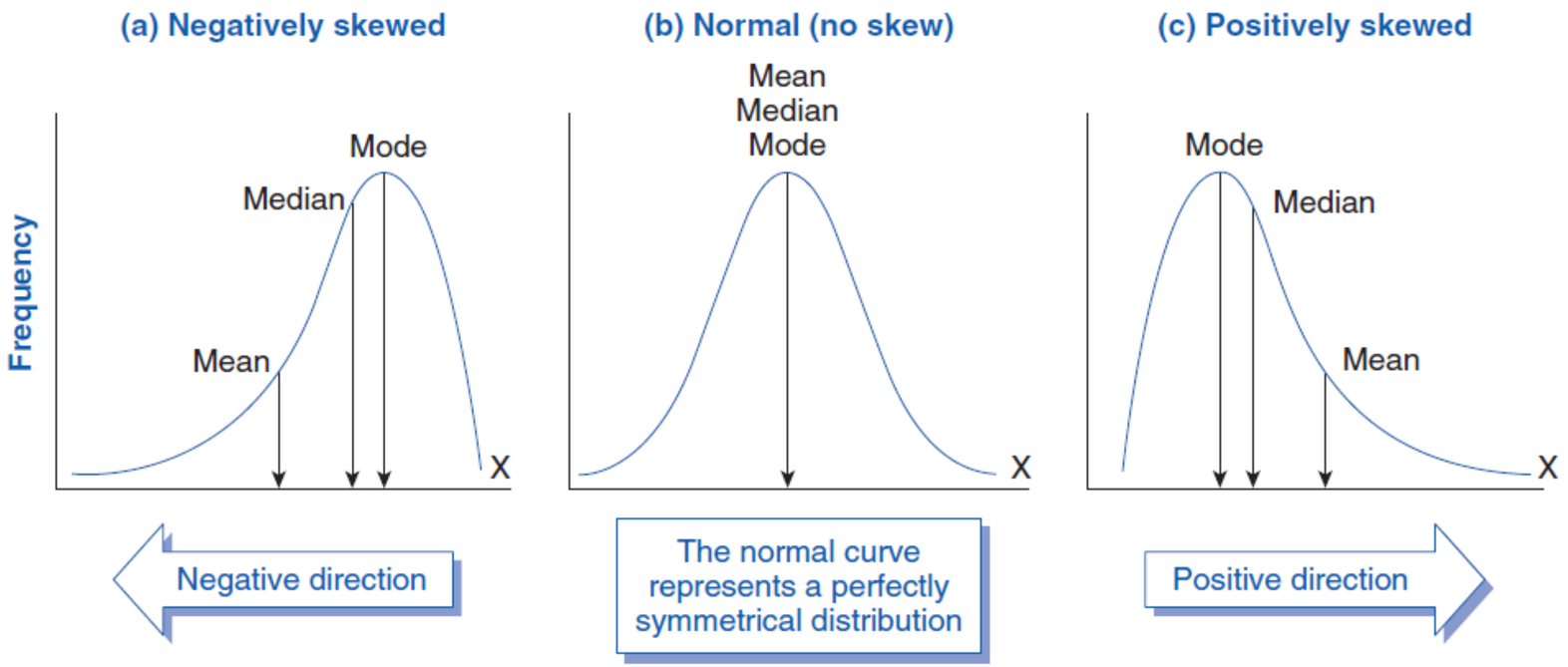
INTRODUCCIÓN A LA **CIENCIA DE DATOS**

QUIZ

Histograma de Notas



Histograma de Notas

Pregunta 01

Los estudiantes de dos clases tomaron un examen. Ambas clases tuvieron la misma puntuación media de 76. Sin embargo, la Clase #1 mostró una desviación estándar de 10, mientras que la Clase #2 mostró una desviación estándar de 16. Esto significa que:

Seleccione una:

a.
Hubo un rango más amplio de puntajes en la Clase #1 que en la Clase #2.

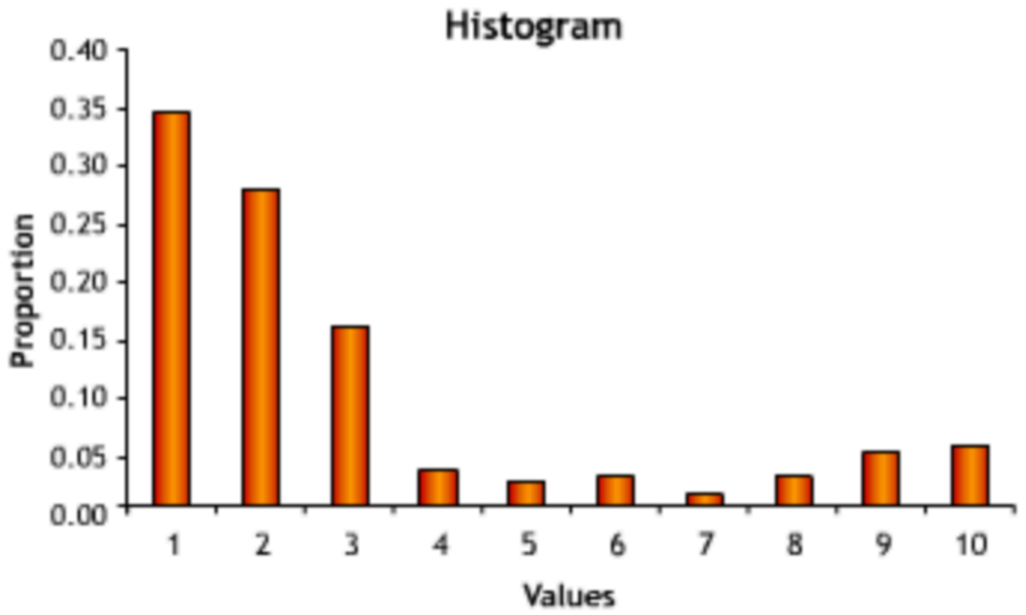
b.
El puntaje promedio de la Clase #2 fue 6 puntos más alto que el puntaje promedio de la Clas

c.
Hubo un rango más amplio de puntajes en la Clase #2 que en la Clase #1.

d.
La clase #2 obtuvo mejores resultados en las pruebas en general que la clase #1.

Pregunta 02

¿Cuál opción clasifica correctamente de menor a mayor las medidas de tendencia central?



Seleccione una:

- a. No hay suficiente información para determinar la respuesta.
- b. Moda < Mediana < Media
- c. Mediana < Moda < Media
- d. Moda < Media < Mediana

Pregunta 03

¿Cuál de los estadísticos se mide en las mismas unidades que la variable analizada?

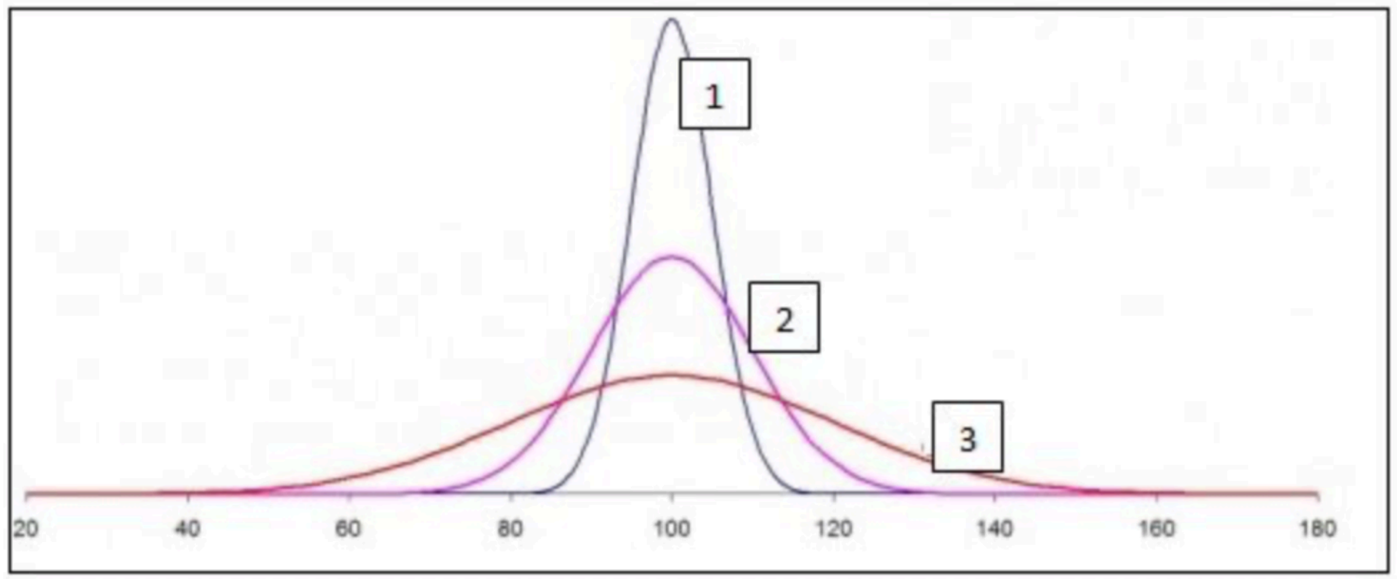
Seleccione una:

- a. Desviación Estándar
- b. Ninguna de las anteriores
- c. Varianza
- d. Coeficiente de Variación

Pregunta 04

Para la distribución normal inferior, ¿cuál de las siguientes opciones es verdadera?

σ1, σ2 y σ3 representan las desviaciones estándar para las curvas 1, 2 y 3 respectivamente.



Seleccione una:

- a. σ1= σ2= σ3
- b. σ1> σ2> σ3
- c. σ1< σ2< σ3
- d. Ninguna de las anteriores.

POBLACIÓN VS. MUESTRA

Población vs. Muestra

Población

Una población es todo el grupo sobre el que desea sacar conclusiones.

Muestra

Una muestra es el grupo específico del que se recopilará datos. El tamaño de la muestra es siempre menor que el tamaño total de la población.

Fuente: *Population vs sample: what's the difference?*, Pritha Bhandari

Población vs. Muestra

A safety inspector conducts air quality tests on a randomly selected group of 7 classrooms at an elementary school.

Identify the population and sample in this setting.

Choose 1 answer:

(A) The population is all classrooms in the district; the sample is the 7 classrooms selected.

(B) The population is all classrooms in the elementary school; the sample is the 7 classrooms selected.

(C) The population is all elementary students in the school; the sample is the students in the 7 classrooms selected.

Fuente: *Identifying the population and sample, Khan Academy.*

Población vs. Muestra

Lucio wants to know whether the food he serves in his restaurant is within a safe range of temperatures. He randomly selects 70 entrees and measures their temperatures just before he serves them to his customers.

Identify the population and sample in this setting.

Choose 1 answer:

(A) The population is all of the hot entrees Lucio serves; the sample is the entrees that are a safe temperature.

(B) The population is the 70 selected entrees; the sample is the entrees that are a safe temperature.

(C) The population is all of the entrees Lucio serves; the sample is the 70 selected entrees.

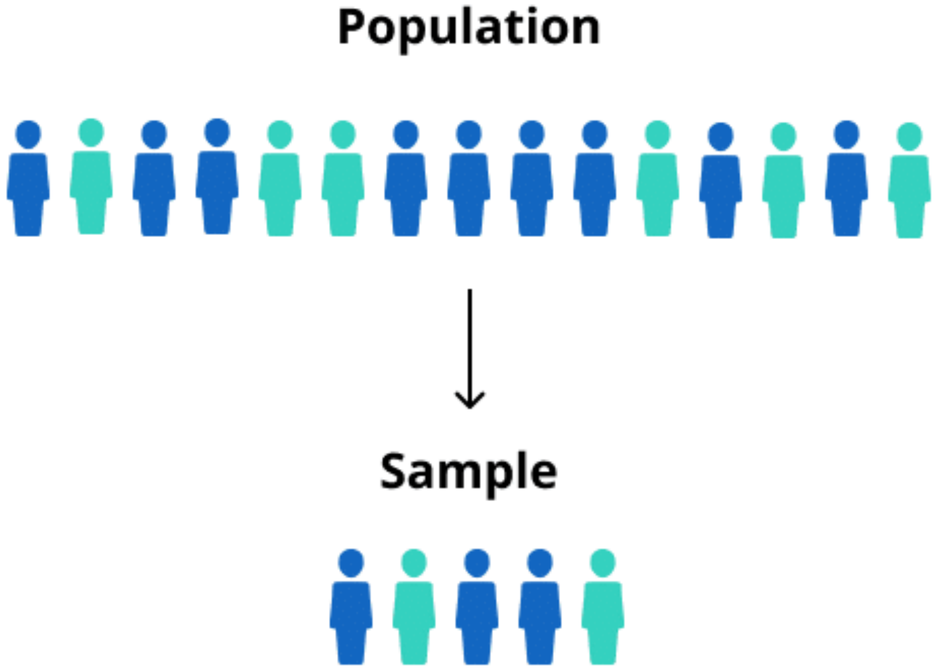
Fuente: *Identifying the population and sample, Khan Academy.*

Población vs. Muestra

Razones para utilizar muestras

- **Necesidad**: a veces simplemente no es posible estudiar a toda la población debido a su tamaño o inaccesibilidad.
- **Practicidad**: es más fácil y eficiente recopilar datos de una muestra.
- **Rentabilidad**: hay menos costos de participantes, laboratorios, equipos e investigadores involucrados.
- **Capacidad de administración**: almacenar y ejecutar análisis estadísticos en conjuntos de datos más pequeños es más fácil.

Por lo general, recopilar datos de una población completa es sencillo cuando es pequeña, accesible y cooperativa.



Fuente: *Population vs sample: what's the difference?*, Pritha Bhandari

Población vs. Muestra

Parámetro de población vs. estadística de muestra

- Cuando se recopilan datos de una población o una muestra, hay varias medidas y números que se pueden calcular a partir de los datos. Un **parámetro** es una medida que **describe a toda la población**. Una **estadística** es una medida que **describe la muestra**.

Fuente: *Population vs sample: what's the difference?*, Pritha Bhandari

INTERVALOS DE CONFIANZA

Sample Distributions

Cada estadístico tiene un valor desconocido correspondiente en la población llamado parámetro que se puede estimar.

- Se selecciona una muestra aleatoria con una media de x barra.
- Si se recolectara otra muestra aleatoria con el mismo número de observaciones, la media también podría calcularse y es probable que las medias de las dos muestras sean cercanas pero no idénticas.
- Si tomamos muchas muestras aleatorias de igual tamaño y calculamos el valor medio de cada muestra, comenzaríamos a formar una distribución de frecuencias.
- Si tuviéramos que tomar un número infinito de muestras de igual tamaño y graficar en un gráfico el valor de la media calculada de cada muestra, se produciría una **distribución de frecuencia normal que refleja la distribución de las medias muestrales para la media poblacional en consideración**.

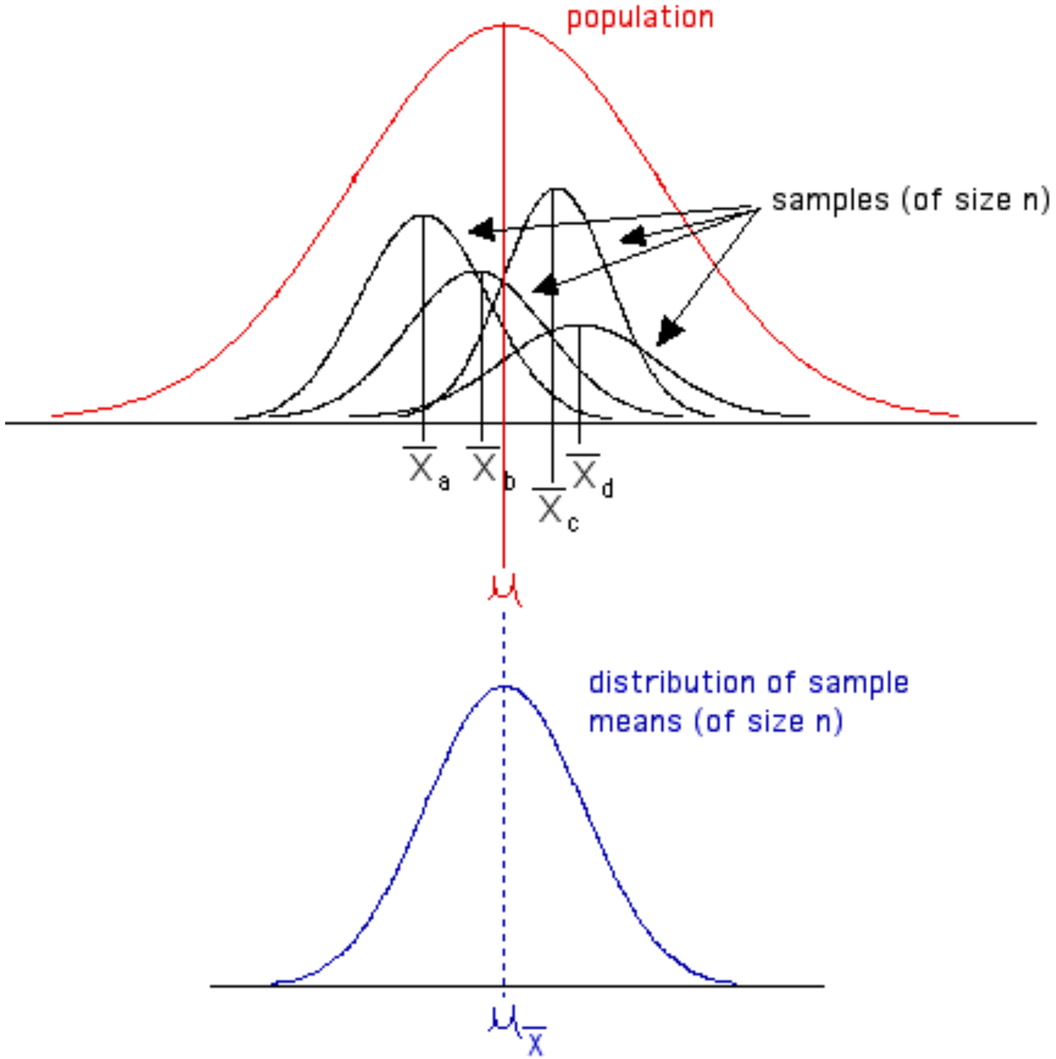
Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Sample Distributions

- La media de una distribución muestral se denomina **valor esperado de la media**: es la media esperada de la población.
- La desviación estándar de la distribución muestral se llama **error estándar**: mide cuánto error esperar de muestras aleatorias de igual tamaño extraídas de la misma población. El error estándar nos informa de la diferencia promedio entre la media de una muestra y el valor esperado.
- Independientemente de cómo se distribuyan los valores de una variable para la población, la distribución muestral de un estadístico calculado sobre muestras de esa variable tendrá una forma normal cuando el tamaño elegido para las muestras tenga al menos 30 observaciones. Esto se conoce como la ley de los grandes números, o más formalmente como el **teorema del límite central.**

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Sample Distributions



Error Estándar

El **error estándar** proporciona un nivel medible de confianza de que tan bien la media muestral estima la media de la población.

El error estándar se puede calcular a partir de una muestra utilizando la siguiente fórmula:

$$standard\ error\ of\ the\ sampling\ distribution = \frac{s}{\sqrt{n}}$$

donde s es la desviación estándar de una muestra y n es el número de observaciones en la muestra. Debido a que el tamaño n está en el denominador y la desviación estándar s está en el numerador, las muestras pequeñas con grandes variaciones aumentan el error estándar, lo que reduce la confianza de que el estadístico de la muestra es una aproximación cercana al parámetro de población que estamos tratando de estimar.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Nivel de Confianza

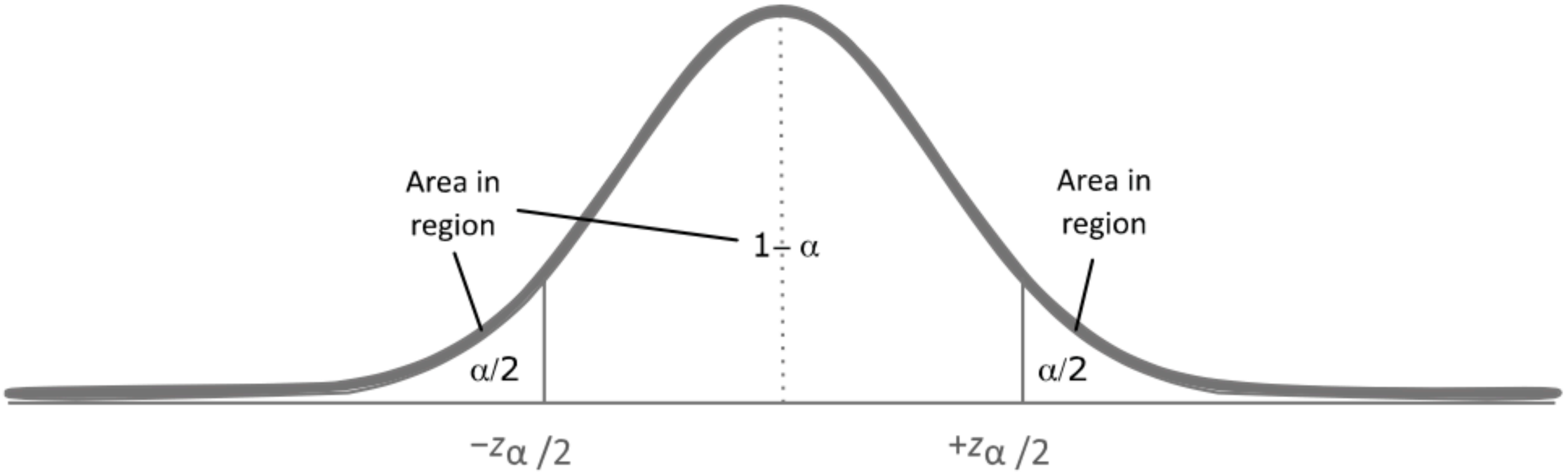
Utilizando un intervalo de confianza del 95%, una forma de interpretar este nivel de confianza es que, **en promedio, el valor de población correcto se encontrará dentro del intervalo de confianza derivado 95 veces de cada 100 muestras recolectadas**. En estas 100 muestras, habrá 5 ocasiones en promedio en las que este valor no se encuentre dentro del rango. El nivel de confianza generalmente se expresa en términos de α de la siguiente ecuación:

$$confidence\ interval = 100 \times (1 - \alpha)$$

El valor utilizado para este nivel de confianza afectará el tamaño del intervalo; es decir, cuanto mayor sea el nivel de confianza deseado, mayor será el intervalo de confianza.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Intervalos de Confianza



Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Intervalos de Confianza

La fórmula para un **intervalo de confianza** para un valor medio es:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

donde $\bar{x}$ es el valor medio, s es la desviación estándar y n es el número de observaciones en la muestra. El valor de $z\alpha / 2$ se basa en el área a la derecha de una distribución z estándar como se ilustra en la Figura anterior. Este número puede derivarse de una tabla estadística estándar o un programa de computadora.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Intervalos de Confianza

- La media muestral es nuestra mejor estimación de la media poblacional. Esto sugiere que la **media muestral debería ser siempre el centro del rango.**
- El ancho del rango depende de:
  - La **desviación estándar** de la muestra.
  - El **tamaño de la muestra.**
  - El **nivel de confianza** deseado.

Ejemplo

Un productor de manzanas está analizando el tamaño de las frutas. Hay cientos de manzanas en los árboles, por lo que elige al azar solo 46 manzanas y obtiene:

- Una media de 86
- Una desviación estándar de 6,2
- Confianza del 95%

$$86 \pm 1{,}960 \times \frac{6{,}2}{\sqrt{46}} = \mathbf{86 \pm 1{,}79}$$

Por lo tanto, es probable que la media real (de todos los cientos de manzanas) esté entre 84,21 y 87,79.

IC para Proporciones

¿Cómo sabe que está lidiando con un problema de proporción?

Primero, la distribución subyacente tiene una variable aleatoria binaria y, por lo tanto, es una distribución binomial. Si X es una variable aleatoria binomial, entonces X ~ B (n, p) donde n es el número de intentos y p es la probabilidad de éxito. Para formar una proporción muestral, tome X, la variable aleatoria para el número de éxitos y divídala por n, el número de ensayos (o el tamaño de la muestra).

$$P' = \frac{X}{n}$$

p′ = la proporción estimada de éxitos o la proporción muestral de éxitos

x = el número de éxitos en la muestra

n = el tamaño de la muestra

IC para Proporciones

El intervalo de confianza para una proporción de población es:

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

Nivel de confianza

Std. Deviation

IC para Proporciones

Ejemplo

Suponga que se contrata a una empresa de investigación de mercado para estimar el porcentaje de adultos que viven en una gran ciudad y que tienen teléfonos móviles. Se encuesta a quinientos residentes adultos seleccionados al azar en esta ciudad para determinar si tienen teléfonos celulares. De las 500 personas de la muestra, 421 respondieron que sí: tienen teléfonos móviles. Con un nivel de confianza del 95%, calcule una estimación del intervalo de confianza para la proporción real de residentes adultos de esta ciudad que tienen teléfonos celulares.

- n = 500
- x = el número de éxitos en la muestra = 421

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

- Estimamos con un 95% de confianza que entre el 81% y el 87,4% de todos los residentes adultos de esta ciudad tienen teléfonos celulares.

PRUEBA DE HIPÓTESIS

Ejemplo

Por ejemplo, una empresa que fabrica productos para el cuidado del cabello desea decir que la cantidad promedio de shampoo dentro de las botellas es de 200 ml. Para probar esta hipótesis, la empresa recolecta una muestra aleatoria de 100 botellas de shampoo y mide con precisión el contenido de la botella. Si se infiere de la muestra que la cantidad promedio de shampoo en cada botella no es de 200 ml, entonces se puede tomar la decisión de detener la producción y rectificar el problema de fabricación.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Formulación de Hipótesis

El primer paso es formular la hipótesis que se probará. Esta hipótesis se conoce como **hipótesis nula (H0)**. La hipótesis nula se establece en términos de lo que se esperaría si no hubiera nada inusual en los valores medidos de las observaciones en los datos de las muestras que recopilamos: “nulo” implica la ausencia de efecto.

- **La hipótesis nula** sería: el volumen medio de shampoo en una botella es de 200 ml.
- Su **hipótesis alternativa** correspondiente (Ha) es que difieren o, expresado de forma medible, que el promedio no es igual a 200 mL.

$$H_0 : \mu = 200$$
$$H_a : \mu \neq 200$$

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Formulación de Hipótesis

$$H_0 : \mu = 200$$
$$H_a : \mu \neq 200$$

- Por tanto, el **resultado** de una prueba de hipótesis es **no rechazar o rechazar la hipótesis nula**. La hipótesis nula sería: el volumen medio de shampoo en una botella es de 200 ml.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Ejercicio

1. Tenemos un medicamento que se está fabricando y se supone que cada pastilla tiene 14 miligramos del ingrediente activo. ¿Cuáles son nuestras hipótesis nulas y alternativas?

2. El director de la escuela quiere probar si es cierto lo que dicen los maestros: que los estudiantes de tercer año de secundaria usan la computadora un promedio de 3.2 horas al día. ¿Cuáles son nuestras hipótesis nulas y alternativas?

Fuente: Hypothesis Testing ,ck12

Estadístico

Debido a que la hipótesis involucra la media, usamos la siguiente fórmula para calcular el estadístico de prueba:

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

donde $\bar{x}$ es el valor medio calculado de la muestra, $\mu 0$ es la media poblacional que es el tema de la prueba de hipótesis, s es la desviación estándar de la muestra y n es el número de observaciones.

Fuente: *Making Dense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. Myatt & Johnson.

Cuándo rechazar H0

Rechazar la hipótesis nula significa encontrar una diferencia suficientemente grande entre la media de la muestra y la media hipotética (nula).

- Si la diferencia entre la media hipotética y la media de la muestra es muy grande , rechazamos la hipótesis nula.
- Si la diferencia es muy pequeña, no lo hacemos.
- En cada prueba de hipótesis, tenemos que decidir de antemano cuál debe ser la magnitud de esa diferencia para permitirnos rechazar la hipótesis nula.

Fuente: Hypothesis Testing ,ck12

Cuándo rechazar H0

- Primero se debe eligir un **nivel de significancia o nivel alfa** (α) para la prueba de hipótesis.
- Este nivel alfa nos dice cuán improbable debe ser una media muestral para que se considere "significativamente diferente" de la media hipotética.
- Los niveles de significancia usados con más frecuencia son 0.05 y 0.01.
- Un nivel alfa de 0.05 significa que consideraremos que nuestra media muestral es significativamente diferente de la media hipotética si las posibilidades de observar esa media muestral son inferiores al 5%.

Fuente: Hypothesis Testing .ck12

Steps

STEP 1: State the hypotheses.
(Population)

α

STEP 2: Set the level of Significance
(Criterion)

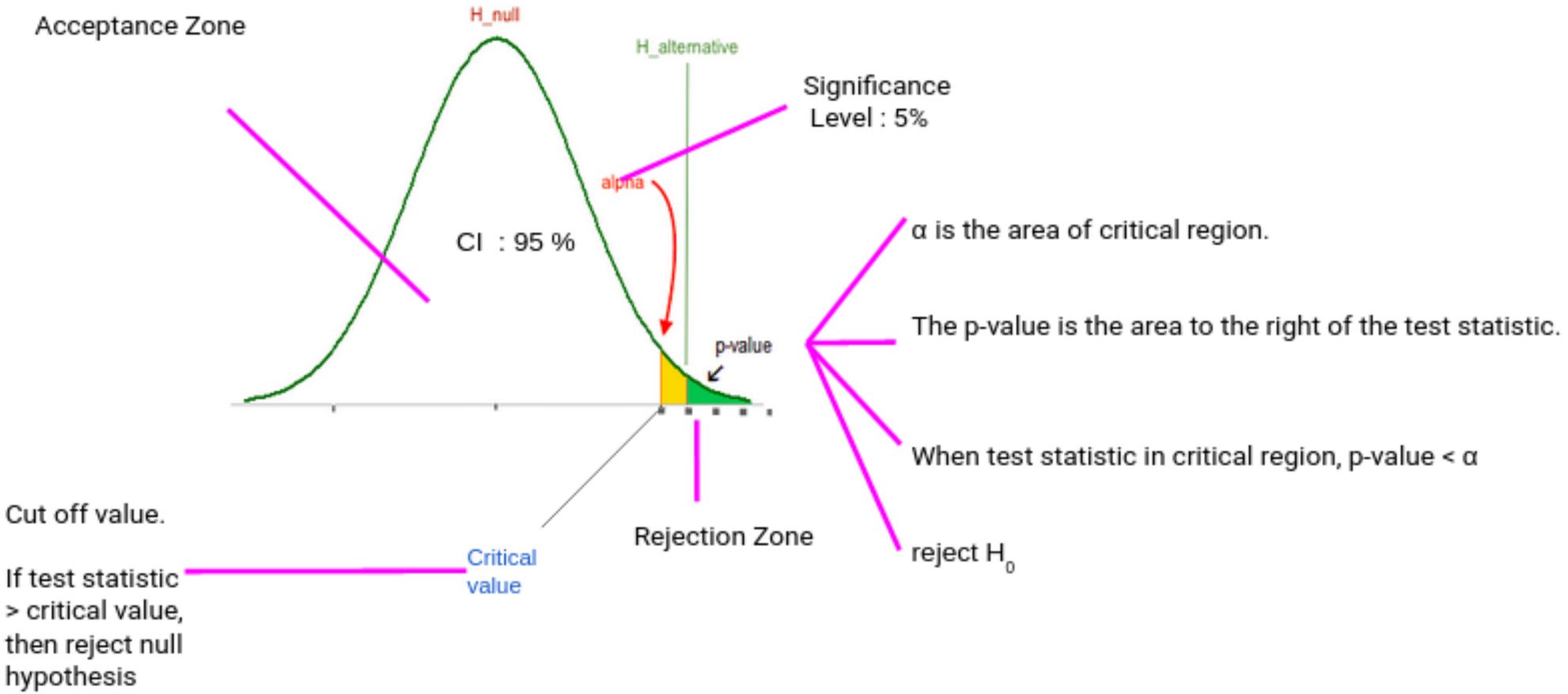
STEP 3: Compute test Statistics
(Sample)

p

STEP 4: Make a decision based on p value

Fuente: Statistics for Analytics and Data Science, SUBHASH MEENA

P-Value



Fuente: Statistics for Analytics and Data Science, SUBHASH MEENA

P-Value

Ventajas

- El valor p tiene la ventaja de que solo necesitamos un valor para tomar una decisión sobre la hipótesis.

  No necesitamos calcular dos valores diferentes, como el valor crítico y el Score de la prueba.

- Otro beneficio de usar el valor p es que podemos probar en cualquier nivel de significancia deseado

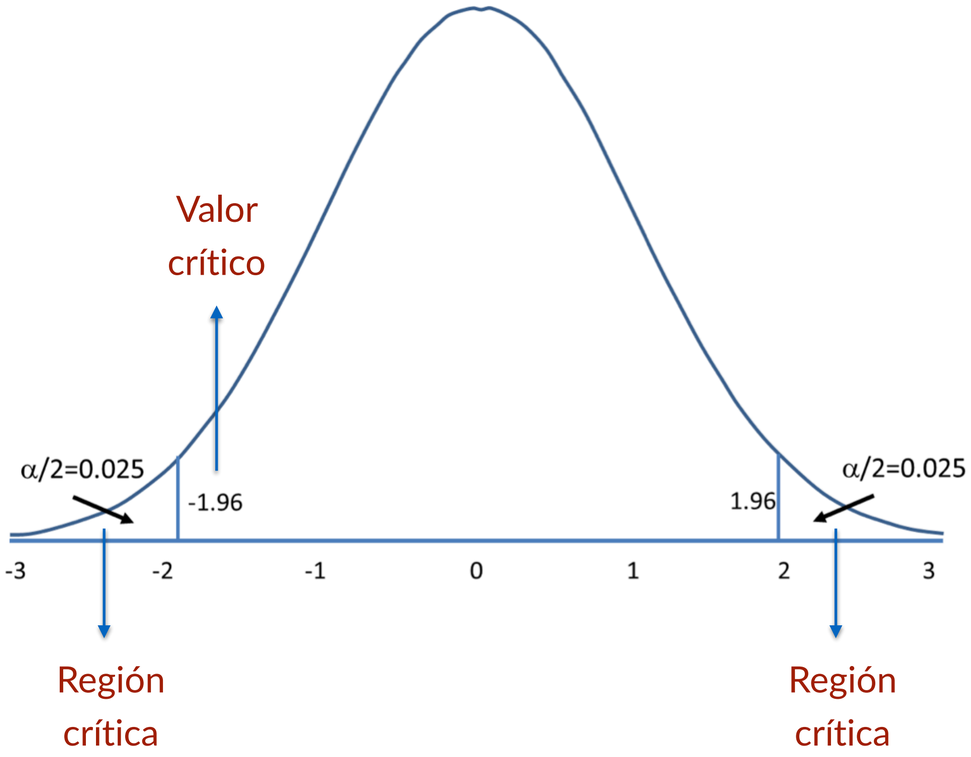
  comparando esto directamente con el nivel de significancia.

Fuente: Statistics for Analytics and Data Science, SUBHASH MEENA

Pruebas de hipótesis de dos colas

- En una prueba de dos colas, se rechazará la hipótesis nula si la media muestral cae en cualquiera de las colas de la distribución.

- Si la media de la muestra tomada de la población cae dentro de estas regiones críticas, o "regiones de rechazo", concluiríamos que hubo demasiada diferencia y rechazaríamos la hipótesis nula. Sin embargo, si la media de la muestra cae en el medio de la distribución (entre las regiones críticas) no rechazaríamos la hipótesis nula.



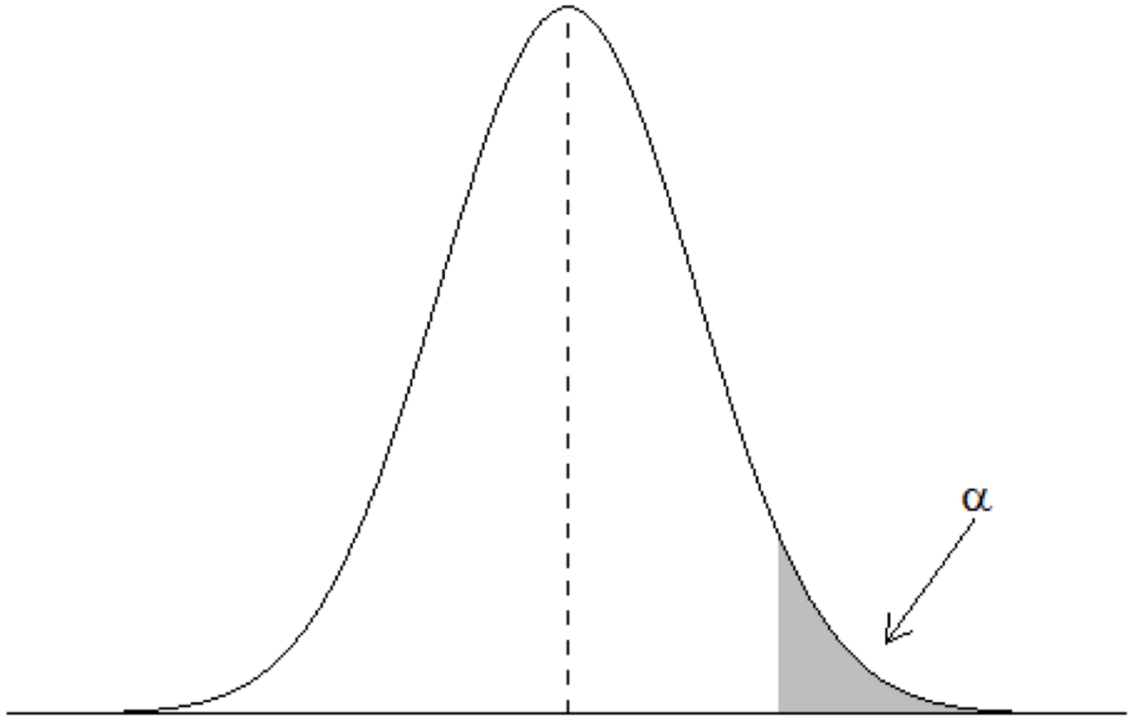
Fuente: Hypothesis Testing ,ck12

Prueba de hipótesis de una cola

- Utilizaríamos una prueba de hipótesis de una sola cola cuando se anticipe la dirección de los resultados o solo nos interese una sola dirección de los resultados.

- Ejemplo: decíamos que el puntaje promedio del SAT de los estudiantes de último año que se gradúan es mayor que 1,110.

$$H_0 : \mu \leq 1100$$
$$H_a : \mu > 1100$$



Visual Representation of Critical Region:
One-tailed

Fuente: Hypothesis Testing .ck12

Prueba de hipótesis para proporciones

- Seleccionamos una muestra y calculamos los estadísticos descriptivas sobre los datos de la muestra. Específicamente, calculamos el tamaño de la muestra (n) y la proporción de la muestra que se calcula tomando la relación entre el número de éxitos y el tamaño de la muestra,

$$\hat{p} = \frac{x}{n}$$

- Luego determinamos el estadístico para la prueba de hipótesis. La fórmula para el estadístico de prueba es:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Válido para muestras grandes

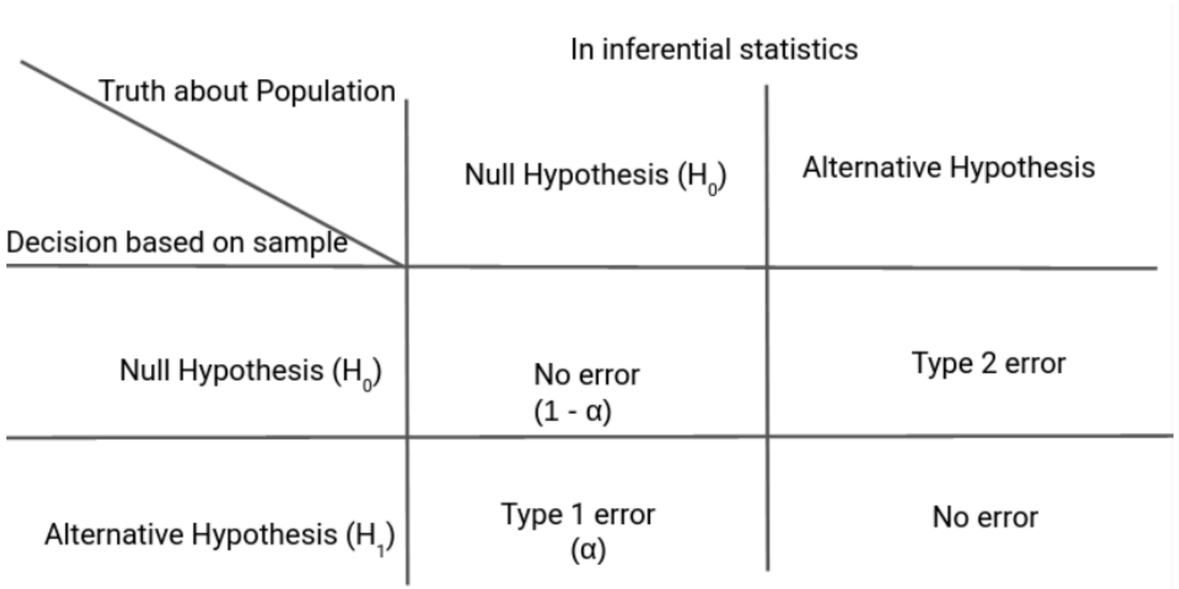
$$\min\left(n\hat{p},\, n\left(1-\hat{p}\right)\right) \geq 5$$

Fuente: Hypothesis Testing ,ck12

Errores Tipo I y Tipo II

Hay dos categorías de errores que se pueden cometer.

- **Tipo I:** rechazar la hipótesis nula cuando, de hecho, la hipótesis nula debería mantenerse.
- **Tipo II:** aceptar la hipótesis nula cuando debería ser rechazada.



Resumen

- La prueba de hipótesis implica realizar conjeturas sobre una población basándose en una muestra extraída de la población. Generamos hipótesis nulas y alternativas basadas en la media de la población para probar estas conjeturas.
- Establecemos regiones críticas en función del nivel de significancia o niveles alfa (α). Si el valor de la estadística de prueba cae en estas regiones críticas, podemos rechazarlo.
- Cuando tomamos una decisión sobre una hipótesis, hay cuatro resultados posibilidades y diferentes y dos tipos diferentes de errores. Un error de Tipo I es cuando rechazamos la hipótesis nula cuando es verdadera y un error de Tipo II es cuando no rechazamos la hipótesis nula, incluso cuando es falsa.

Fuente: Hypothesis Testing ,ck12

Conceptos

UNIVERSIDAD DE MONTEVIDEO

- El Z-Score es una prueba de significación estadística que ayuda a decidir si rechazar o no la hipótesis nula.
- El Z-Score son medidas de desviación estándar.
- El valor p es la probabilidad de que haya rechazado falsamente la hipótesis nula.
- Los valores p son probabilidades.
- Las puntuaciones Z muy altas o muy bajas (negativas), asociadas con valores p muy pequeños, se encuentran en las colas de la distribución normal.
- Para rechazar la hipótesis nula, debe realizar un juicio subjetivo sobre el grado de riesgo que está dispuesto a aceptar por estar equivocado. Este grado de riesgo a menudo se expresa en términos de valores críticos y / o niveles de confianza.

TWO SAMPLE Z TEST

Ejemplo

Se sabe que la cantidad de un determinado elemento en la sangre varía con una desviación estándar de 14,1 ppm (partes por millón) para los donantes de sangre masculinos y de 9,5 ppm para las donantes femeninas. Las muestras aleatorias de 75 donantes masculinos y 50 femeninos arrojan medias de concentración de 28 y 33 ppm, respectivamente. ¿Cuál es la probabilidad de que las medias poblacionales de concentraciones del elemento sean las mismas para hombres y mujeres?

Null hypothesis: $H_0: \mu_1 = \mu_2$

or $H_0: \mu_1 - \mu_2 = 0$

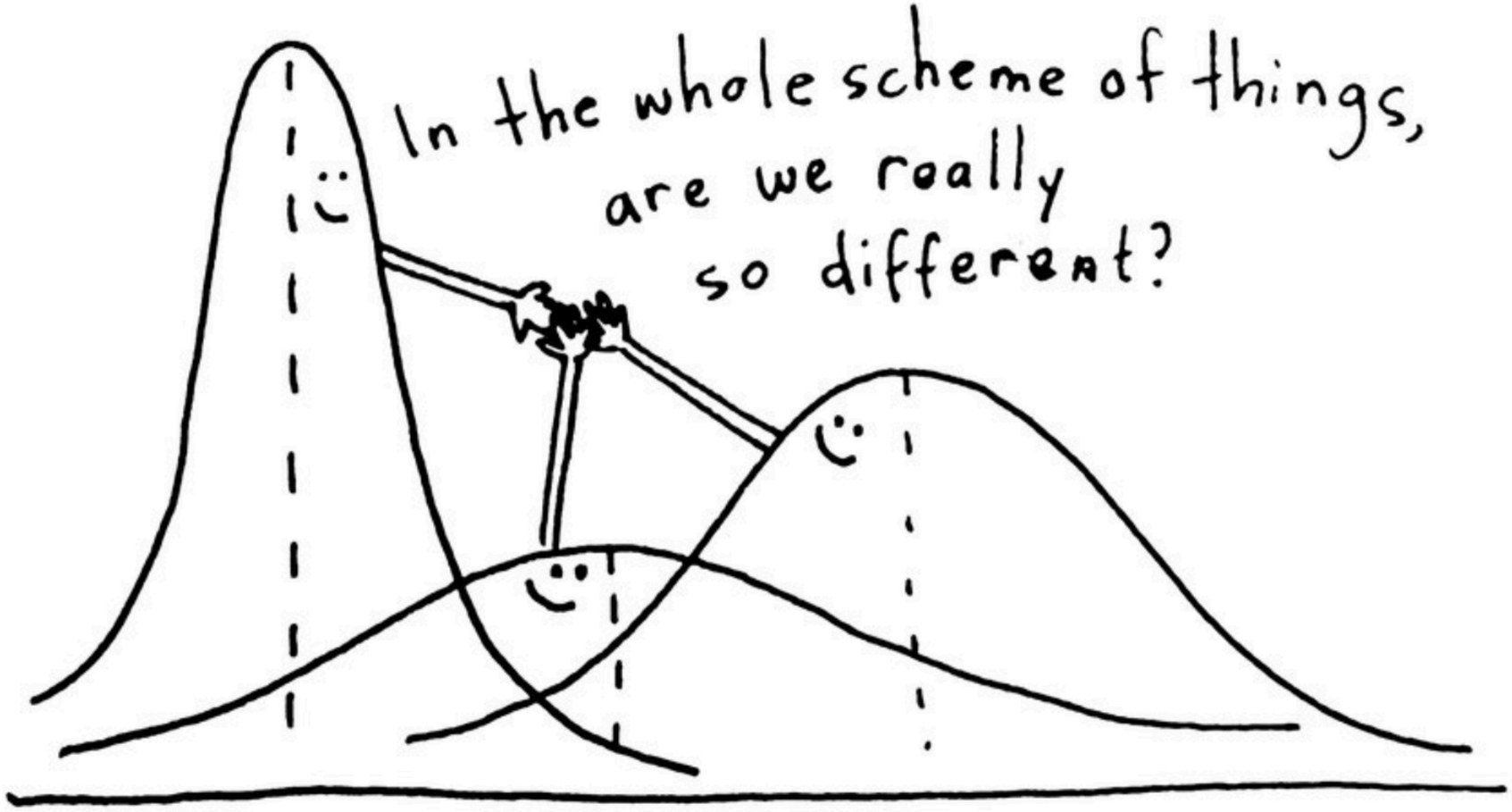
alternative hypothesis: $H_a: \mu_1 \neq \mu_2$

or: $H_a: \mu_1 - \mu_2 \neq 0$ $z = \frac{28-33-0}{\sqrt{\frac{14.1^2}{75}+\frac{9.5^2}{50}}} = \frac{-5}{\sqrt{2.65+1.81}} = -2.37$ ⟶

0,0089 x 2 (dos colas) = 0.0178 < 0.05 (alfa) -> no se rechaza H0.

ANOVA TEST

ANOVA



ANOVA

ANOVA, que significa análisis de varianza, es una prueba estadística que se utiliza para analizar la diferencia entre las medias de más de dos grupos.

- Null Hypothesis – There is no significant difference among the groups
- Alternate Hypothesis – There is a significant difference among the groups

Ejemplo

- Un grupo de pacientes psiquiátricos está probando tres terapias diferentes y se quiere ver si una terapia es mejor que las otras.
- Un fabricante tiene dos procesos diferentes para hacer bombillas y se quiere saber si un proceso es mejor que el otro.
- Los estudiantes de diferentes universidades toman el mismo examen y se quiere ver si una universidad supera a la otra.

ANOVA

Un one-way ANOVA Test se utiliza cuando se haya recopilado datos sobre una variable independiente categórica y una variable dependiente cuantitativa. La variable independiente debe tener al menos tres niveles (es decir, al menos tres grupos o categorías diferentes).

Ejemplo

- Su variable independiente es el uso de las redes sociales, y asigna grupos a niveles bajo, medio y alto de uso de las redes sociales para averiguar si hay una diferencia en las horas de sueño por noche.

- Su variable independiente es el tipo de fertilizante y trata los campos de cultivo con las mezclas 1, 2 y 3 para averiguar si hay una diferencia en el rendimiento del cultivo.

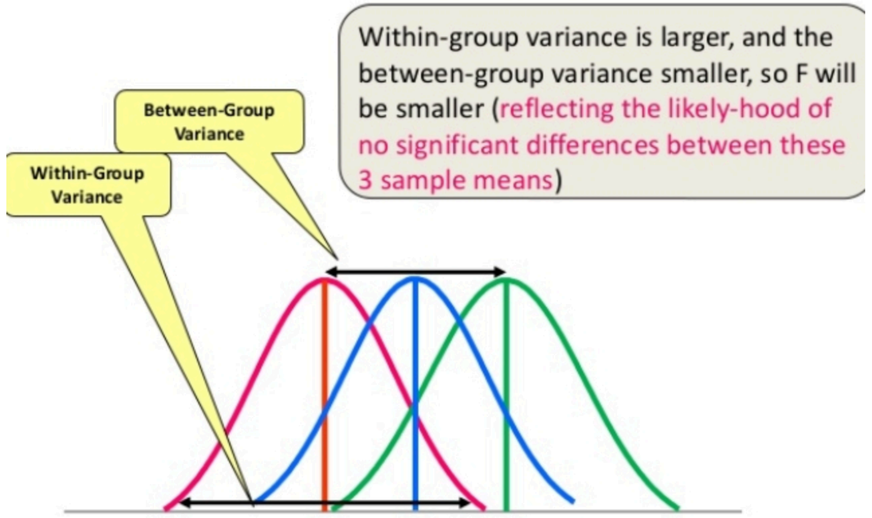
ANOVA

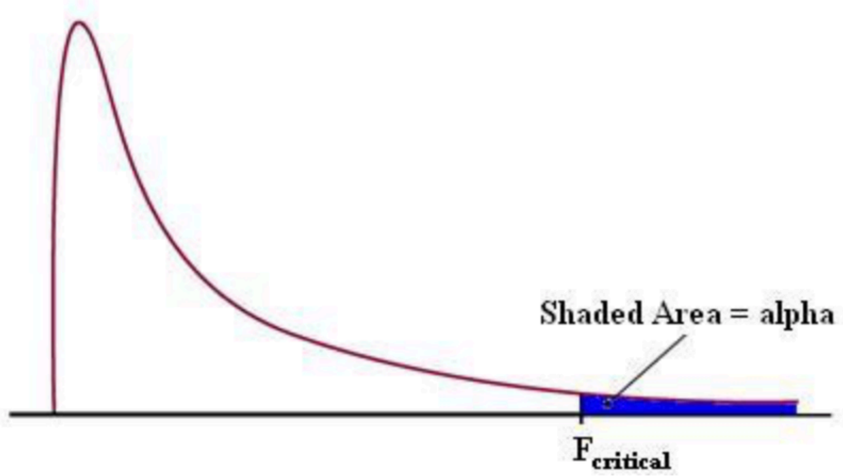
UNIVERSIDAD DE MONTEVIDEO

Estadística F

La estadística que mide si las medias de diferentes muestras son significativamente diferentes o no se llama F-Ratio. Cuanto más bajo el F-Ratio, más similares son las medias muestrales. En ese caso, no podemos rechazar la hipótesis nula.

F = Variabilidad entre grupos / Variabilidad dentro del grupo





X2 TEST

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

- El **test estadístico chi-cuadrado** de Pearson es un ejemplo de una **prueba de independencia entre variables categóricas.**

- La prueba de chi-cuadrado compara una tabla de contingencia observada con una tabla de valores esperados y determina si las variables categóricas son independientes.

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

Tabla de contingencia

	Pass	Fail	Total
Attended	30	20	50
Skip	20	30	50
Total	50	50	100

$$\frac{\text{row total x column total}}{\text{grand total}}$$

Tabla de valores esperados

	Pass	Fail	Total
Attended	25	25	50
Skip	25	25	50
Total	50	50	100

- H0: Salvar el curso no está asociada con la asistencia a clase.
- HA: Salvar el curso está asociada con la asistencia a clase.

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

Tabla de contingencia

	Pass	Fail	Total
Attended	30	20	50
Skip	20	30	50
Total	50	50	100

$$\frac{\text{row total x column total}}{\text{grand total}}$$

Tabla de valores esperados

	Pass	Fail	Total
Attended	25	25	50
Skip	25	25	50
Total	50	50	100

Estadístico X2

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

=

((30-25)^2/25) + ((20-25)^2/25) +

((20-25)^2/25) + ((30-25)^2/25)

=

1 + 1 + 1 + 1 = 4

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

Grados de libertad

- Los grados de libertad indican cuántos números de la tabla son realmente independientes.
- (R-1)*(C-1)
- En una tabla de 2x2 -> (2-1)*(2-1) = 1
- Una vez que se ponga un número en una celda de una tabla de 2x2, los totales determinan el resto de los valores.
- Utilizar los grados de libertad para decidir la probabilidad o el valor p de que las variables sean independientes.

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

- Si Estadístico >= Valor crítico: resultado significativo, se rechaza la hipótesis nula (H0), dependiente.
- Si Estadística < Valor crítico: resultado no significativo, no se rechaza la hipótesis nula (H0), independiente.

$$\chi^2 = \sum \frac{(n_{xy} - e_{xy})^2}{e_{xy}}$$

$$\chi^2 \sim \text{chi-squared}\left((r-1)(c-1)\ \text{df}\right)$$

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

Resultados e Interpretación

- Se supone un nivel de significancia igual a 0.05.
- => valor p de 0.0455.
- Esto se interpreta como una **probabilidad del 4,6% de que la hipótesis nula sea correcta**. En otras palabras, si la distribución de estos datos se debe completamente al azar, entonces se tiene un 4,6% de posibilidades de encontrar una discrepancia entre las distribuciones observadas y esperadas que sea al menos tan extrema.

X2 Test

Datos Categóricos

UNIVERSIDAD DE MONTEVIDEO

- La prueba de Chi-cuadrado solo está destinada a **probar la probabilidad de independencia** de una distribución de datos.

- **NO explica ningún detalle sobre la relación entre las variables.** Si se desea calcular cuánto más probable es que alguien que asista a clase salve el examen, la prueba de Chi-cuadrado no será muy útil.

- Sin embargo, una vez que haya determinado la probabilidad de que las dos variables estén relacionadas (usando la prueba de Chi-cuadrado), se pueden usar otros métodos para explorar su interacción con más detalle.