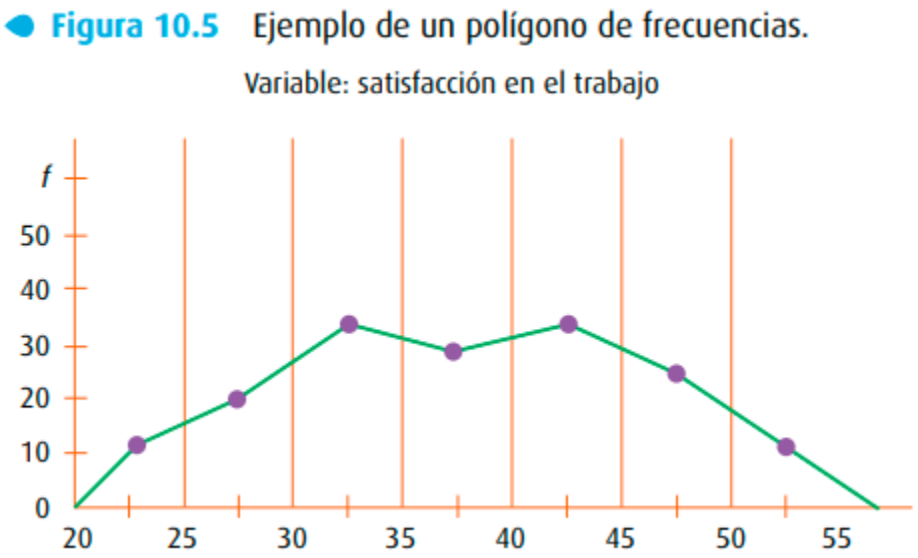
Tallo	Hoja
4	4 5 9
5	0 2 3 3 4 4 6 7 7 7 8
6	1 2 2 3 4 7 8 9
7	0 1 1 2 3 4 4 5 6 6 8 9
8	0 1 3 5



Gráficas

El polígono puede presentarse con frecuencias como en la figura 10.5 o con porcentajes como con este segundo ejemplo. Pero además de la distribución o polígono de frecuencias, deben calcularse las *medidas de tendencia central* y de *variabilidad o dispersión.*

Figura 10.5 Ejemplo de un polígono de frecuencias.



Medidas de tendencia central

¿Cuáles son las medidas de tendencia central?
Las **medidas de tendencia central** son puntos en una distribución obtenida, los valores medios o centrales de ésta, y nos ayudan a ubicarla dentro de la escala de medición de la variable analizada. Las principales medidas de tendencia central son tres: **moda, mediana y media**. El nivel de medición de la variable determina cuál es la medida de tendencia central apropiada para interpretar (Graham, 2013, Kwok, 2008a y Platt, 2003a).

Medidas de tendencia central

La **media** es el promedio aritmético de un conjunto de datos. Se calcula sumando todos los valores de la distribución y dividiendo el resultado por el número total de observaciones. Es una medida útil cuando los datos no tienen valores extremos que distorsionen el promedio.
La **mediana** es el valor central de un conjunto de datos ordenados de menor a mayor. Si el número de observaciones es impar, la mediana es el valor del medio. Si es par, la mediana se obtiene promediando los dos valores centrales. Es especialmente útil cuando hay valores atípicos, ya que no se ve afectada por ellos.
La **moda** es el valor que aparece con mayor frecuencia en un conjunto de datos. Puede haber una o más modas si varios valores tienen la misma frecuencia máxima. La moda es útil para describir conjuntos de datos categóricos y también se puede usar con datos cuantitativos.

Medidas de variabilidad

¿Cuáles son las medidas de la variabilidad?
Las medidas de la variabilidad indican la dispersión de los datos en la escala de medición de la variable considerada y responden a la pregunta: ¿dónde están diseminadas las puntuaciones o los valores obtenidos? Las medidas de tendencia central son valores en una distribución y las medidas de la variabilidad son intervalos que designan distancias o un número de unidades en la escala de medición (Kon y Rai, 2013 y O'Brien, 2007). Las medidas de la variabilidad más utilizadas son rango, desviación estándar y varianza.

Medidas de variabilidad

La fórmula de la varianza (σ^2) para una población es:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Donde:

- σ^2 es la varianza de la población.
- X_i representa cada valor individual en el conjunto de datos.
- μ es la media de la población.
- N es el número total de elementos en la población.

Para una muestra, la fórmula de la varianza muestral (s^2) es:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Donde:

- s^2 es la varianza de la muestra.
- X_i representa cada valor individual en la muestra.
- $\bar{X}$ es la media de la muestra.
- n es el número de elementos en la muestra.

La varianza sirve para medir la **dispersión** de los datos en un conjunto. Es una forma de cuantificar cuánto se separan o dispersan los valores individuales respecto a la media de un conjunto de datos.

Medidas de variabilidad

Para la desviación estándar de una población (σ):

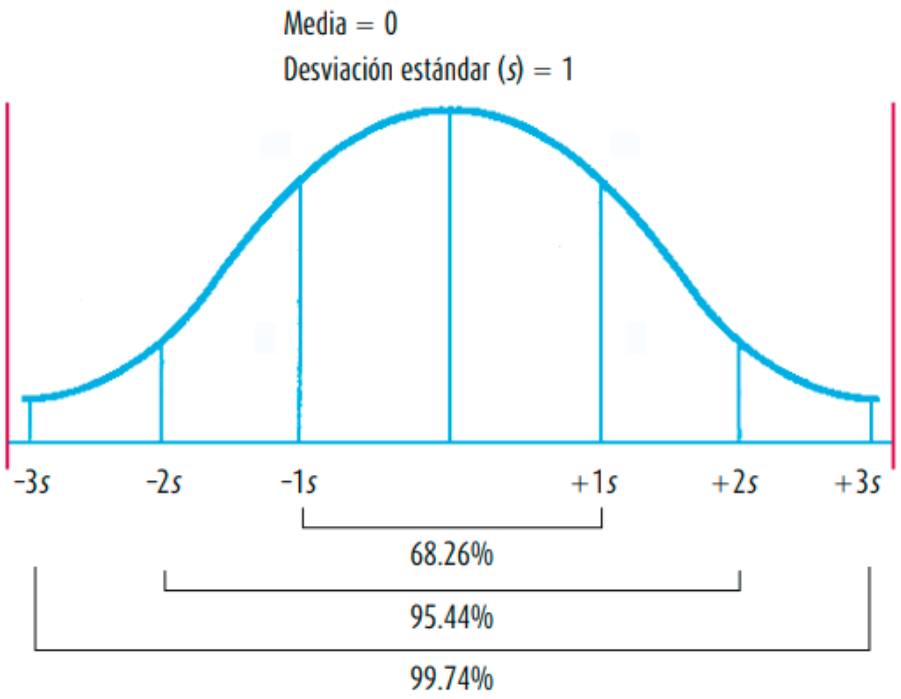
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Donde:

- σ es la desviación estándar de la población.
- s es la desviación estándar de la muestra.
- X_i representa cada valor individual en el conjunto de datos.
- μ es la media de la población.
- $\bar{X}$ es la media de la muestra.
- N es el número total de elementos en la población.
- n es el número de elementos en la muestra.

La desviación estándar mide cuánto se desvían, en promedio, los valores de un conjunto de datos respecto a su media.

Medidas de variabilidad



68.26% del área de la curva normal es cubierta entre $-1s$ y $+1s$, 95.44% del área de esta curva es cubierta entre $-2s$ y $+2s$ y 99.74% se cubre con $-3s$ y $+3s$.

¿Cómo se traducen las estadísticas descriptivas al inglés?

Algunos programas y paquetes estadísticos computacionales pueden realizar el cálculo de las estadísticas descriptivas, cuyos resultados aparecen junto al nombre respectivo de éstas, muchas veces en inglés.

A continuación se indican las diferentes estadísticas y su equivalente en inglés.

Estadística	Equivalente en inglés
• Moda	• *Mode*
• Mediana	• *Median*
• Media	• *Mean*
• Desviación estándar	• *Standard deviation*
• Varianza	• *Variance*
• Máximo	• *Maximum*
• Mínimo	• *Minimum*
• Rango	• *Range*
• Asimetría	• *Skewness*
• Curtosis	• *Kurtosis*

Razones y tasas

Una razón es la relación entre dos categorías. Por ejemplo:

Categorías	Frecuencia
Masculino	60
Femenino	30

La razón de hombres a mujeres es de $\frac{60}{30} = 2$. Es decir, por cada dos hombres hay una mujer.

Una **tasa** es la relación entre el número de casos, frecuencias o eventos de una categoría y el número total de observaciones, multiplicada por un múltiplo de 10, generalmente 100 o 1 000. La fórmula es:

$$\text{Tasa} = \frac{\text{Número de eventos}}{\text{Número total de eventos posibles}} \times 100 \text{ o } 1\,000$$

Universidad de la Ciudad de Aguascalientes

Análisis paramétrico

Mentes que transforman el mundo

Coeficientes de correlación: Pearson

Fórmula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = coeficiente de correlación

x_i = valores de la variable x en una muestra

$\bar{x}$ = media de los valores de la variable x

y_i = valores de la variable y en una muestra

$\bar{y}$ = media de los valores de la variable y

Coeficientes de correlación: Pearson

El coeficiente de correlación de Pearson es una prueba que mide la relación estadística entre dos variables continuas. Si la asociación entre los elementos no es lineal, entonces el coeficiente no se encuentra representado adecuadamente.

El coeficiente de correlación puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva. Es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra disminuye.

Coeficientes de correlación: Pearson

Universidad de la Ciudad de Aguascalientes

Tabla 10.12 Correlaciones entre moral y dirección

Correlaciones			
		Moral	**Dirección**
Moral	Correlación de Pearson	1	0.557**
	Sig. (bilateral)		0.000
	N	362	335
Dirección	Correlación de Pearson	0.557**	1
	Sig. (bilateral)	0.000	
	N	335	373

** La correlación es significativa al nivel 0.01 (bilateral, en ambos sentidos entre las variables).

Consideraciones: cuando el coeficiente r de Pearson se eleva al cuadrado (r2), se obtiene el coeficiente de determinación y el resultado indica la varianza de factores comunes. Esto es, el porcentaje de la variación de una variable debido a la variación de la otra variable y viceversa (o cuánto explica o determina una variable la variación de la otra).

Regresión lineal

La **regresión lineal** es un método estadístico y de aprendizaje automático que se utiliza para modelar la relación entre una variable dependiente (o respuesta) y una o más variables independientes (o predictoras). El objetivo es encontrar la línea recta (en el caso de una regresión lineal simple) o el hiperplano (en el caso de múltiples variables independientes) que mejor ajuste los datos observados.

Objetivo: El objetivo de la regresión lineal es minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (esto se llama **mínimos cuadrados**).

Regresión lineal

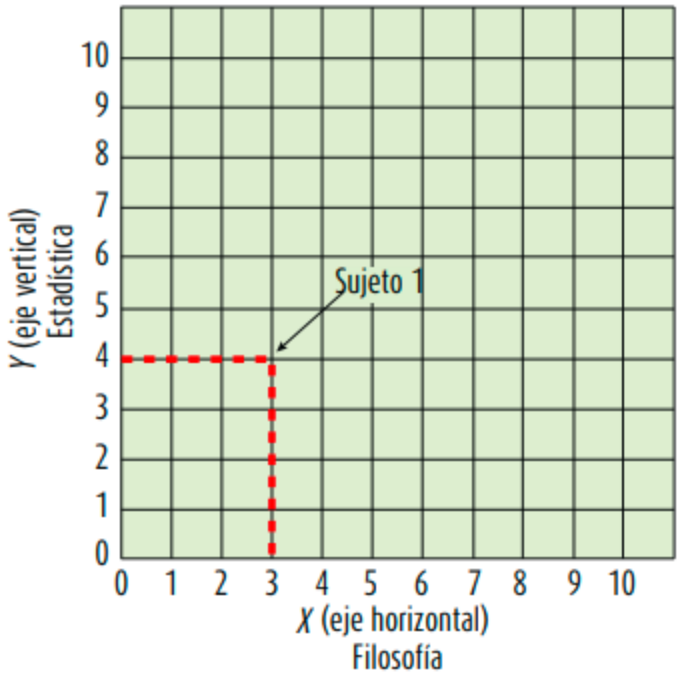
2. **Ecuación:** La fórmula de la regresión lineal simple es:

$$Y = a + bX + \epsilon$$

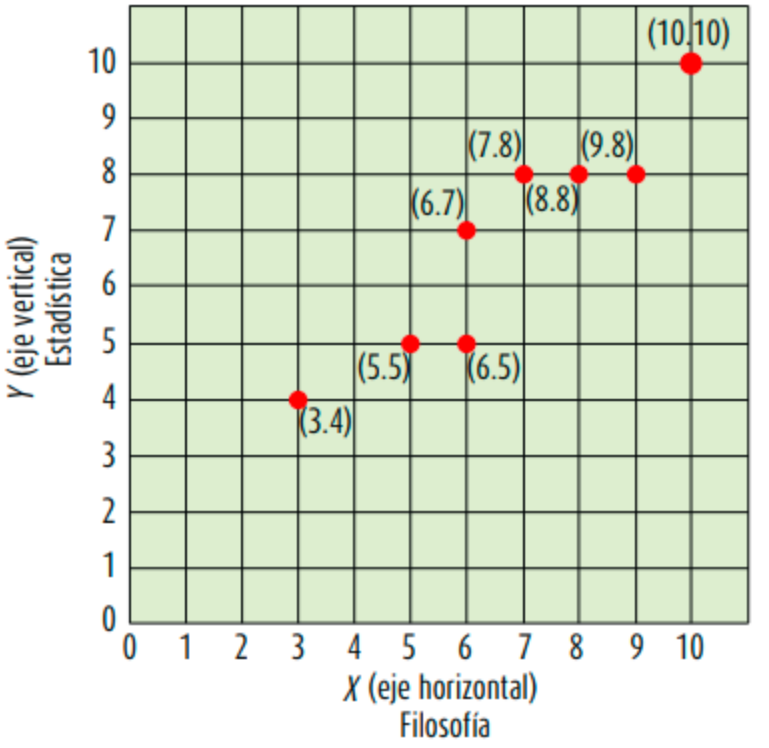
- Y es la variable dependiente.
- X es la variable independiente.
- a es la intersección (o término independiente).
- b es el coeficiente de regresión (la pendiente de la línea).
- ϵ es el término de error.

Regresión lineal

El *diagrama de dispersión* se construye graficando cada par de puntuaciones en un espacio o plano bidimensional. Sujeto "1" tuvo 3 en *X* (filosofía) y 4 en *Y* (estadística):



Así se grafican todos los pares:



(*continúa*)

Análisis de varianza

¿Qué es el análisis de varianza unidireccional o de un factor? (ANOVA one-way)

El análisis de varianza unidireccional o de un factor (ANOVA de un factor) es una técnica estadística que se utiliza para comparar las medias de tres o más grupos independientes para determinar si existe una diferencia significativa entre ellas. Esta prueba se emplea cuando hay una sola variable independiente (factor) con diferentes niveles o categorías, y se desea ver si esa variable tiene un efecto sobre una variable dependiente continua.

Análisis de covarianza

El **análisis de covarianza (ANCOVA)** es un método estadístico que combina características del análisis de varianza (ANOVA) y la regresión lineal. Su objetivo es evaluar si existen diferencias significativas entre los grupos de una variable dependiente mientras se controla el efecto de una o más variables continuas llamadas **covariables** o **covariantes**.

Chi cuadrada

La **chi cuadrada (χ^2)** es una prueba estadística que se utiliza para determinar si existe una diferencia significativa entre las frecuencias observadas y las frecuencias esperadas en uno o más grupos de categorías. Esta prueba se emplea principalmente en análisis de datos categóricos y es útil para evaluar si la distribución de datos observados difiere de una distribución hipotética.

Chi cuadrada

La **chi cuadrada (χ^2)** es una prueba estadística que se utiliza para determinar si existe una diferencia significativa entre las frecuencias observadas y las frecuencias esperadas en uno o más grupos de categorías. Esta prueba se emplea principalmente en análisis de datos categóricos y es útil para evaluar si la distribución de datos observados difiere de una distribución hipotética.

La fórmula básica de la chi cuadrada es:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde:

- O_i representa las frecuencias observadas.
- E_i representa las frecuencias esperadas bajo la hipótesis nula.

Chi cuadrada

Interpretación:

•Un valor de chi cuadrada bajo indica que las diferencias entre las frecuencias observadas y esperadas son pequeñas, lo que sugiere que los datos se ajustan bien a la hipótesis nula.

•Un valor de chi cuadrada alto indica que hay una diferencia significativa entre las frecuencias observadas y las esperadas, lo que podría llevar al rechazo de la hipótesis nula.

Ejemplo sencillo:

Imagina que lanzas un dado 60 veces y esperas que cada número (1 a 6) salga aproximadamente 10 veces si el dado es justo. Si observas que algunos números salen con más frecuencia que otros, puedes usar la prueba de chi cuadrada para determinar si esta diferencia es significativa o podría haber ocurrido por azar.

En resumen, la chi cuadrada es una herramienta importante en la estadística inferencial para probar hipótesis sobre la distribución de datos en variables categóricas.

Chi cuadrada

1. Prueba de independencia

Funcionamiento:

•Se construye una tabla de contingencia donde se registran las frecuencias observadas de cada combinación de las categorías de las dos variables.

•Se calcula el valor de la chi cuadrada comparando las frecuencias observadas con las frecuencias esperadas, que se calculan bajo la hipótesis nula de que las variables son independientes.

2. **Prueba de bondad de ajuste**

Funcionamiento:

•Compara las frecuencias observadas de una muestra con las frecuencias esperadas calculadas bajo una distribución teórica (por ejemplo, uniforme o binomial).

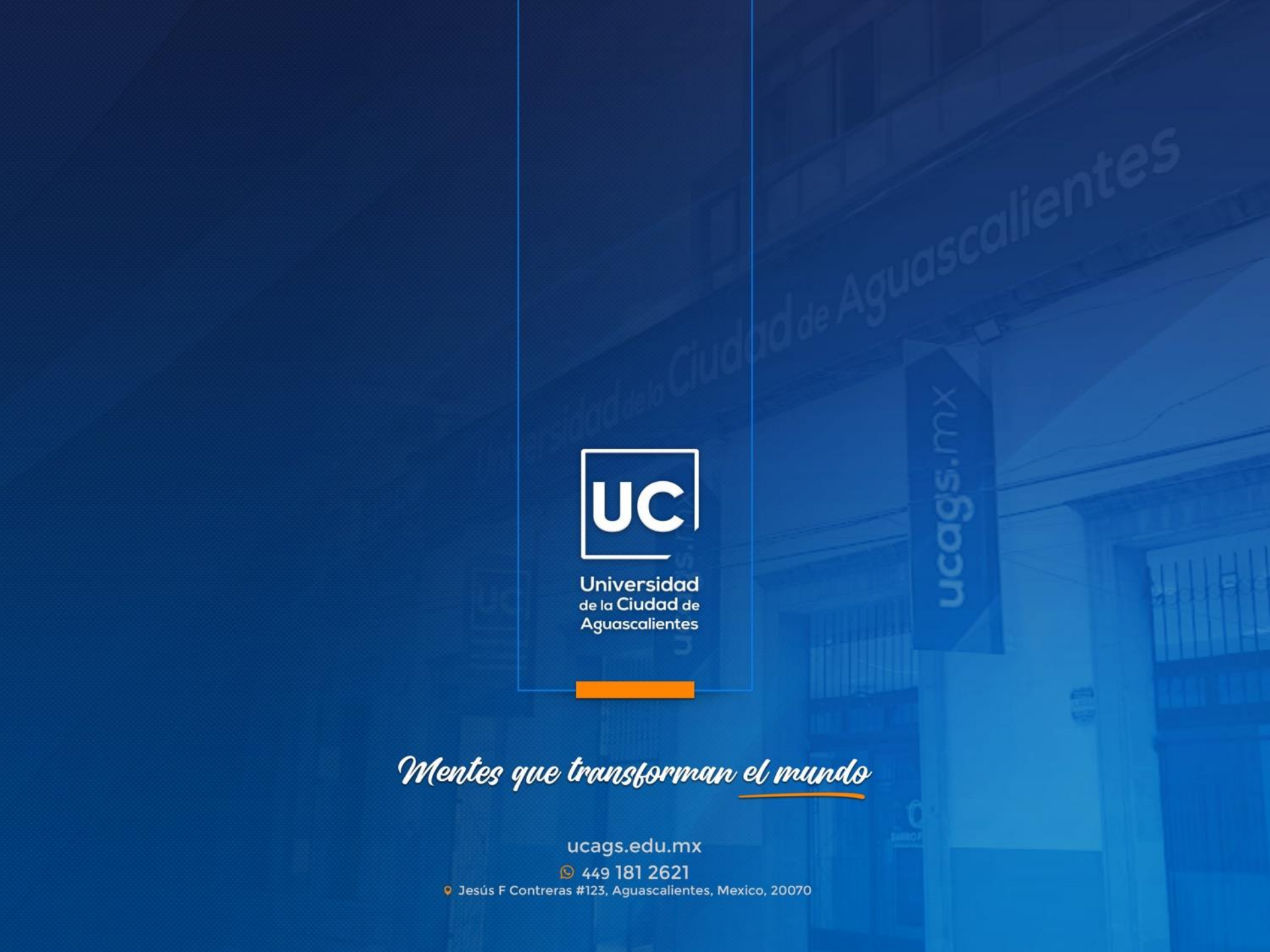
Chi cuadrada

Interpretación de los resultados

En ambas pruebas:

- Si el valor calculado de χ^2 es mayor que el valor crítico de chi cuadrada correspondiente al nivel de significancia (α) y los grados de libertad, se rechaza la hipótesis nula.
- La **hipótesis nula** en la prueba de independencia es que las dos variables son independientes.
- La **hipótesis nula** en la prueba de bondad de ajuste es que los datos siguen la distribución teórica esperada.

Estas pruebas ayudan a evaluar si las diferencias observadas pueden atribuirse al azar o si son lo suficientemente significativas como para sugerir una relación o un ajuste inadecuado.


Universidad
de la Ciudad de
Aguascalientes
Mentes que transforman el mundo
ucags.edu.mx
449 181 2621
Jesús F Contreras #123, Aguascalientes, Mexico, 20070