

Análisis Inteligente de Datos

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Profesora: Mónica Cantoni

Ayudantes: Cecilia Oliva

Fabiana Rossi

Pamela Pairo

Clase 02 – Probabilidades, Teorema de Bayes , Muestreo, Test de hipótesis



Agenda

- Probabilidades
- Teorema de Bayes
- Muestreo
- Test de Hipótesis

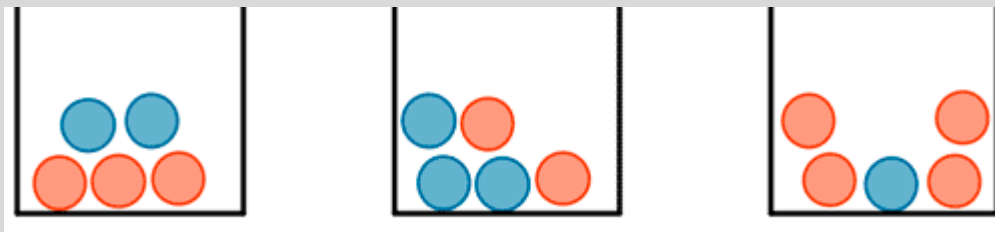
Revisión del Concepto de Probabilidad

Enfoques:

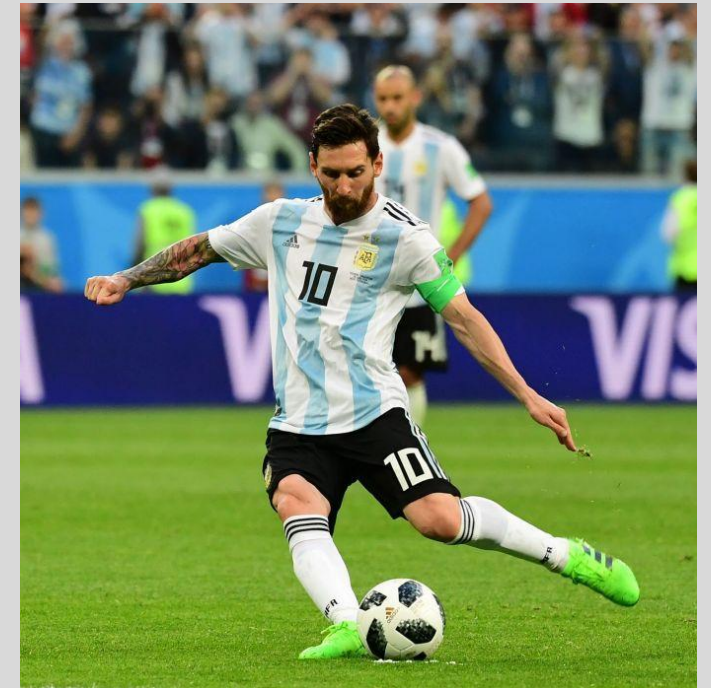
Clásico



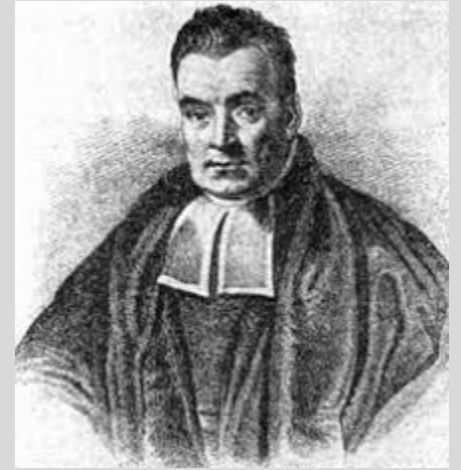
Frecuencista



Subjetivo



Probabilidades a priori y a posteriori



A priori Verosimilitud

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(B)}$$

A posteriori

Marginal

Teorema de Bayes

- Si la realización de un acontecimiento aleatorio B depende necesariamente que se produzca uno de los acontecimientos excluyentes A_1, A_2, \dots, A_k y se sabe que B se ha cumplido, la probabilidad para que sea uno determinado, por ejemplo A_i , el que se haya cumplido conjuntamente con B , está dada por:

Teorema de Bayes o probabilidad de las causas

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(B)}$$

- $P(A_i)$ son las probabilidades **a priori**. Representan las probabilidades de los estados del suceso antes de tomar información adicional.
- $P(B/A_i)$ es la **verosimilitud** de la hipótesis. Representan la probabilidad condicional de observar información adicional B cuando el estado del suceso que se presente sea A_i .
- $P(A_i/B)$ es la probabilidad **a posteriori**. Representa la probabilidad de ocurrencia del estado del suceso A_i dada la información adicional B.

El enfoque Bayesiano

- Este enfoque hace una revisión de la información probabilística a priori disponible a fin de tomar en cuenta información adicional que sea posible recopilar.
- La información a priori puede ser objetiva o subjetiva. Objetiva estimando la probabilidad por la frecuencia relativa (enfoque frecuentista) y subjetiva por conocimiento y/o experiencia del decisor o expertos (enfoque subjetivo).
- La probabilidad subjetiva asignada por un individuo a un conjunto de sucesos futuros inciertos puede diferir de la que asignaría otro individuo.
- La información adicional se puede adquirir ya elaborada o mediante un experimento muestral.

Ejemplo:

- Existen en el mercado tres modelos de máquinas lavavajillas. La marca A1 provee el 50% del mercado, la A2 provee el 30% y la A3 el 20% restante.
- Se sabe además, que el 20% de todos los modelos de la marca A1 requieren un ajuste después de instalados, lo mismo ocurre con el 15% de los modelos de la marca A2 y con el 5% de la marca A3.
- Un usuario determinado que no puede distinguir exteriormente una máquina de otra advierte que la suya necesita ajuste. Como es una persona muy racional decide usar el Teorema de Bayes para corregir las probabilidades a priori que corresponden a la distribución porcentual del mercado usando como verosimilitud de la hipótesis la información adicional referente a la proporción de lavavajillas que necesitan ajuste según la marca.

Cuadro de cálculos:

Marca	$P(A_i)$	$P(B/A_i)$	$P(A_i)P(B/A_i)$	$P(A_i/B)$
A1	0,50	0,20	0,10	0,645
A2	0,30	0,15	0,045	0,290
A3	0,20	0,05	0,010	0,065
			0,155	

$P(A_i)$ son las probabilidades **a priori**. Representan las probabilidades de haber adquirido cada uno de los lavavajillas.

$P(B/A_i)$ es la **verosimilitud** de la hipótesis. Representan la probabilidad que las máquinas requieran service según sean de marcas A1, A2 o A3.

$P(A_i)P(B/A_i)$ es la probabilidad **conjunta** y representa la probabilidad que se presenten, en conjunto, máquinas que requieran service y sean de las marcas A1, A2 o A3.

Las probabilidades a posteriori obtenidas en la última columna indican que los lavavajillas de marca A1 necesitarían más service.

Aplicaciones

- Naïve Bayes o Bayes ingenuo es un clasificador bayesiano. Es un algoritmo de aprendizaje automático basado en la probabilidad condicional y en el teorema de Bayes.

Muestreo

- El muestreo es la principal técnica empleada para la reducción de datos.
- Es utilizada tanto para la investigación preliminar de los datos como para el análisis de datos final.
- Trabajar con muestras, en lugar de trabajar con el conjunto completo de datos de interés, reduce costos y tiempo.
- El muestreo en la minería de datos se utiliza porque procesar todo el conjunto de datos de interés es demasiado caro o requiere mucho tiempo.

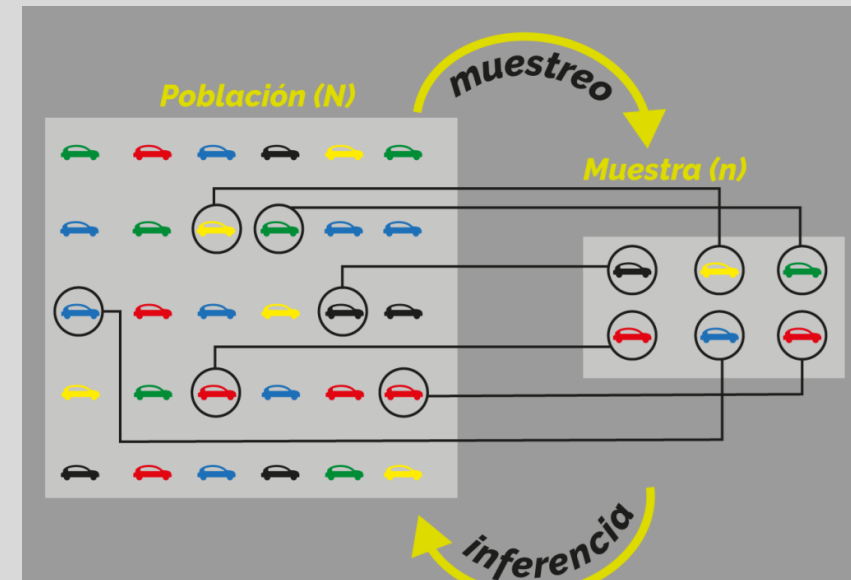
Muestreo

El principio clave para un muestreo eficaz es el siguiente:

- Usar una muestra funcionará casi tan bien como usar todo el conjunto de datos, si la muestra es representativa.
- Una muestra es representativa si tiene aproximadamente las mismas propiedades (de interés) que el conjunto original de datos.

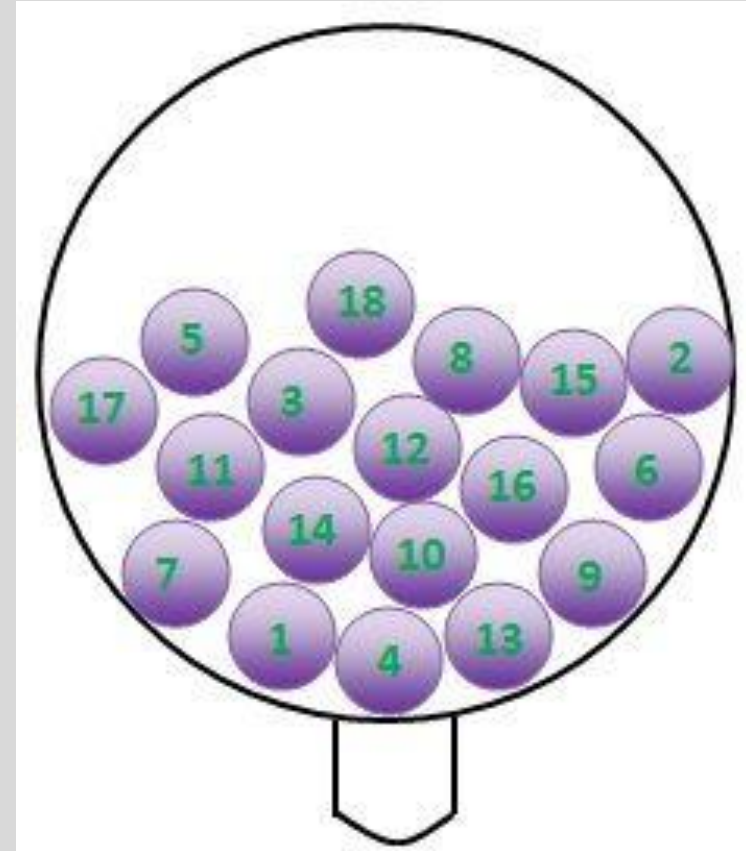


Técnicas de muestreo.



Muestreo aleatorio simple

- Cada elemento de la población tiene la misma probabilidad de ser elegido para formar parte de la muestra.
- Si se dispone de la lista enumerativa de todos los elementos de la población, se obtiene la muestra con la ayuda de números aleatorios.
- La selección de los elementos es similar a realizar un sorteo.

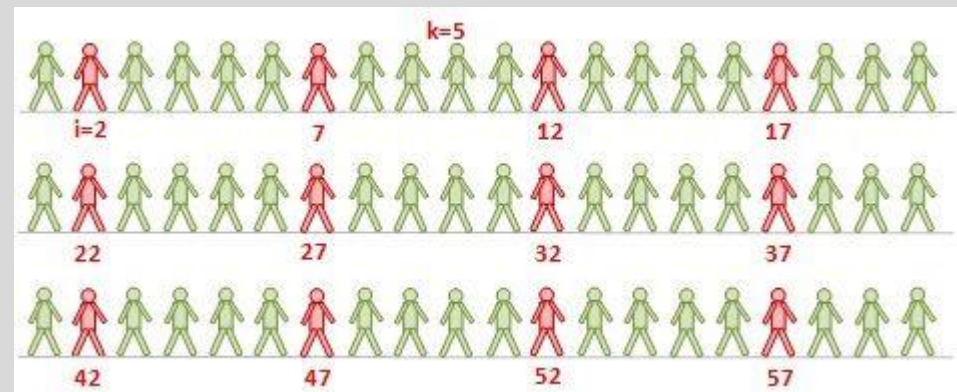


Muestreo sistemático

- Los elementos se seleccionan de la población en un intervalo uniforme que se mide respecto de tiempo, orden o espacio.
- Se emplea si existe una lista ordenada de los elementos de la población o cuando se sabe cuántos elementos componen esa población.
- La técnica consiste en tomar cada **k** elementos de una lista que contiene todos los elementos de una población, eligiendo al azar el primer elemento de la muestra.
- Se puede elegir aleatoriamente por dónde comenzar.

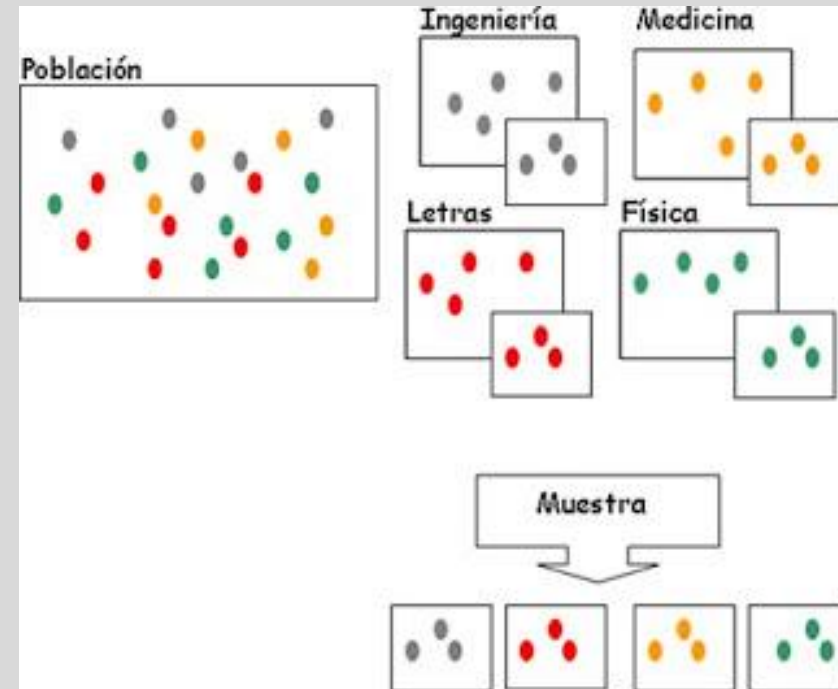
$$k = \frac{N}{n}$$

- K es la constante de elevación
- N es el tamaño de la población
- n es el tamaño de la muestra



Muestreo estratificado

- Es una variante del muestreo aleatorio simple.
- Se divide a la población en grupos homogéneos o estratos. Y los elementos dentro de cada estrato se seleccionan aleatoriamente de acuerdo con una de las siguientes reglas:
 1. Un número específico de elementos se extrae de cada estrato, y corresponde a la proporción de ese estrato en la población
 2. Igual número de elementos se extraen de cada estrato, y los resultados son valorados de acuerdo con la porción del estrato de la población total.
- Permite efectuar comparaciones entre los estratos conformados.



Muestreo con y sin reemplazo

- Muestreo con reemplazo
 - Los objetos no se eliminan de la población a medida que se seleccionan para la muestra.
 - En el muestreo con reemplazo, el mismo objeto se puede recoger más de una vez.
- Muestreo sin reemplazo
 - A medida que se selecciona cada elemento, se elimina de la población

Aplicaciones de Muestreo

- Datos desbalanceados
- Entrenamiento y prueba de modelos
- Validación cruzada

Test de Hipótesis

- **Test de Hipótesis**
- Nivel de Significación
- Errores tipo I y tipo II
- Potencia
- P-valor



Test de Hipótesis

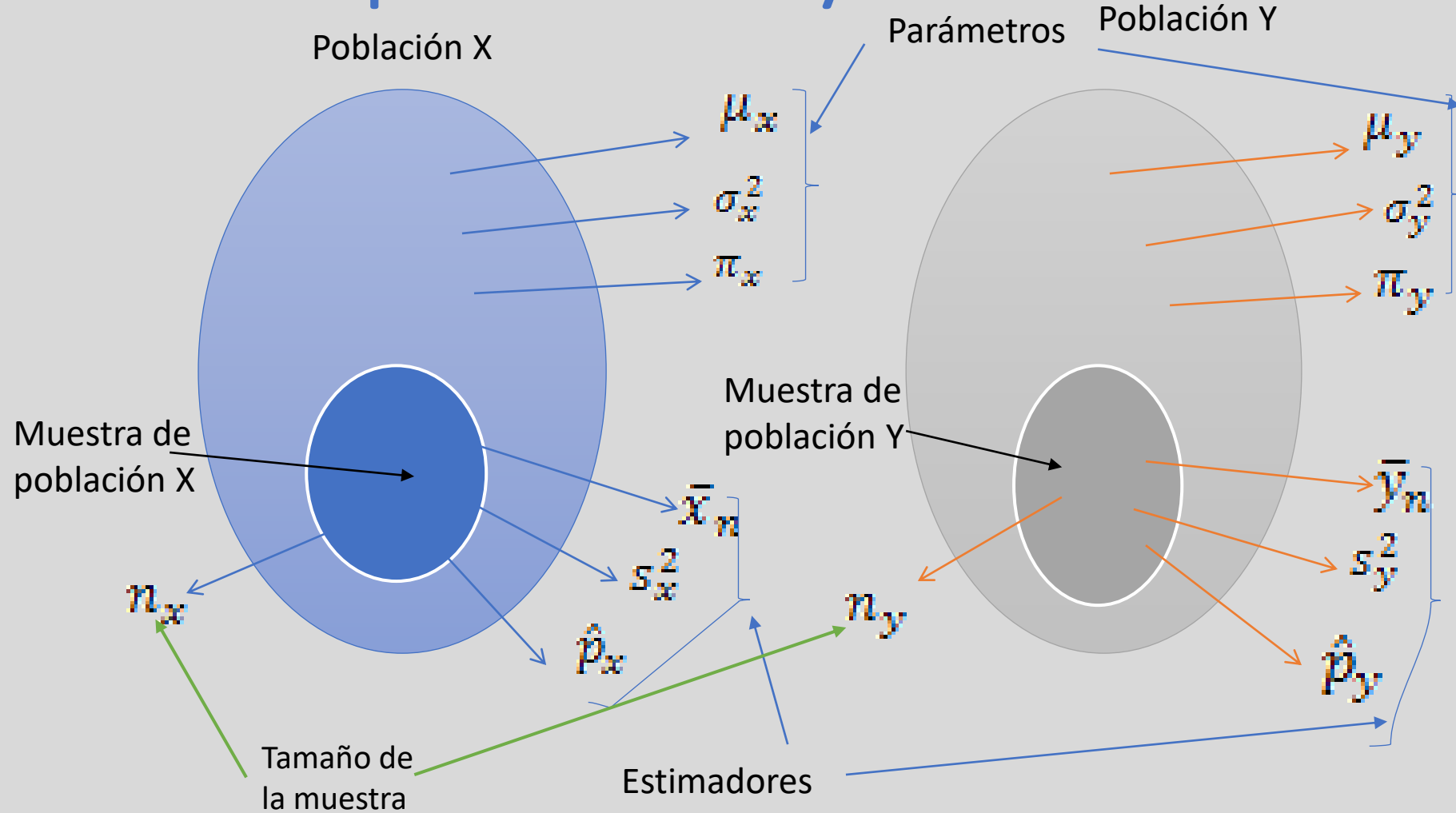
- H_0 Hipótesis nula
- H_1 Hipótesis alternativa
- La hipótesis nula H_0 es la que se contrasta
- La hipótesis alternativa H_1 es la que se investiga
- Hipótesis paramétricas: se refieren al valor de un parámetro
- Hipótesis no paramétricas o de libre distribución: no se refieren al valor del parámetro

Test de Hipótesis

- Parámetros
- Estimadores o estadísticos

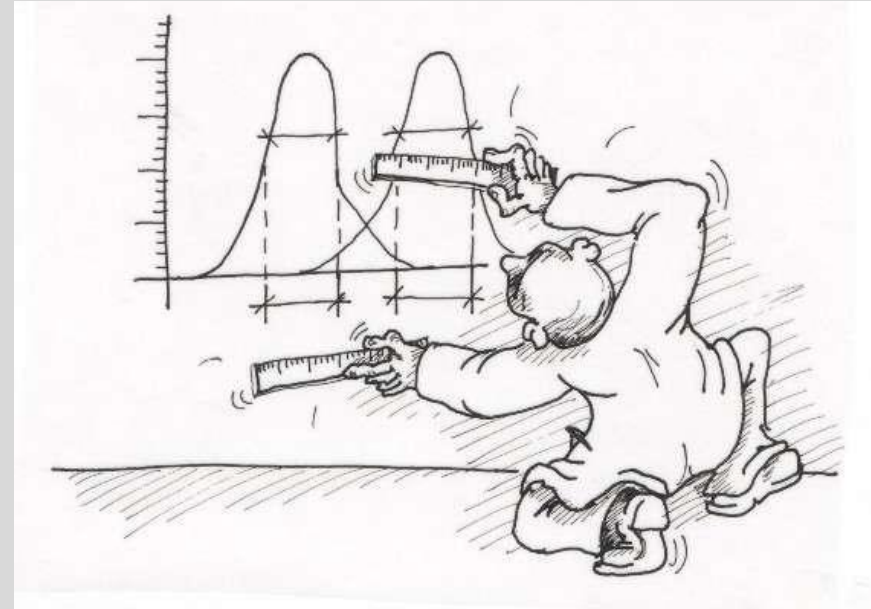
Test de Hipótesis

Notación de parámetros y estimadores



Agenda

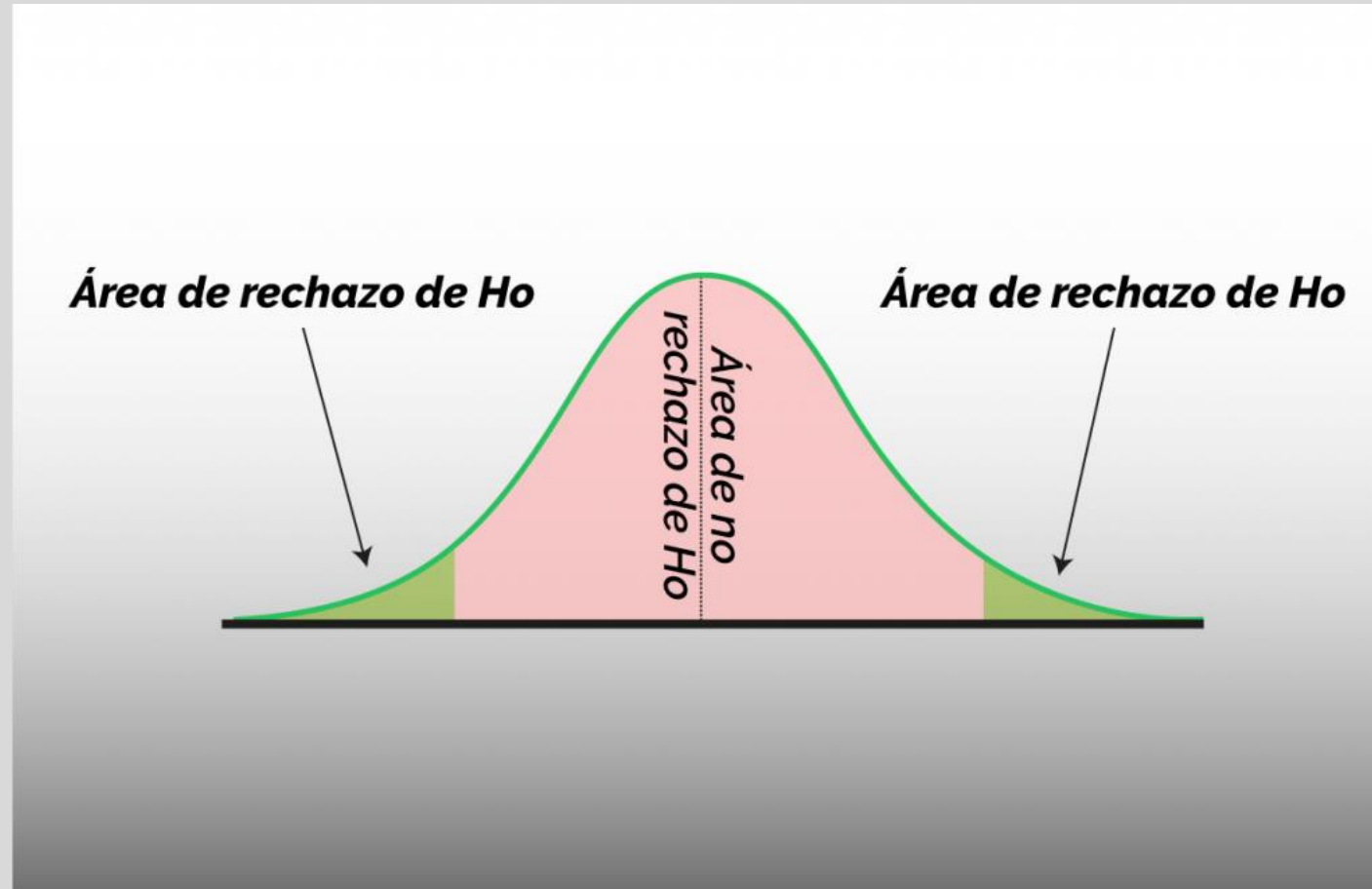
- Test de Hipótesis
- **Nivel de Significación**
- Errores tipo I y tipo II
- Potencia
- P-valor



Nivel de Significación

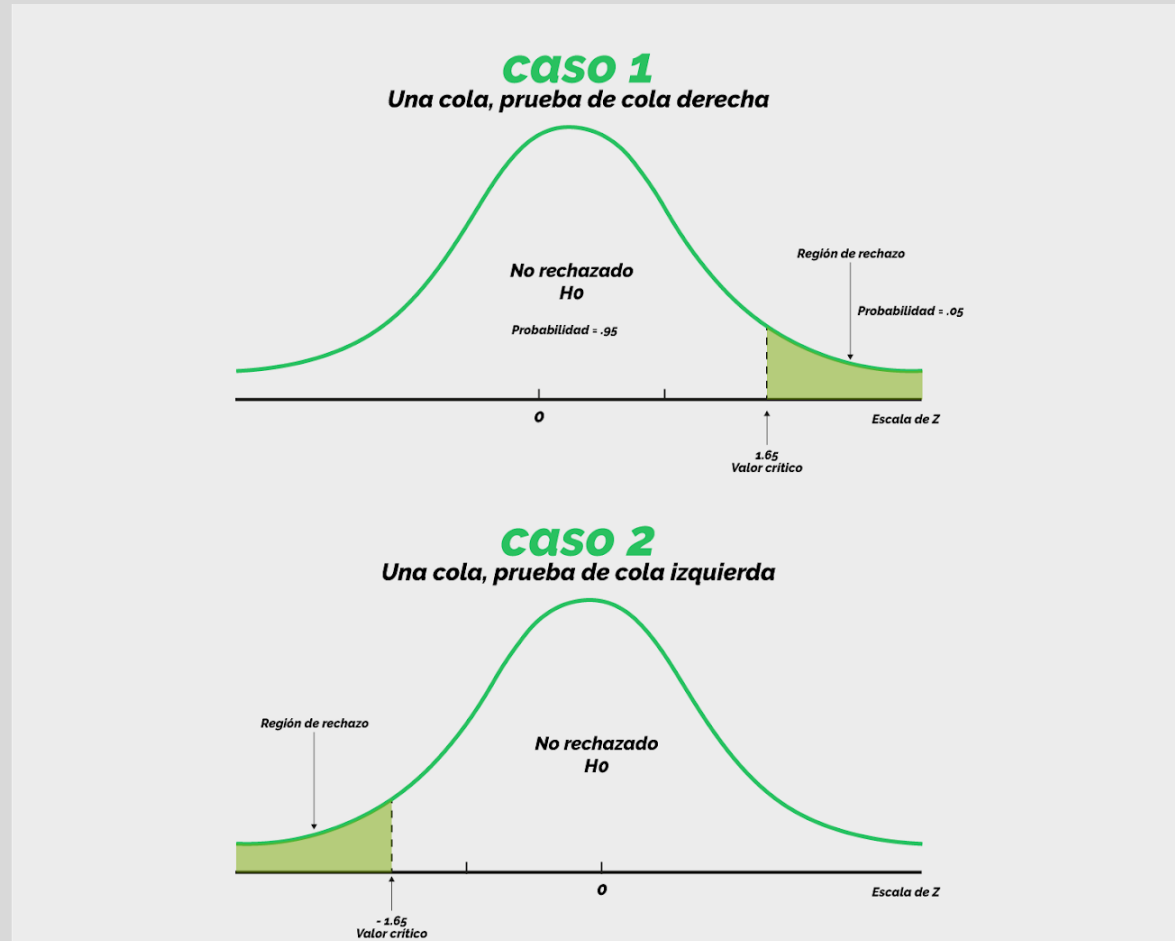
- El nivel de significación de un test es la máxima probabilidad de rechazar la hipótesis nula H_0 cuando ésta es cierta
- Riesgo máximo admisible para rechazar H_0 cuando es cierta
- Lo elige el investigador se denota con α
- Generalmente es un valor entre 0.01 y 0.10

Criterios de rechazo



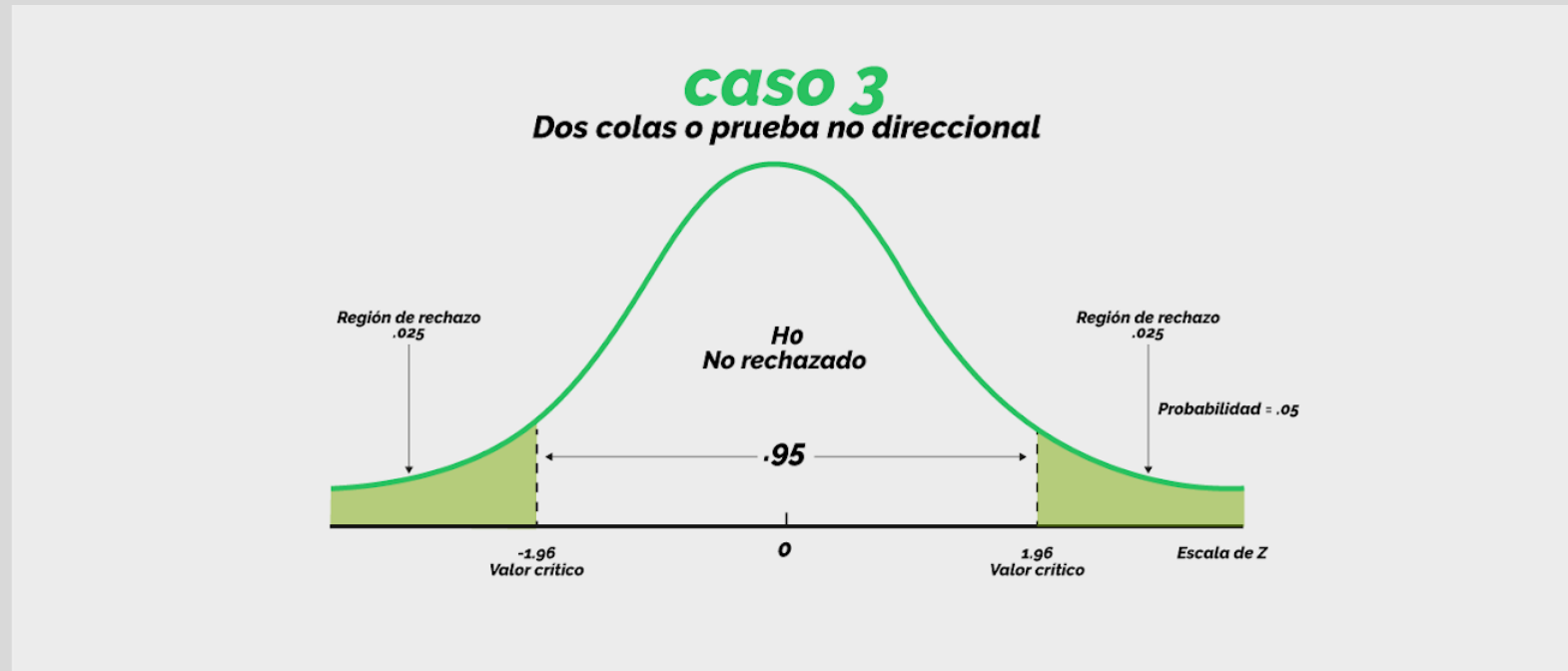
Tipos de contraste

- Unilateral



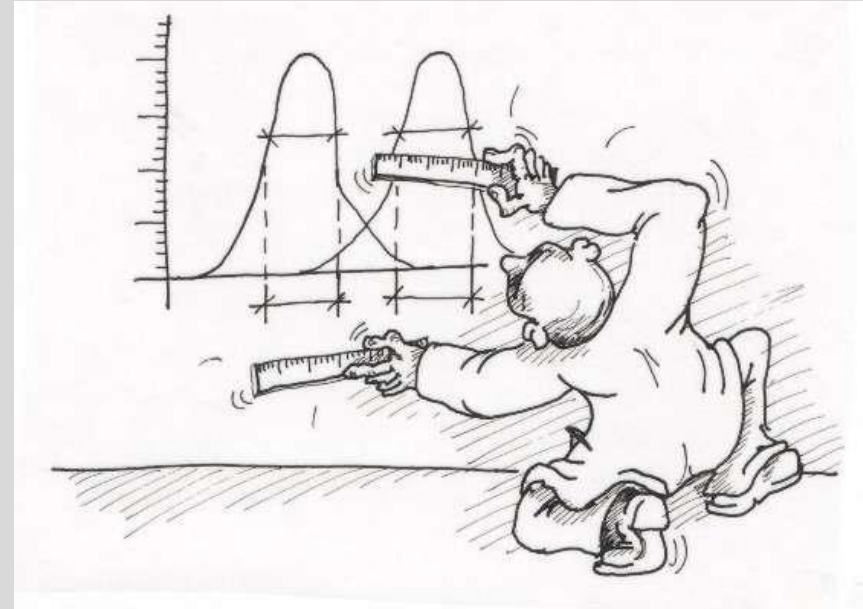
Tipos de contraste

- Bilateral



Agenda





- Test de Hipótesis
- Nivel de Significación
- **Errores tipo I y tipo II**
- Potencia
- P-valor



Decisiones en los test de hipótesis

		La verdadera situación puede ser	
		H_0 es verdadera	H_0 es falsa
Acción	No rechazar H_0	Decisión correcta	Decisión incorrecta (error de tipo II)
	Rechazar H_0	Decisión incorrecta (error de tipo I)	Decisión correcta

Decisiones en los test de hipótesis

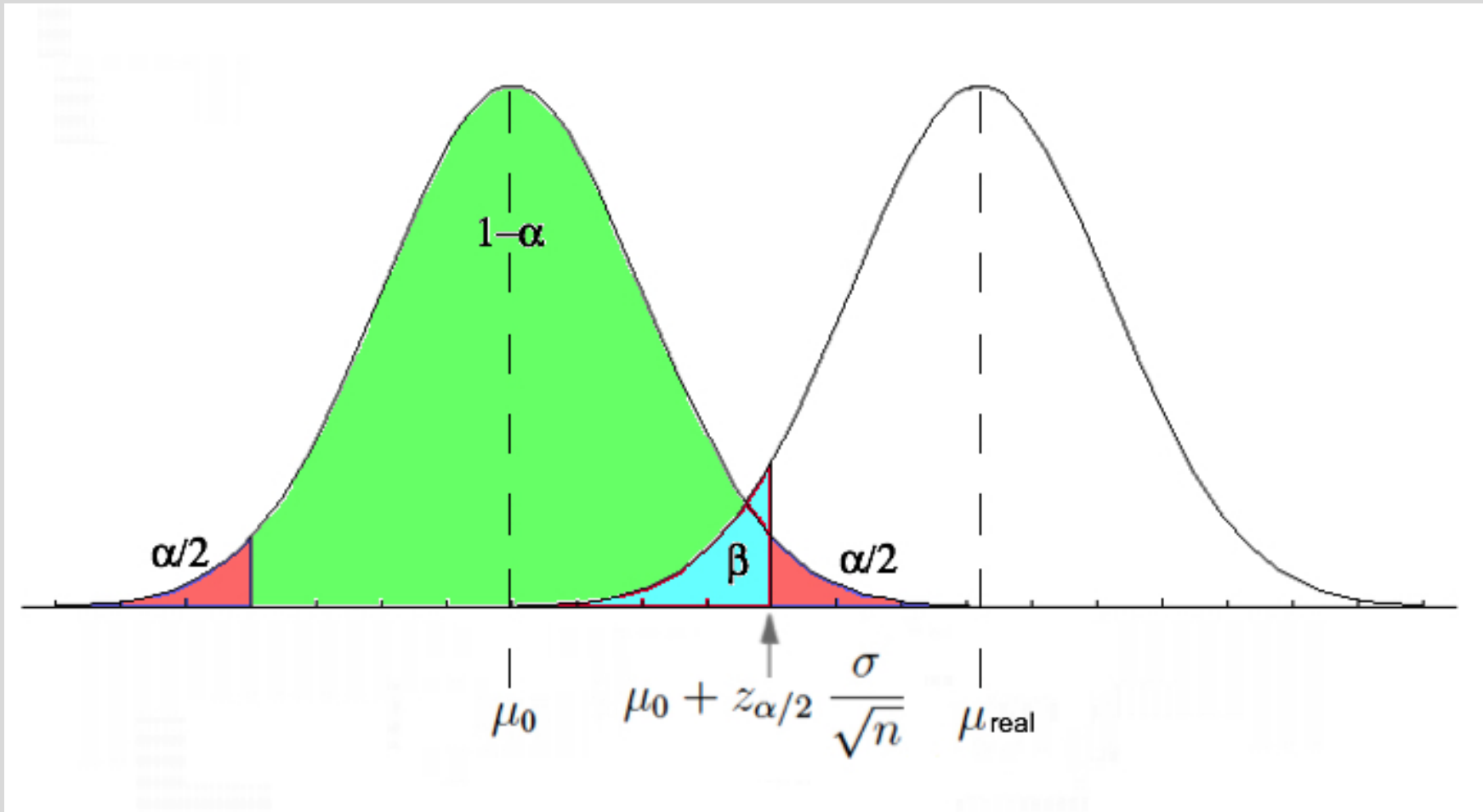
La verdadera situación puede ser			
		H_0 es verdadera	H_0 es falsa
Acción	No rechazar H_0	 Verdadero negativo	 Falso negativo
	Rechazar H_0	 Falso positivo Falso rechazo	 Verdadero positivo Rechazo cierto

Relación entre los Errores tipo I y tipo II

		La verdadera situación puede ser	
		H_0 es verdadera	H_0 es falsa
Acción	No rechazar H_0	$1 - \alpha$ (Nivel de confianza)	β
	Rechazar H_0	α (Nivel de significación)	$1 - \beta$ (Potencia de la prueba)

Relación entre los Errores tipo I y tipo II

Test bilateral

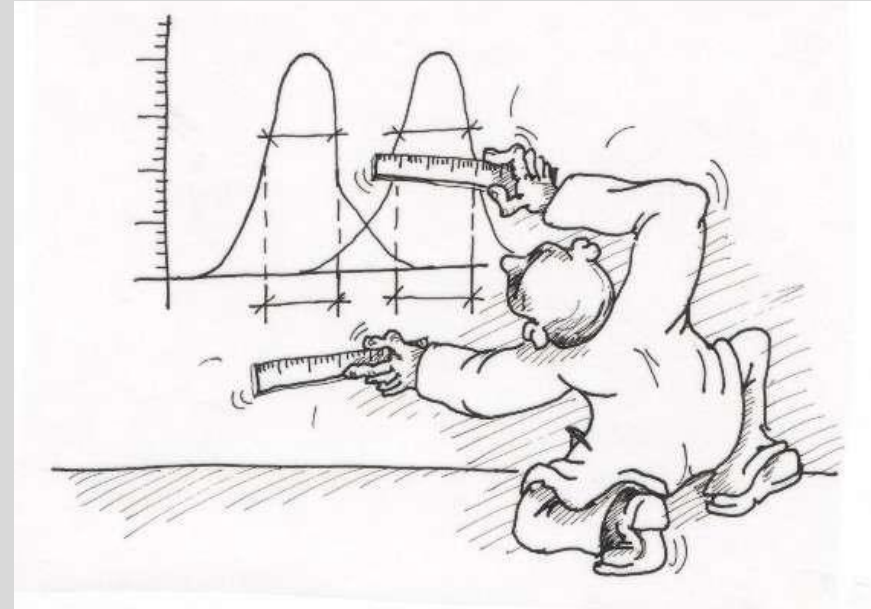


Errores Tipo I y Tipo II

- Error tipo I: rechazar la hipótesis nula H_0 cuando es verdadera.
- Error tipo II: no rechazar la hipótesis nula H_0 cuando es falsa.
- α = Nivel de significación = $P(\text{error de tipo I}) = P(\text{se rechaza } H_0/H_0 \text{ es verdadera})$
- $1 - \alpha$ = Nivel de confianza = $1 - P(\text{error de tipo I}) = P(\text{no se rechaza } H_0/H_0 \text{ es verdadera})$
- β = $P(\text{error de tipo II}) = P(\text{no se rechaza } H_0/H_0 \text{ es falsa})$
- $1 - \beta$ = Potencia de la prueba = $1 - P(\text{error de tipo II}) = P(\text{se rechaza } H_0/H_0 \text{ es falsa})$
- α = Probabilidad de rechazar H_0 / H_0 verdadera
- β = Probabilidad de no rechazar H_0 / H_0 falsa
- Con los valores de β se grafica la Curva de Operación Característica.

Agenda

- Test de Hipótesis
- Nivel de Significación
- Errores tipo I y tipo II
- **Potencia**
- P-valor

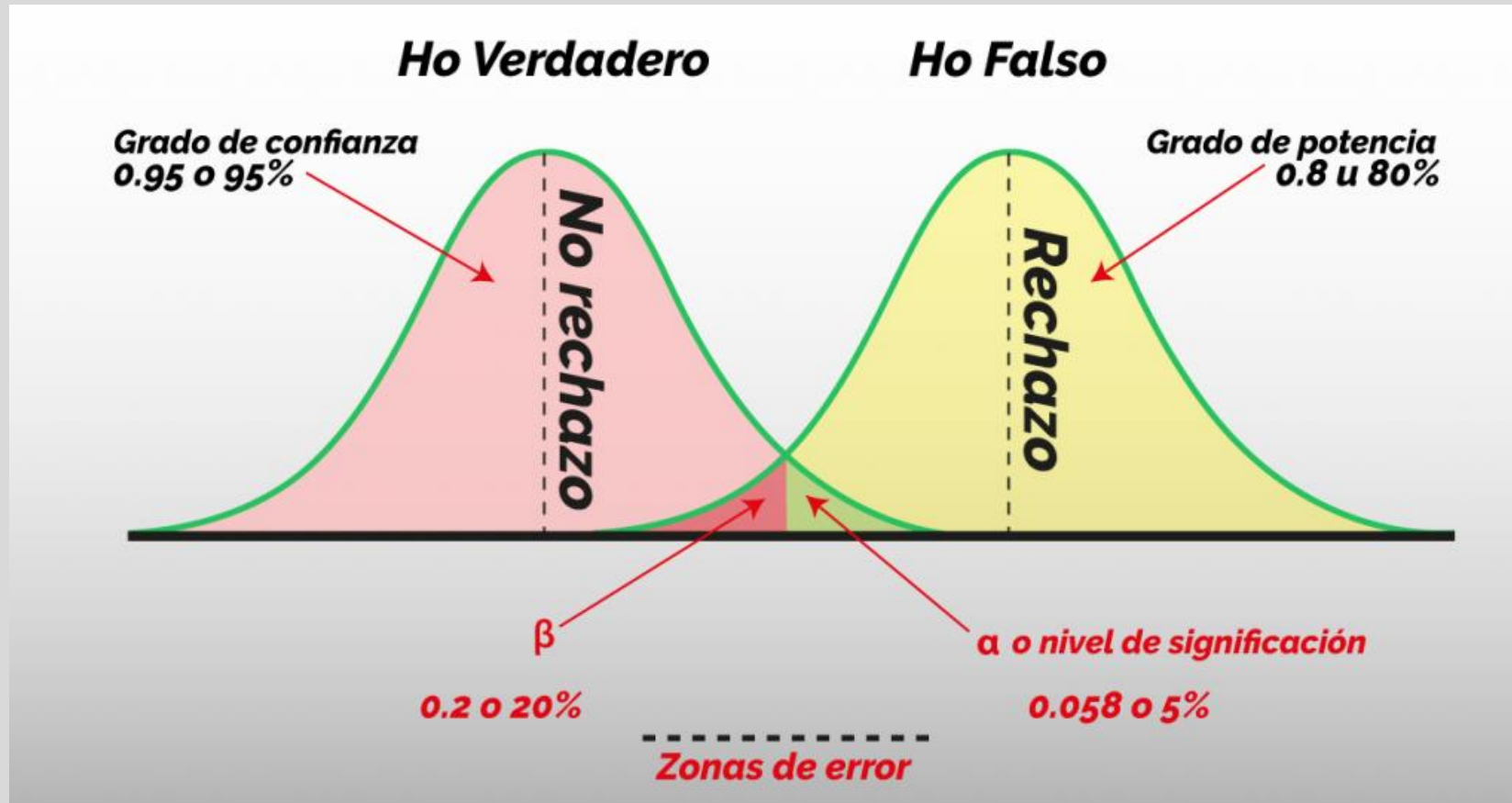


Potencia de un contraste

- $1 - \beta$ = Potencia de la prueba.
- Con los valores de $1 - \beta$ se grafica la Curva de Potencia.
- Es la medida de la sensibilidad del procedimiento de las pruebas de hipótesis, puesto que determinar la probabilidad de rechazo correcto de la hipótesis nula en diferentes circunstancias.

Relación entre los Errores tipo I y tipo II

Test unilateral



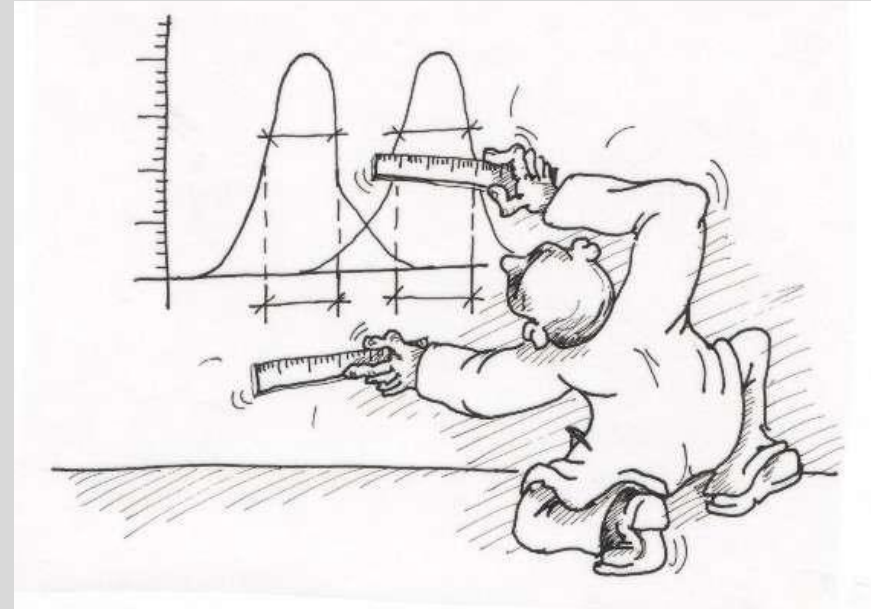
Observaciones entre errores tipo I, tipo II y tamaño de la muestra

Relación entre α , β y tamaño de muestra:

- Una vez conocidos dos de los tres valores, se puede calcular el otro.
- Para un determinado nivel de significación α , al aumentar el tamaño de la muestra se reduce β .
- Para un determinado tamaño de muestra, al disminuir α aumenta β , y viceversa.

Agenda

- Test de Hipótesis
- Nivel de Significación
- Errores tipo I y tipo II
- Potencia
- **P-valor**



Contraste de hipótesis: el p-valor

- El contraste de hipótesis es la técnica estadística utilizada frecuentemente en la literatura científica.
- La idea inicial de las pruebas de significación se debe a Fisher.
- La idea de Fisher consiste en realizar una comparación calculando una probabilidad, el famoso p-valor o nivel de significación.
- Usar ese valor como índice de la fuerza probatoria de los datos contra la hipótesis nula, cuando menor sea p-valor mayor será la carga de la prueba en contra de la hipótesis nula.
- Propone además un valor de 0,05 como punto de corte conveniente.

Utilización del p-valor

- El p-valor, se utiliza en pruebas de hipótesis de una sola población (media, varianza y proporción), para comparar dos poblaciones (diferencia de medias, cociente de varianzas, diferencia entre proporciones) o más de dos poblaciones (análisis de la varianza), en pruebas de bondad de ajuste, independencia y en pruebas de significancia en los modelos de regresión.

Concepto de p-valor

- **¿Qué es el p-valor?**
- Según diferentes autores:
- **(Newbold)**
- Es un método para examinar el contraste de la hipótesis nula.
- El ***p-valor*** es la probabilidad de obtener un valor del estadístico del contraste igual de extremo o más que el valor efectivo obtenido cuando la hipótesis nula es verdadera.
- El ***p-valor*** es el menor nivel de significación al que puede rechazarse una hipótesis nula, dado el estadístico muestral observado

Rechazar H_0 si $p\text{-valor} < \alpha$

- Es el nivel de significación más bajo al que puede rechazarse la hipótesis nula.

Concepto de p-valor

- **(Anderson)**
- El p-valor es una probabilidad que aporta una medida de la evidencia suministrada por la muestra contra la hipótesis nula.
- P-valores pequeños indican una evidencia mayor contra la hipótesis nula.
- Se utiliza para determinar si la hipótesis nula debe ser rechazada.
- Para su cálculo se utiliza el valor del estadístico de prueba.

Resumiendo: P-valor

- Es el peso de la evidencia de la prueba o nivel observado de significancia.
- El p-valor se basa en una probabilidad y otro método que permite decidir si se rechaza o no la hipótesis nula.
- Si suponemos que la hipótesis nula es verdadera, el p-valor es la probabilidad de obtener un resultado de la muestra que sea al menos tan improbable como lo que se observa.
- Fundamentación de la utilización del valor p en una prueba de hipótesis:
- En la medida que el p-valor sea menor, más fuerte es la conclusión del rechazo de la hipótesis nula; o sea si p es muy pequeño nos está diciendo que el valor observado ha sido un valor muy lejano al valor esperado propuesto, y que siendo válida dicha hipótesis es muy poco probable que aparezcan valores del orden del valor observado. Por el contrario si p-valor es grande, lo que está diciendo es que el valor observado es parecido al valor hipotético propuesto, y que son valores muy probables de presentación cuando la hipótesis propuesta es válida.

¿Preguntas?

