Estadística

Clase 6

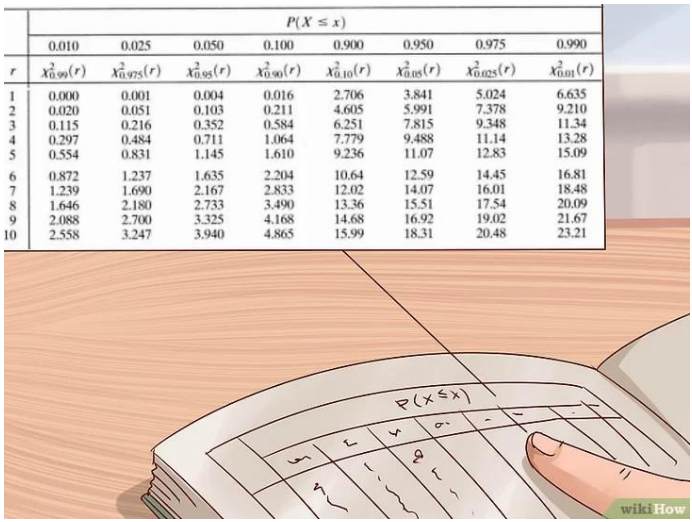
Rodrigo Del Rosso

20 de Mayo de 2022

Prueba de Hipótesis

Para una Población

	$P(X \leq x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21



En esta parte de la materia se estudiará el proceso para decidir si una determinada afirmación acerca de una población (o de varias poblaciones) está sustentada por una evidencia experimental obtenida mediante una o más muestras extraídas de dichas poblaciones bajo estudio.

En forma general, las afirmaciones se refieren al valor numérico desconocido de los parámetros estadísticos o formas funcionales desconocidas.

Dado que se emplea un Muestreo Probabilístico, la decisión se toma basándose en el valor de probabilidad de cometer errores en dicha decisión y consecuentemente en la acción que se realice.

Definiciones Básicas

Hipótesis Estadística

Se denomina a cualquier afirmación o aseveración que se formula acerca de cualquier característica poblacional (el valor numérico de un parámetro, la forma funcional de una población, etc)

Hipótesis Paramétrica

Se denomina así aquella hipótesis estadística planteada para controlar o verificar el valor numérica de un parámetro.

Se consideran 3 posibles situaciones del valor numérico del parámetro, a saber,

- El valor de θ es igual a un determinado valor postulado θ_0
- El valor de θ es mayor a un determinado valor postulado θ_0
- El valor de θ es menor a un determinado valor postulado θ_0

En función de los posibles valores del parámetro se realizará una determinada acción.

Cursos de Acción

Se denomina así a la acción que se llevaría a cabo, si se conociese el verdadero valor del parámetro θ.

Desigualdad equivalente a la Igualdad

Se denomina así aquella desigualdad entre el parámetro θ y el valor postulado θ_0 que provoca el mismo curso de acción que se llevaría a cabo con la igualdad entre el valor del parámetro θ y el valor postulado θ_0.

Ejemplo

Si el parámetro poblacional es $\theta = \mu$ y se poseen los siguientes cursos de acción,

$$Si\ \mu = \begin{cases} > 1500 \Rightarrow Se\ implementa\ el\ IMF \\ = 1500 \Rightarrow Se\ implementa\ el\ IMF \\ < 1500 \Rightarrow No\ Se\ implementa\ el\ IMF \end{cases}$$

La desigualdad $>$ posee el mismo curso de acción que la igualdad $=$.

Hipótesis Nula

Se denomina aquella hipótesis que establece que la diferencia entre el verdadero valor del parámetro y el valor que se postula es cero. Formalmente,

$$H_0: \theta = \theta_0$$

A fines prácticos esta igualdad puede estar acompañada o no por alguna de las dos desigualdades, según sea el curso de acción a seguir y la existencia o no de alguna desigualdad equivalente. Se pueden distinguir dos tipos de hipótesis nula,

- **Hipótesis Nula Única**

  Cuando no hay desigualdad equivalente

$$H_0: \theta = \theta_0$$

- **Hipótesis Nula Múltiple**

  Cuando hay desigualdad equivalente

$$H_0: \theta \geq \theta_0 \quad \text{ó} \quad H_0: \theta \leq \theta_0$$

Hipótesis Alternativa

Se denomina aquella hipótesis que debería cumplirse si la hipótesis nula no es cierta. Formalmente,

$$H_1: \theta = \theta_1$$

A fines prácticos, se pueden distinguir dos tipos de planteos,

- **Planteo Bilateral (a dos colas)**

$$H_0: \theta = \theta_0 \qquad H_1: \theta \neq \theta_0$$

- **Planteo Unilateral (a una cola)**

$$H_0: \theta \geq \theta_0 \qquad H_1: \theta < \theta_0$$

$$H_0: \theta \leq \theta_0 \qquad H_1: \theta > \theta_0$$

Prueba de la Hipótesis Nula

Se denomina así a un método estadístico con el cual, a partir de los datos de una muestra aleatoria, se decide acerca de la veracidad o falsedad de la Hipótesis Nula formulada, pudiéndose calcular la probabilidad de cometer un error en la decisión tomada.

La hipótesis que se prueba para decidir si debe ser rechazada o no, siempre es la hipótesis nula. En vez de decir “Prueba de la Hipótesis Nula” se dirá “Prueba de Hipótesis”.

Estadígrafo de Prueba

Se denomina a un estadígrafo $\boldsymbol{ep}$ apropiado con el que se realiza la prueba de hipótesis que mida la discrepancia $\boldsymbol{d}$ entre el parámetro a probar y el estimador correspondiente y, además, tiene una distribución de probabilidad conocida. Es una variable aleatoria que se genera transformando al estimador del parámetro.

$$\boldsymbol{ep} = \boldsymbol{d}(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta})$$

Región Crítica y de No Rechazo

Consiste en particionar al Dominio del Estadígrafo de Prueba en dos subconjuntos o regiones mutuamente excluyentes. Según a cuál de las dos regiones pertenezca el valor numérico del Estadígrafo de Prueba se rechaza o no la hipótesis nula.

Se denomina Región Crítica R_c al subconjunto del dominio del Estadígrafo de Prueba con el cual se rechaza la Hipótesis Nula.

Se denomina Región de No Rechazo R_a al subconjunto del dominio del Estadígrafo de Prueba con el cual no se rechaza la Hipótesis Nula.

Si hay una desigualdad equivalente, la R_c está formada por un subconjunto semicerrado. Se dice que la prueba es unilateral.

Si no hay desigualdad equivalente, la R_c está formada por dos subconjuntos semicerrados mutuamente excluyentes de igual tamaño. Se dice que la prueba es bilateral.

Se denomina punto crítico p_c a la frontera de la R_c. Es un punto de rechazo de la hipótesis.

Regla de Decisión

Se denomina aquella regla que establece las pautas para rechazar la hipótesis nula y se enuncia:

$$Si\ ep \in R_c \Rightarrow RH_0\ (Rechazo\ la\ Hipótesis\ Nula)$$

$$Si\ ep \notin R_c \Rightarrow No\ RH_0\ (No\ Rechazo\ la\ Hipótesis\ Nula)$$

Dada la Hipótesis Nula,

$$H_0: \theta = \theta_0$$

la regla de decisión establece que hay que rechazarla si, luego de obtener la muestra, hacer las mediciones correspondientes y calcular el valor numérico del Estadígrafo de Prueba, éste pertenece a la Región Crítica, y que no hay que rechazarla si el Estadígrafo de Prueba, no pertenece a la Región Crítica.

Error de Tipo I y II

Se denomina **Error de Tipo I (ETI)** al hecho de **rechazar la hipótesis nula cuando ésta es cierta**.

Puede suceder que a pesar de que el valor numérico del Estadígrafo de Prueba pertenezca a la Región Crítica, entonces se rechaza y dicha hipótesis sea verdadera. En esta situación estaríamos cometiendo un error en la decisión tomada. La probabilidad de cometer **este error se denomina α o nivel de significación y representa el tamaño de la región crítica**. Formalmente,

$$\alpha = P(ETI) = P(RH_0/H_0V)$$

Se denomina **Error de Tipo II (ETII)** al hecho de **no rechazar la hipótesis nula cuando ésta es falsa**. Puede suceder que a pesar de que el valor numérico del Estadígrafo de Prueba no pertenezca a la Región Crítica, entonces no se rechaza y dicha hipótesis sea falsa. En esta situación estaríamos cometiendo un error en la decisión tomada. La probabilidad de cometer **este error se denomina β**. Formalmente,

$$\beta = P(ETII) = P(No\ RH_0/H_0F)$$

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

¿Puede dar un falso positivo una prueba de embarazo?

Obtener un falso positivo real (uno en el que no estaba embarazada en ningún momento) es increíblemente raro. Los **tests de embarazo** detectan la hormona hCG (gonadotropina coriónica humana) en su cuerpo, que solo está presente cuando está embarazada. Por eso un resultado positivo en el test de embarazo significará que está embarazada casi con total seguridad.

Sin embargo, en casos raros, puede obtener un test de embarazo falso positivo a raíz de lo siguiente:

- Un embarazo reciente (por ejemplo, después de un aborto natural, nacimiento reciente o aborto voluntario)
- Algunos quistes ováricos raros
- Algunos fármacos que contienen la hormona hCG, como algunos tratamientos de fertilidad

Sin embargo, hay casos en los que puede obtener un resultado positivo y descubrir más tarde que ya no está embarazada, pero sí lo estaba. Esto puede ocurrir si ha tenido un embarazo químico o aborto espontáneo, o un embarazo ectópico. Esto no son falsos positivos; son resultados de embarazo positivos reales aunque este no continúe.

Decision Matrix

		DECISION	
		Reject H_o	*Fail to Reject H_o*
ACTUAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α
	H_a True	**Correct Decision** **Power = 1- β**	Type II Error *Consumer Risk* β-Risk False Negative

H_o: Null Hypothesis H_a: Alternative Hypothesi

		Truth		
		Positive	**Negative**	
Test	**Positive**	True Positive	False Positive Type I α	Total Testing Positive
	Negative	False Negative Type II β	True Negative	Total Testing Negative
		Total Truly Positive	Total Truly Negative	Total

Potencia de la Prueba

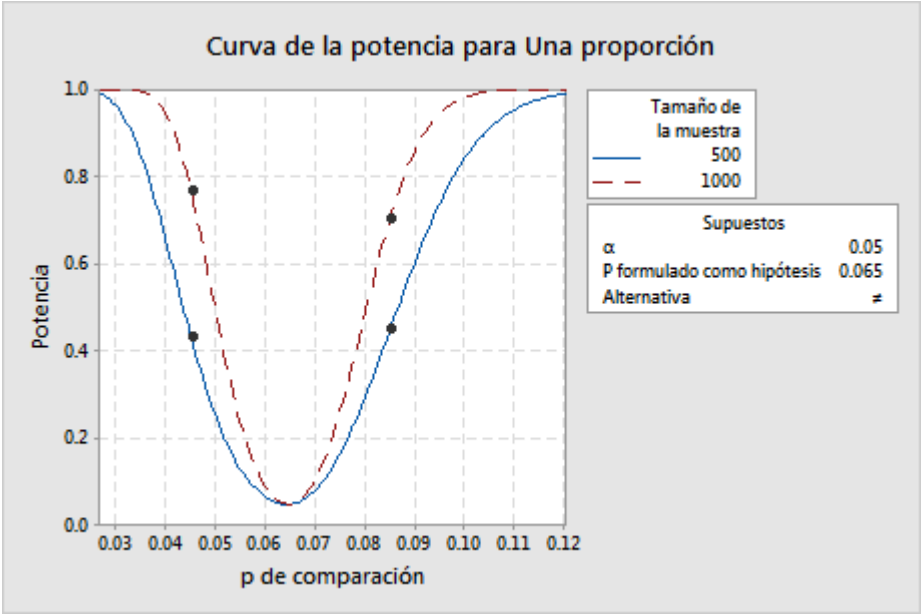
Se denomina **a la probabilidad de no cometer el Error de Tipo II, o sea, la probabilidad de rechazar la hipótesis nula cuando es falsa.** En términos formales,

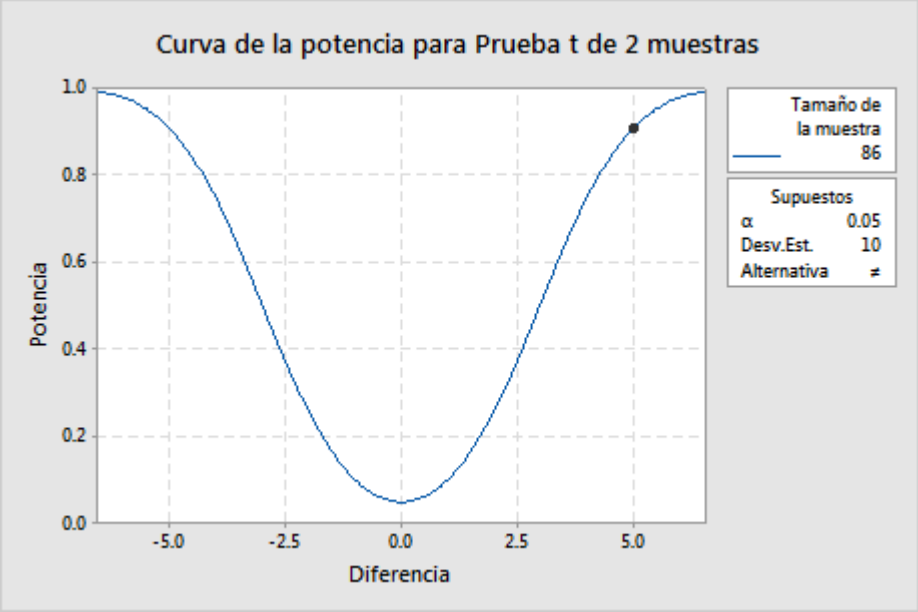
$$\Gamma = 1 - \beta = P(RH_0/H_0F)$$

Para plantear la potencia de la prueba, es necesario plantear una hipótesis alternativa única, o sea, establecer un solo valor alternativo para el parámetro.

Si se calculan las potencias correspondientes a todos los valores posibles del parámetro y se las representan gráficamente se obtiene la gráfica de una función denominada “Curva de Potencia”.

Curva de Potencia





Acción Derivada

El rechazo de la Hipótesis Nula inducirá al investigador a realizar una determinada acción con respecto al objeto de su investigación.

Si por el contrario, no se rechaza la hipótesis nula, entonces el investigador estará también inducido a realizar una acción, pero distinta. Cualesquiera de estas dos acciones se derivan del resultado de la prueba de hipótesis.

Se denomina Acción Derivada a la acción que se lleva a cabo según el resultado de la decisión estadística que se tome, rechazar o no la hipótesis nula.

Pasos a seguir

1. Plantear las Hipótesis
2. Indicar el Estadígrafo de Prueba a utilizar y su correspondiente Distribución de Probabilidad
3. Establecer la Región Crítica
4. Plantear la Regla de Decisión Estadística
5. Calcular el Valor Numérico del Estadígrafo de Prueba y verificar a que Región pertenece
6. Tomar la Decisión Estadística
7. Llevar a cabo la Acción Derivada

Condición	Enfermos	Sanos	Total
Test positivos	PV	FP	PV + FP
Test negativos	FN	NV	NV + FN
Total	PV + FN	NV + FP	PV + FP + NV + FN = n

PV = Positivos verdaderos, **FP** = Falsos positivos, **NV** = Negativos verdaderos, **FN** = Falsos negativos, **n** = total de individuos.

Positivos verdaderos: Individuos que habiendo dado positivos a la prueba diagnóstica, están verdaderamente enfermos.

Falsos positivos: Individuos que habiendo dado positivos a la prueba diagnóstica, están verdaderamente sanos.

Negativos verdaderos: Individuos que habiendo dado negativos a la prueba diagnóstica, están verdaderamente sanos.

Falsos negativos: Individuos que habiendo dado negativos a la prueba diagnóstica, están verdaderamente enfermos.

Tabla I **Tabla de contingencia 2 x 2 para la evaluación de una prueba diagnóstica**

		Patrón de referencia +	Patrón de referencia −	
Prueba diagnóstica	+	Verdaderos positivos (a)	Falsos positivos (b)	a+b
	−	Falsos negativos (c)	Verdaderos negativos (d)	c+d
		a+c	b+d	Total=a+b+c+d

Claves:

a Verdaderos positivos (VP): enfermos con la prueba positiva
b Falsos positivos (FP): no enfermos con la prueba positiva
c Falsos negativos (FN): enfermos con la prueba negativa
d Verdaderos negativos (VN): no enfermos con la prueba negativa
a+c Casos con patrón de referencia positivo (enfermos)
b+d Casos con patrón de referencia negativo (no enfermos)
a+b Casos con la prueba diagnóstica positiva
c+d Casos con la prueba diagnóstica negativa

La **sensibilidad** (Se) es la probabilidad de que la prueba dé positiva si la condición de estudio está presente (paciente enfermo o con patrón de referencia positivo). También se puede definir como la proporción de verdaderos positivos respecto al total de enfermos.

Sensibilidad: $$Se = \frac{VP}{Enfermos} = \frac{a}{a + c}$$

La **especificidad** (Es) es la probabilidad de que la prueba dé negativa si la enfermedad está ausente (paciente sano o con patrón de referencia negativo). También se puede definir como la proporción de verdaderos negativos respecto al total de sujetos sanos.

Especificidad: $$Es = \frac{VN}{Sanos} = \frac{d}{b + d}$$

6.1.1. Teorema de Bayes.

$$Pr\,(E/+) = \frac{Pr\,(E) \times Pr\,(+/E)}{Pr\,(E) \times Pr\,(+/E) + Pr\,(NE) \times Pr\,(+/NE)}$$

Donde:

Pr (E) = Probabilidad de Enfermedad.
Pr (S) = Probabilidad de Sano.
Pr (+/E) = Probabilidad de que estando Enfermo sea Positivo a la prueba diagnóstica.
Pr (+/S) = Probabilidad de que estando Sano sea Positivo a la prueba diagnóstica.

Reemplazando:

Pr (E/+) = VP+ = Valor predictivo positivo.
Pr (E) = PR = Prevalencia real de la enfermedad.
Pr (S) = 1 - PR = Prevalencia de sanos.
Pr (+/E) = Sens. = Sensibilidad de la prueba diagn\stica.
Pr (+/S) = 1 - Esp. = Proporci\n de positivos sanos.

El avance de la pandemia

Coronavirus en Argentina: la fórmula matemática que explica por qué no es lo mismo un test positivo que un infectado

Dos matemáticos del Conicet, Alicia Dickenstein y Pablo Groisman, explican que los testeos rápidos masivos no dan un "certificado de inmunidad" porque entre los testeos que dan positivo, una gran cantidad son falsos. Pero los datos epidemiológicos que ofrecen son muy valiosos.

Si tomamos -por poner un ejemplo- el test rápido que tiene una sensibilidad del 93,8% y una especificidad del 95,6%, esa cuenta da 6%. Es decir que si nos testearon y nos dio positivo, la probabilidad de que tengamos los anticuerpos **es de 6 en 100**.

Bajísima. A pesar de haber tenido un test positivo.

¿Eso quiere decir que el test es malo? No. El problema es que **la probabilidad de estar infectado es muy baja.** Pero cabe observar que saber que tenemos un test positivo aumenta muchísimo la probabilidad de estar infectados: pasó de 3 en 1.000 (antes de contar con la información que provee el test) a 3 en 50 (luego de tener un resultado positivo en el test). El test hizo bien su trabajo, al dar positivo, hizo que la probabilidad de tener los anticuerpos aumentara **20 veces.** El problema es que sigue siendo baja.

Artículo

By Sharon Theimer

Falsos negativos en prueba de COVID-19 pueden llevar a errónea sensación de seguridad

April 13, 2020

Hasta en las pruebas con valores de sensibilidad altos del 90 por ciento, la magnitud del riesgo de un resultado de falso positivo puede ser grande, a medida que aumenta la cantidad de gente sometida al análisis. “En California, se calcula que la tasa de la infección por COVID-19 exceda del 50 por ciento hacia mediados de mayo del 2020. Con una población de 40 millones de personas, se anticiparía que con las pruebas integrales haya 2 millones de falsos positivos en California. Aunque el análisis se hiciera solamente en el 1 por ciento de la población, se anticiparían 20 000 resultados falsos positivos”, añade la médica.

Los autores del trabajo también mencionan los efectos de esto sobre el personal de la salud. Si la tasa de la infección por COVID-19 de las más de 4 millones de personas que en Estados Unidos prestan cuidados médicos directos a los pacientes fuese del 10 por ciento (lo que está muy por debajo de las predicciones), se anticiparía obtener más de 40 000 falsos positivos al hacer la prueba a todos los proveedores de atención médica.

https://newsnetwork.mayoclinic.org/discussion/falsos-negativos-en-prueba-de-covid-19-pueden-llevar-a-erronea-sensacion-de-seguridad/

Machine Learning: Confusion Matrix

Confusión o error Matrix es una tabla que describe el rendimiento de un modelo supervisado de Machine Learning en los datos de prueba, donde se desconocen los verdaderos valores. Se llama “matriz de confusión” porque hace que sea fácil detectar dónde el sistema está confundiendo dos clases.

1. **True Positives (TP):** cuando la clase real del punto de datos era 1 (Verdadero) y la predicha es también 1 (Verdadero)
2. **Verdaderos Negativos (TN):** cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).
3. **False Positives (FP):** cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True).
4. **False Negatives (FN):** Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso).

Vamos a dar una etiqueta a nuestra variable de destino:

1 *: cuando una persona tiene cáncer*

0: *Cuando una persona NO está teniendo cáncer.*

	Actual - Cancer	Actual - NOT Cancer	Total
Predicted - Cancer	TP = 20	FP = 70	90
Predicted - NOT Cancer	FN = 10	TN = 200	210
Total	30	270	300

La matriz de confusión en en sí mismo no es una medida de desempeño como tal, pero casi todas las métricas de desempeño se basan en la matriz de confusión y los números dentro de ella.

Minimizando falsos negativos VS falsos positivos

En el ejemplo del problema de detección de cáncer digamos que de 100 personas, solo 5 personas tienen cáncer Definitivamente queremos capturar todos los casos de cáncer y podríamos terminar haciendo una clasificación cuando la persona que realmente NO tiene cáncer se clasifica como cancerosa.

Esto podría estar bien ya que perder a un paciente con cáncer será un gran error ya que no se realizarán más exámenes. Por lo tanto, es mejor **minimizar los falsos negativos** en este caso

Consideremos ahora un problema de detección de spam por correo electrónico y que está esperando un correo electrónico importante como una respuesta de un reclutador o esperando una carta de admisión de una universidad.

Asigne una etiqueta a la variable de destino y diga:

1:"El correo electrónico es un correo no deseado" y

0:"El correo electrónico no es un correo no deseado"

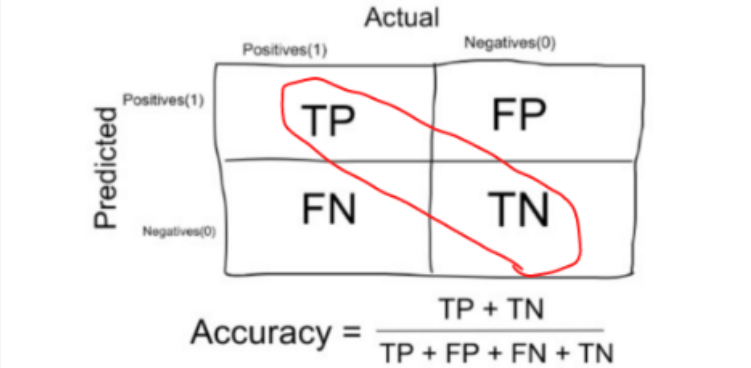
Supongamos que el Modelo clasifica ese correo electrónico importante que usted están esperando desesperadamente, como Spam (caso de falso positivo).

Por lo tanto, en el caso de la clasificación de correos electrónicos no deseados, **minimizar los falsos positivos** es más importante que los falsos negativos.

https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

Machine Learning: Accuracy (Precisión)

Es el porcentaje total de elementos clasificados correctamente.



Por lo tanto, para nuestro ejemplo: Precisión = (20 + 200) / (20 + 10 + 70 + 200) = 220/300.

Es la medida más directa de la calidad de los clasificadores. Es un valor entre 0 y 1. Cuanto más alto, mejor.

Esta métrica es tan intuitiva y natural, de hecho, que las personas a menudo la utilizan sin pensarlo dos veces, aunque ciertamente no es apropiado en muchos casos.

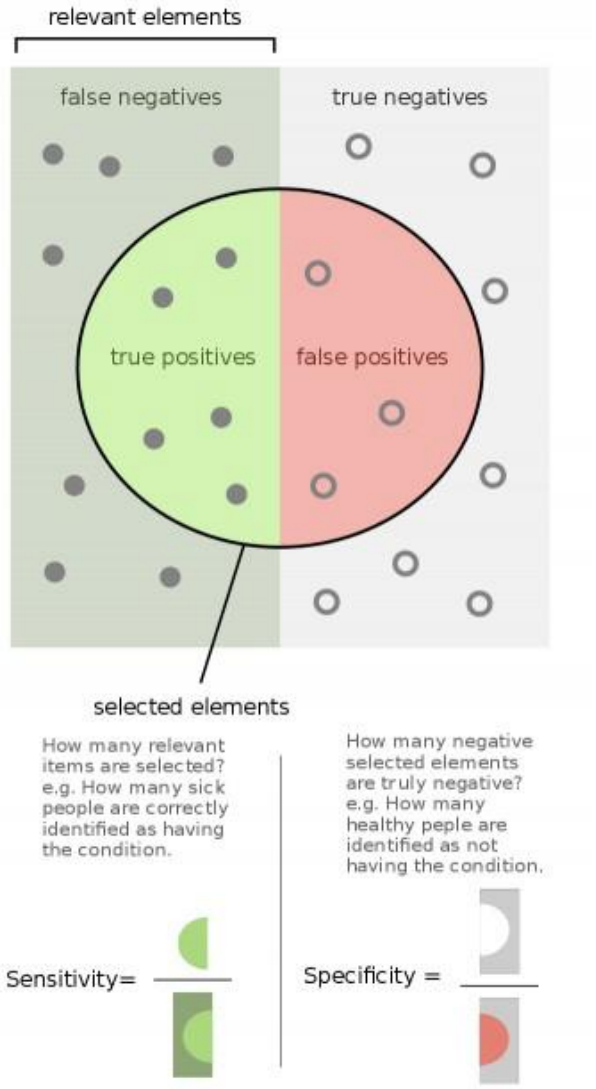
1. Tasa de falso positivo o Error tipo I: *Número de elementos identificados erróneamente como positivos de total negativos verdaderos- FP / (FP + TN)*

2. Tasa falso negativo o error de tipo II: *Número de elementos identificados erróneamente como s negativo del total de verdaderos positivos: FN / (FN + TP)*

Ejemplo: En el ejemplo de detección de cáncer, consideremos que contiene 100 personas, solo 5 personas tienen cáncer. Predicemos que todos los pacientes tienen cáncer:

Entonces, TP=5, FP=95 and TN=FN=0

Precision = 5/(5+95)=5%
Recall = 5/(5+0)=100%
Specificity=0/(0+5) =0%


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.
Sensitivity=
How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.
Specificity =

P-Value (P-Valor para lxs amigxs)

La elección del nivel de significación, es en cierta forma arbitraria en base al criterio analista. Se corresponde al nivel de significación más pequeño posible que puede escogerse, para el cual todavía se rechazaría la hipótesis nula con las observaciones actuales. Cualquier nivel de significación escogido inferior al p-valor indica el rechazo de la hipótesis nula.

El p-valor es una medida directa de lo verosímil que resulta obtener una muestra como la actual si es cierta H_0. Los valores pequeños indican que es muy infrecuente obtener una muestra como la actual, en cambio, los valores altos que es frecuente. El p-valor se emplea para indicar cuánto (o cuán poco) contradice la muestra actual la hipótesis alternativa. Informar sobre cual es el p-valor tiene la ventaja de permitir que cualquiera decida qué hipótesis no rechaza basándose en su propio nivel de riesgo α.

$$Si\ P-Valor > \alpha \Rightarrow No\ Rechazo\ H_0$$

$$Si\ P-Valor \leq \alpha \Rightarrow Rechazo\ H_0$$

http://www.ub.edu/stat/GrupsInnovacio/Statmedia/demo/Temas/Capitulo9/B0C9m1t18.htm

P-Value (o P-Valor)

$$Si\ P - Valor > \alpha \Rightarrow No\ Rechazo\ H_0$$

$$Si\ P - Valor \leq \alpha \Rightarrow Rechazo\ H_0$$

H_0

H_a

When the p value is low

Estadisticos de Transformación

Para una Población

POBLACIÓN NORMAL	Población Infinita	Población Finita
Media Poblacional μ σ^2 *conocida*	$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$	$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$
Media Poblacional μ σ^2 *desconocida*	$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$	$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \sim t_{n-1}$
Varianza Poblacional σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	
Proporción Poblacional p	$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \longrightarrow N(0,1)$	$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \longrightarrow N(0,1)$

Resolución de Ejercicio de la Guía

Ejercicio n° 1

Se pudo comprobar que el año pasado, los precios de una determinada canasta de productos se distribuye normalmente con media $1780, y un desvío estándar de $110. Este año, una muestra de 40 ventas, proporcionó un precio promedio de $1900. Con un nivel de significación del 5%, ¿se puede afirmar que el precio promedio de estos productos, ha aumentado, con respecto al precio promedio del año pasado?

$X = Precios\ de\ una\ determinada\ canasta\ de\ productos\ (En\ \$)$

$X \sim N(1780,110)$

$n = 40 \Rightarrow \bar{X} = 1900$

$\alpha = 0.05$

Planteo de Hipótesis

$$H_0: \mu \leq 1780$$

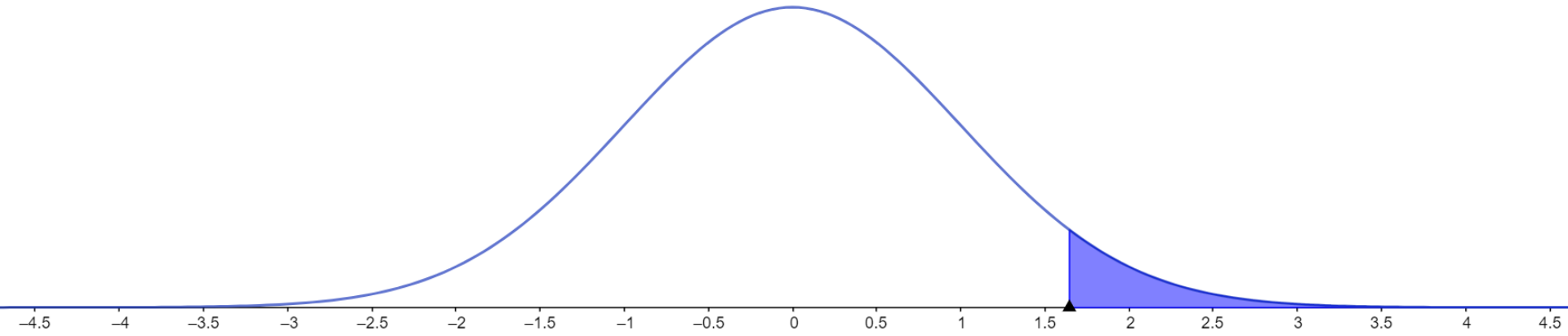
$$H_1: \mu > 1780$$

Determinación del Estadígrafo de Prueba

$$Z^e = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Determinación de la Región Crítica

El signo positivo de la hipótesis alternativa me provee evidencia que la región crítica se ubica a derecha. $H_1: \mu > 1780$



$$R_c: Z \geq 1{,}645$$

Regla de Decisión

$$Si\ Z^e \in R_c \mid Z^e \geq 1{,}645 \Rightarrow RH_0\ (Rechazo\ la\ Hipótesis\ Nula)$$

$$Si\ Z^e \notin R_c \mid Z^e < 1{,}645 \Rightarrow No\ RH_0\ (No\ Rechazo\ la\ Hipótesis\ Nula)$$

Cálculo del Valor Empírico

$$Z^e = \frac{1900 - 1780}{\frac{110}{\sqrt{40}}} = 6{,}899$$

Decisión Estadística

$$Z^e \geq 1{,}645 \Rightarrow RH_0\ (Rechazo\ la\ Hipótesis\ Nula)$$

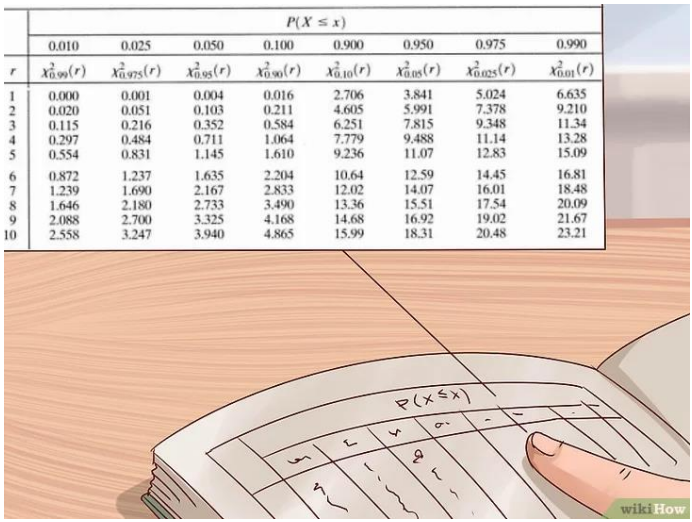
Acción Derivada

Bajo la evidencia empírica se puede afirmar que los precios de la canasta básica han aumentado respecto al año anterior.

Prueba de Hipótesis

Para Dos Poblaciones

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21



Fundamentación de testear dos Poblaciones

En la clase anterior interesaba poner a prueba alguna conjetura acerca de un parámetro poblacional determinada.

En esta parte interesa poner a prueba la conjetura sobre parámetros que involucran información sobre dos poblaciones específicas. Alguno de los parámetros que se pondrán a prueba son,

- Diferencia de Medias Poblacionales: $\mu_1 - \mu_2$
- Cociente de Varianzas Poblacionales: $\frac{\sigma_2^2}{\sigma_1^2}$
- Diferencia de Proporciones Poblacionales: $p_1 - p_2$

El supuesto fuerte que **se asume** es que las **Poblaciones son Normales e Infinitas**.

Diferencia de Medias Poblacionales: $\mu_1 - \mu_2$

El parámetro es probar es la diferencia de las medias poblacionales de cada población.

El estimador de dicho parámetro poblacional es: $\bar{X}_1 - \bar{X}_2$

Se demuestra que bajo las condiciones descriptas

$$\bar{X}_1 \sim N\left(\mu_1; \sqrt{\frac{\sigma_1^2}{n_1}}\right) \quad \bar{X}_2 \sim N\left(\mu_2; \sqrt{\frac{\sigma_2^2}{n_2}}\right)$$

Por lo tanto, bajo estas condiciones el estimador del parámetro poblacional se distribuye con los siguientes parámetros,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Diferencia de Medias Poblacionales: $\mu_1 - \mu_2$

El estadígrafo de transformación utilizado para poner a prueba dicho parámetro será,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Las posibilidades de hipótesis nula son,

- $H_0: \mu_1 - \mu_2 = k$
- $H_0: \mu_1 - \mu_2 \leq k$
- $H_0: \mu_1 - \mu_2 \geq k$

Diferencia de Medias Poblacionales: $\mu_1 - \mu_2$

Ahora bien si las varianzas poblacionales son desconocidas $\sigma_1^2 =? \; y \; \sigma_2^2 =?$ y la población es normal, entonces deberá utilizarse la Distribución T-Student tal como se hace con una Población.

El problema es las varianzas poblacionales desconocidas pueden ser:

1. Iguales $\sigma_1^2 = \sigma_2^2$
2. Distintas $\sigma_1^2 \neq \sigma_2^2$

Estas conjeturas deben ser validadas mediante una prueba de hipótesis que testee la condición de que las varianzas poblacionales desconocidas son iguales (hipótesis nula) contra una hipótesis de que son distintas (alternativa). Este test se denomina “Prueba de Homocedasticidad”. En términos formales,

$$H_0: \frac{\sigma_2^2}{\sigma_1^2} = 1 \qquad H_1: \frac{\sigma_2^2}{\sigma_1^2} \neq 1$$

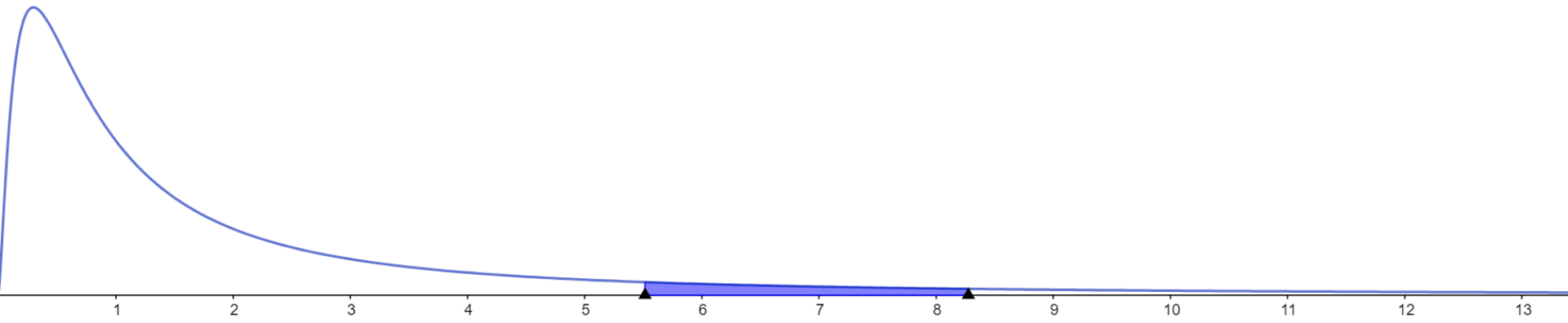
Cociente de Varianzas Poblacionales: $\frac{\sigma_2^2}{\sigma_1^2}$

$$H_0: \frac{\sigma_2^2}{\sigma_1^2} = 1 \qquad H_1: \frac{\sigma_2^2}{\sigma_1^2} \neq 1$$

El estadístico de esta prueba es,

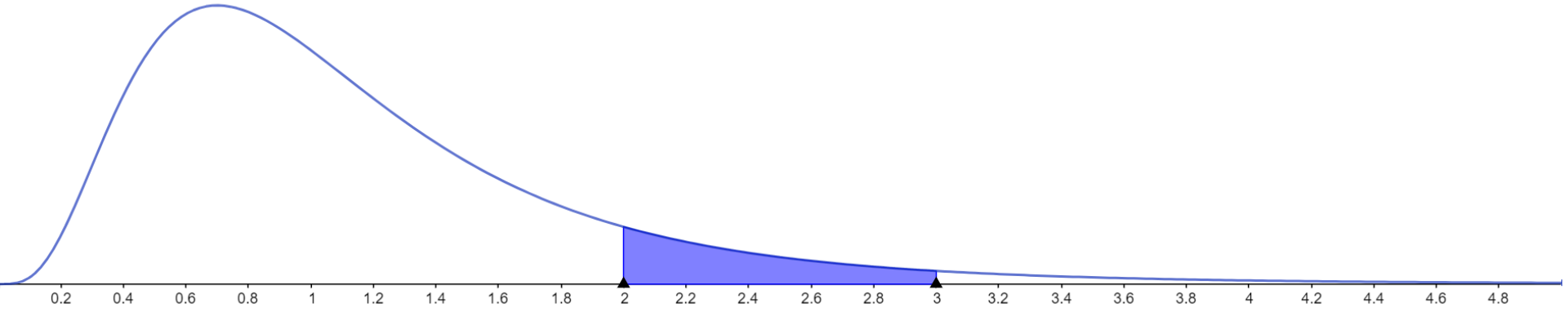
$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;n_2-1}$$

Es una Distribución F de Snedecor con dos parámetros, grados de libertad en el numerador y en el denominador. Surge como el cociente de dos variables Chi Cuadrado, cada una con sus correspondientes grados de libertad. La forma funcional de la distribución se asemeja a la Chi Cuadrada.



F-Distribution df1 5 df2 2

P(5.52 ≤ X ≤ 8.28) = 0.0492



F-Distribution df1 10 df2 14

P(2 ≤ X ≤ 3) = 0.0845

Demostración del Estadístico F de Snedecor

La F de Snedecor se determina de la forma siguiente,

$$F = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_n^2}{m}} \sim F_{n;m}$$

Para una población se había obtenido el siguiente estadístico

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Si se reemplaza por este estadístico para cada población se obtiene lo siguiente,

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}}{\frac{(n_2-1)S_2^2}{\sigma_2^2}} = \frac{\frac{(n_1-1)}{(n_2-1)}}{\frac{(n_2-1)}{(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;n_2-1}$$

Cociente de Varianzas Poblacionales: $\frac{\sigma_2^2}{\sigma_1^2}$

$$H_0: \frac{\sigma_2^2}{\sigma_1^2} = 1 \quad "Homocedasticidad"$$

$$H_1: \frac{\sigma_2^2}{\sigma_1^2} \neq 1 \quad "Heterocedasticidad"$$

Del resultado de este test se determina si las varianzas poblacionales son iguales o no bajo la evidencia proporcionada por la muestra.

Si **se rechaza la hipótesis nula**, entonces bajo la evidencia empírica se afirma que las varianzas poblacionales desconocidas son distintas (se utilizará para testear la diferencia de medias poblacionales el **test de Welch**).

Si **NO se rechaza la hipótesis nula**, entonces bajo la evidencia empírica se afirma que las varianzas poblacionales desconocidas son iguales (se utilizará para testear la diferencia de medias poblacionales el test que emplea **una varianza amalgamada o ponderada**).

Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ con Varianzas Poblacionales Desconocidas Iguales $\sigma_1^2 = \sigma_2^2$

Del test de Homocedasticidad se decidió bajo la evidencia empírica no rechaza la hipótesis nula.

$$H_0: \frac{\sigma_2^2}{\sigma_1^2} = 1 \qquad H_1: \frac{\sigma_2^2}{\sigma_1^2} \neq 1$$

Por lo tanto, se asume que las varianzas poblacionales desconocidas son iguales. Dado que la población es normal, entonces se utilizará la T de Student.

Recordemos:

Sea X una Población que se distribuye normalmente con Esperanza Matemática μ y Varianza Poblacional (desconocida) σ^2 respectivamente. Bajo estas condiciones se demuestra que la distribución de probabilidad utilizada es la T-Student,

$$\boldsymbol{T = \frac{N(0,1)}{\sqrt{\frac{\chi_n^2}{n}}} \sim t_n}$$

Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ con Varianzas Poblacionales Desconocidas Iguales $\sigma_1^2 = \sigma_2^2$

$$T = \frac{N(0,1)}{\sqrt{\frac{\chi_n^2}{n}}} \sim t_n$$

Se dispone de la siguiente variable normal estándar,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

El cociente de la primera fórmula se construye para dos poblaciones de la forma siguiente,

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}$$

Se asume que ambos estadísticos son independientes.

$$T = \frac{\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}}{\sqrt{\dfrac{\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2} + \dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}$$

Al reordenar los términos y realizar una simplificación y asumir varianzas iguales se llega a la siguiente fórmula,

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_a^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

$$S_a^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ con Varianzas Poblacionales Desconocidas Distintas $\sigma_1^2 \neq \sigma_2^2$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v$$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Encontrarán la fórmula de v de la forma siguiente $v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

Diferencia de Proporciones Poblacionales $p_1 - p_2$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \longrightarrow N(0,1)$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$$

$$\hat{q} = 1 - \hat{p}$$

POBLACIÓN NORMAL	Estadígrafos de Prueba Utilizados	Observaciones
Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ σ_1^2 *y* σ_2^2 *conocidas*	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	
Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ σ_1^2 *y* σ_2^2 *desconocidas e iguales*	$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$	$S_a^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$
Diferencia de Medias Poblacionales $\mu_1 - \mu_2$ σ_1^2 *y* σ_2^2 *desconocidas y distintas* **(Test de Welch)**	$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v$	$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$
Cociente de Varianzas Poblacionales $\frac{\sigma_2^2}{\sigma_1^2}$	$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;n_2-1}$	
Diferencia de Proporciones Poblacionales $p_1 - p_2$	$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \longrightarrow N(0,1)$	$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$ $\hat{q} = 1 - \hat{p}$

Intervalos de Confianza para dos Poblaciones

IC para estimar $\mu_1 - \mu_2$, cuando σ_1^2 y σ_2^2 son Conocidas y la Población es Normal e Infinita

$$P\left(\bar{X}_1 - \bar{X}_2 - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Intervalos de Confianza para dos Poblaciones

IC para estimar $\mu_1 - \mu_2$, cuando σ_1^2 y σ_2^2 son Desconocidas y Distintas

$$P\left(\bar{X}_1 - \bar{X}_2 - t_{1-\frac{\alpha}{2};v}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{1-\frac{\alpha}{2};v}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right) = 1-\alpha$$

$$v = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Intervalos de Confianza para dos Poblaciones

IC para estimar $\mu_1 - \mu_2$, cuando σ_1^2 y σ_2^2 son Desconocidas e Iguales

$$P\left(\bar{X}_1 - \bar{X}_2 - t_{1-\frac{\alpha}{2};v}\sqrt{S_a^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{1-\frac{\alpha}{2};v}\sqrt{S_a^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = 1-\alpha$$

$$v = n_1 + n_2 - 2$$

$$S_a^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Intervalos de Confianza para dos Poblaciones

IC para estimar $\frac{\sigma_2^2}{\sigma_1^2}$

$$P\left(\frac{S_2^2}{S_1^2}\frac{1}{B} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{S_2^2}{S_1^2}\frac{1}{A}\right) = 1-\alpha$$

$$B = F_{1-\frac{\alpha}{2};n_1-1;n_2-1} \qquad A = F_{\frac{\alpha}{2};n_1-1;n_2-1}$$

IC para estimar $p_1 - p_2$

$$P\left(\bar{p}_1 - \bar{p}_2 - Z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq p_1 - p_2 \leq \bar{p}_1 - \bar{p}_2 + Z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = 1-\alpha$$

$$\hat{p} = \frac{X_1 + X_1}{n_1 + n_2} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \qquad \hat{q} = 1 - \hat{p}$$

Resolución de Ejercicio de la Guía

Ejercicio n° 11

En un negocio de ropa se desea comparar la eficiencia (medida en el monto medio de sus ventas) de sus dos vendedores, Hernán y María Julia. Se sabe que el monto de las ventas para ambos vendedores tiene distribución normal. El desvío estándar de la distribución de Hernán es igual a \$75 y el desvío estándar de la distribución de María Julia es \$50. Una muestra de 36 ventas de Hernán proporcionó un monto promedio diario de \$300, mientras que una muestra de 40 ventas de María Julia, proporcionó un monto promedio de \$350. Con un nivel de significación del 5%, verificar si María Julia es más eficiente que Hernán.

Respuesta: $R_c : Z^t \geq 1{,}645 \;\; Z^e = 3{,}38$. Se rechaza H_0

$$X = Monto\ de\ las\ Ventas\ para\ ambos\ vendedores\ (En\ \$)$$

$$X_1 \sim N(?, 75)\ (Hernán)\ X_2 \sim N(?, 50)(Ma.Julia) \qquad \alpha = 0.05$$

$$n_1 = 36 \Rightarrow \bar{X}_1 = 300$$

$$n_2 = 40 \Rightarrow \bar{X}_2 = 350$$

Planteo de Hipótesis

$$H_0: \mu_2 - \mu_1 \leq 0$$

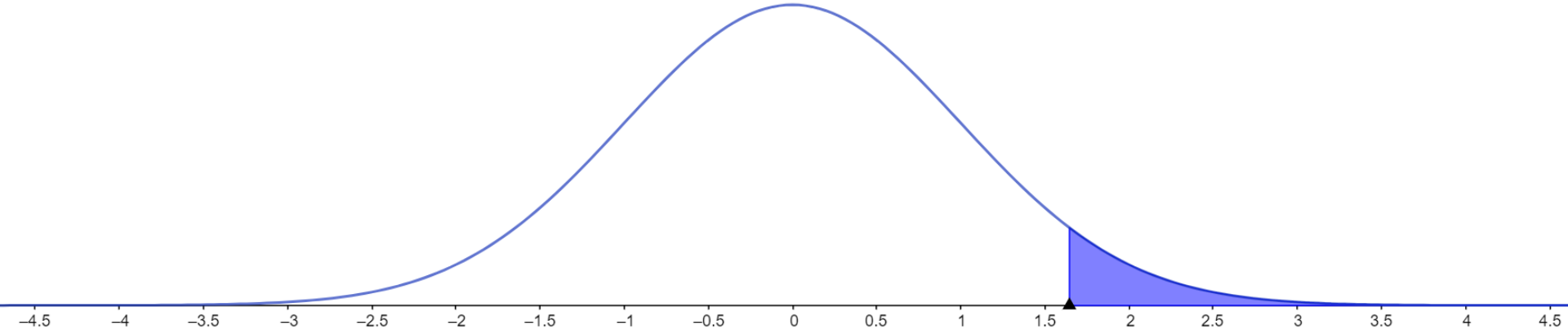
$$H_1: \mu_2 - \mu_1 > 0$$

Determinación del Estadígrafo de Prueba

$$Z^e = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} \sim N(0,1)$$

Determinación de la Región Crítica

El signo positivo de la hipótesis alternativa me provee evidencia que la región crítica se



$$R_c: Z \geq 1{,}645$$

Regla de Decisión

$$Si\ Z^e \in R_c \mid Z^e \geq 1{,}645 \Rightarrow RH_0\ (Rechazo\ la\ Hipótesis\ Nula)$$

$$Si\ Z^e \notin R_c \mid Z^e < 1{,}645 \Rightarrow No\ RH_0\ (No\ Rechazo\ la\ Hipótesis\ Nula)$$

Cálculo del Valor Empírico

$$Z^e = \frac{(350 - 300) - 0}{\sqrt{\frac{50^2}{40} + \frac{75^2}{36}}} = 3{,}38$$

Decisión Estadística

$$Z^e \geq 1{,}645 \Rightarrow RH_0\ (Rechazo\ la\ Hipótesis\ Nula)$$

Acción Derivada

Bajo la evidencia empírica se puede afirmar que María Julia es más eficiente que Hernán.

Próxima Clase

En la próxima semana veremos

- **Regresión Lineal Simple**

Por tal motivo es necesario e indispensable que tengan en claro cada uno de los temas, antes de avanzar con la segunda parte de la materia.

Recomendación

Leer al menos alguno de los siguientes capítulos

Capítulo 13 del libro de Canavos

Capítulo 10 del libro de Chao

Capítulo 14 del libro de Anderson

Consultar dudas en el foro del Campus Virtual


TO BE
CONTINUED...

Preguntas, Sugerencias y Comentarios

RDelRosso-ext@austral.edu.ar

¡Muchas Gracias!