SESIÓN 4 PROGRAMACIÓN EN R

Inferencia estadística

Hugo Andrés Dorado B.

Contenido

- Definiciones de inferencia.
- Bases de probabilidad.
- Teorema central de límite.
- Estimaciones puntuales o por intervalos.
- Pruebas de hipótesis y valor p.
- Pruebas no paramétricas.

Inferencia estadística

- Es el conjunto de métodos y técnicas que permiten sacar conclusiones de una **población** a partir de información de una **muestra**.

Homogeneidad
Tiempo
Espacio
Cantidad

Aleatorio
Estratificado
Sistemático

Inferencia estadística

La inferencia estadística navega sobre un conjunto de supuestos y herramientas sobre como extraer información sobre los datos.

- Estrategia de muestreo.
- Control de variables de confusión.
- Sesgo en el diseño del estudio.
- Distribución de probabilidad en los datos.

Probabilidad

Una medida de probabilidad, P, es función de una colección de posible eventos que siguen las siguientes condiciones:

1. Para un evento $E \subset \Omega, 0 \leq P(E) \leq 1$
2. $P(\Omega) = 1$
3. Sí E_1 y E_2 son mutuamente excluyentes $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Variable aleatoria

Es una **función** que asigna un valor usualmente **numérico**, al resultado de un **experimento aleatorio**.

- Pueden ser discretas o continuas
- Variables aleatorias discretas si solo pueden tomar un valor contable en un número de posibilidades.
- Variables continuas si pueden tomar cualquier valor en una línea real de tiempo.

Ejemplos de variable aleatoria.

- Lanzamiento de una moneda o un dado.
- El tiempo de vida de un celular después de comprado.
- El número de llegadas de clientes a una tienda.
- El peso de una persona que es elegida aleatoriamente en el salón de clase.

Función de masa o función de densidad

Una función de masa o densidad corresponde a una función que calcula la probabilidad de que una variable aleatoria tome cierto valor satisface que:

Discretas

1. $p(X = x) \geq 0$ para todo x
2. $\sum_i^{\infty} p(X = i) = 1$

Continuas

1. $\mathrm{f}(X = x) \geq 0$ para todo x
2. $\int_{-\infty}^{\infty} f(X = x) = 1$

Función de distribución $\mathrm{F(X)} = P(x \leq x) = 1 - P(X > x)$

Distribuciones discretas.

Bernoulli: Consiste en realizar un experimento aleatorio una sola vez y observar si cierto suceso ocurre o no.

Proviene de una salida binaria [0,1], con respectivas probabilidades p, 1-p

$$P(X = x) = p^{\,x}(1-p)^{1-x}$$

$$E[x] = p$$

$$VAR[x] = p(1-P)$$

Ejemplos:
Se evalua si una persona tiene cierta enfermedad.
Se lanza una moneda para ver si sale cara.

Distribuciones discretas.

Binomial: Describe el número de éxitos obtenidos después de n ensayos.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\mathbb{E}[X] = np$$
$$\mathbb{V}\mathrm{ar}[X] = np(1-p)$$

Supongamos que se lanza un dado (con 6 caras) 51 veces y queremos conocer la probabilidad de que el número 3 salga 20 veces. En este caso tenemos una X ~ B(51, 1/6) y la probabilidad sería P(X=20):

$$P(X = 20) = \binom{51}{20} (1/6)^{20} (1-1/6)^{51-20}$$

Distribuciones discretas.

Poisson: Se expresa partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo.

Concretamente, se especializa en la probabilidad de ocurrencia de sucesos con probabilidades muy pequeñas, o sucesos "raros".

$$f(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

$$E[X] = \lambda$$
$$\mathrm{VAR}[X] = \lambda$$

Distribuciones discretas

Ejemplo poisson: En una clínica el promedio de atención es 16 pacientes por 4 horas, encuentre la probabilidad que en 30 minutos se atiendan menos de 3 personas.

Fórmula de Poisson:

$$P = (x = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} = \frac{e^{-\lambda}(\lambda)^k}{k!}$$

La probabilidad que en 30 minutos se atiendan menos de 3 personas.

$$\lambda = 16\frac{pacientes}{4\ horas} \quad \lambda = 4\frac{pacientes}{1\ hora} \quad \lambda = 2\frac{pacientes}{media\ hora}$$

$$P = (x < 3) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$P = (x < 3) = \frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^2}{2!}$$

$$P = (x < 3) = 0.1353 + 0.2707 + 0.2707$$

Resultado final:

$$P = (x < 3) = 0.6767$$

Distribuciones continuas

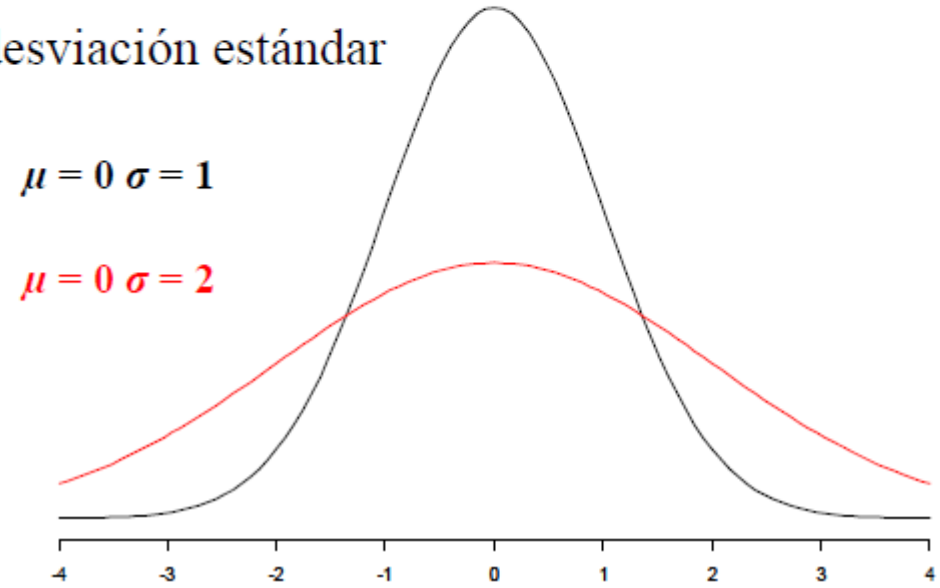
Distribución normal:

Definición

Se dice que una variable aleatoria X tiene una distribución Normal, si su función de densidad es:

$$f(x) = N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

Donde μ es la media y σ la desviación estándar



Distribuciones continuas

La distribución de probabilidad normal es una distribución continua de probabilidad.

Propiedades de la Normal:

1. La familia completa de distribución de probabilidad normales se diferencia por su media μ y desviación estándar σ.

2. El punto más alto de la curva normal es la media, que también es la mediana y la moda de la distribución.

3. La media de la distribución puede ser cualquier valor numérico: negativo, cero o positivo.

4. La distribución de probabilidad normal es simétrica, y su forma a la izquierda de la media es una imagen especular de la forma a la derecha de la media.

Distribuciones continuas

Propiedades de la Normal:

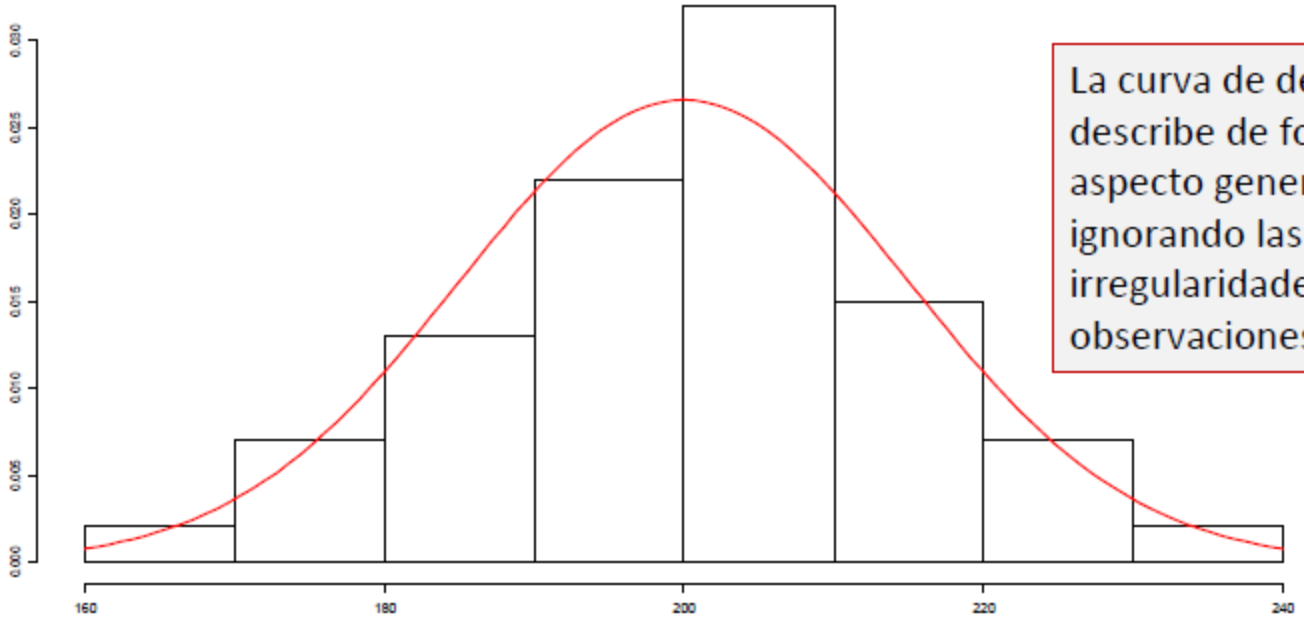
5. La desviación estándar (σ) determina el ancho de la curva. A valores mayores de σ se tienen curvas mas anchas y bajas, que muestran una mayor dispersión en los datos.

6. Las probabilidades para la variable aleatoria normal están dadas por áreas bajo la curva. El área total bajo la curva para la distribución de probabilidad normal es 1.

Distribuciones continuas

Ejemplo:

Se desea saber como se distribuye la cantidad de llenado de una maquina despachadora de gaseosas. Para esto se toma una muestra de 200 botellas y se observa la cantidad de liquido contenido



La curva de densidad normal describe de forma compacta el aspecto general de los datos, ignorando las pequeñas irregularidades así como las observaciones atípicas

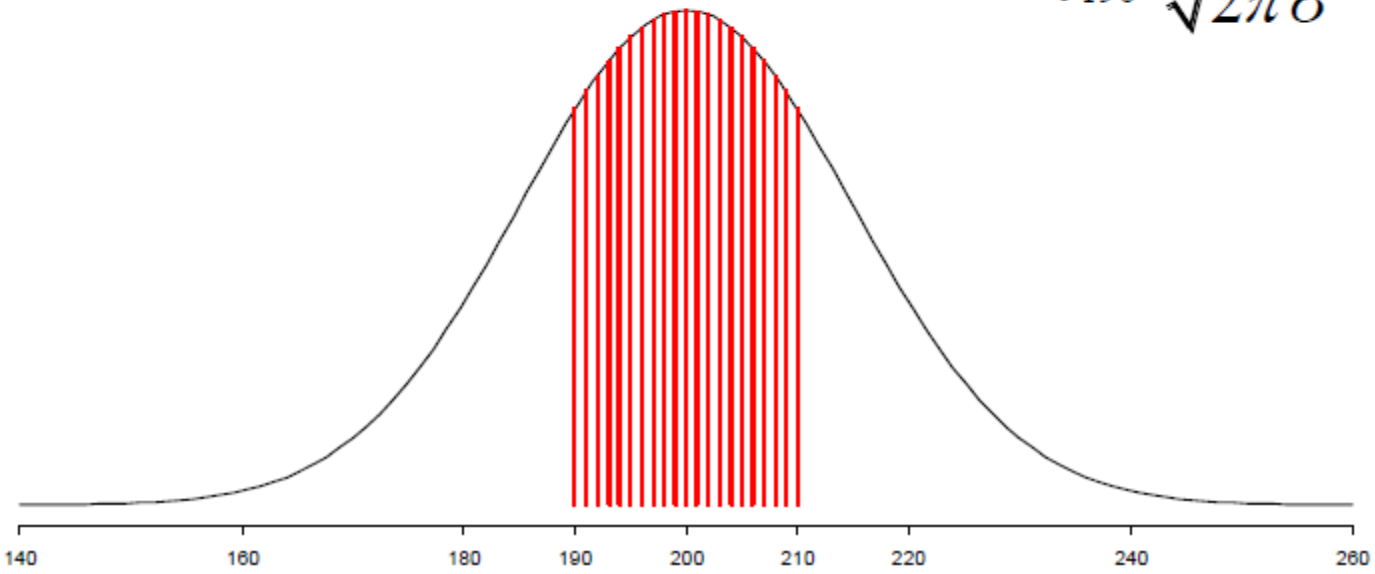
La cantidad de llenado de la maquina despachadora puede describirse por una distribución normal con media $200ml$ y desviación estándar $15ml$.

Distribuciones continuas

Ejemplo:

Una maquina despachadora de gaseosa está ajustada para servir un promedio de 200 ml por vaso. Si la cantidad de gaseosa es normalmente distribuida con una desviación estándar de 15 ml ¿Cuál es la probabilidad de que un vaso contenga entre 190 y 210 ml?

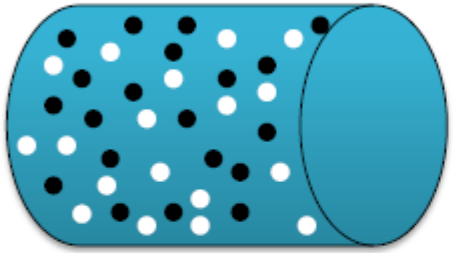
$$P(190 < X < 210) = \int_{190}^{210} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$



Definición de terminos

Parámetro

Característica medible sobre la población.

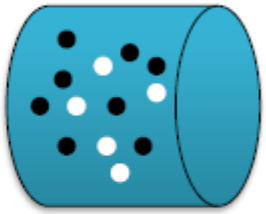


Ejemplo:

- Edad promedio de los estudiantes de ingeniería.
- Diámetro promedio de los tornillos fabricados en una empresa.

Estimador

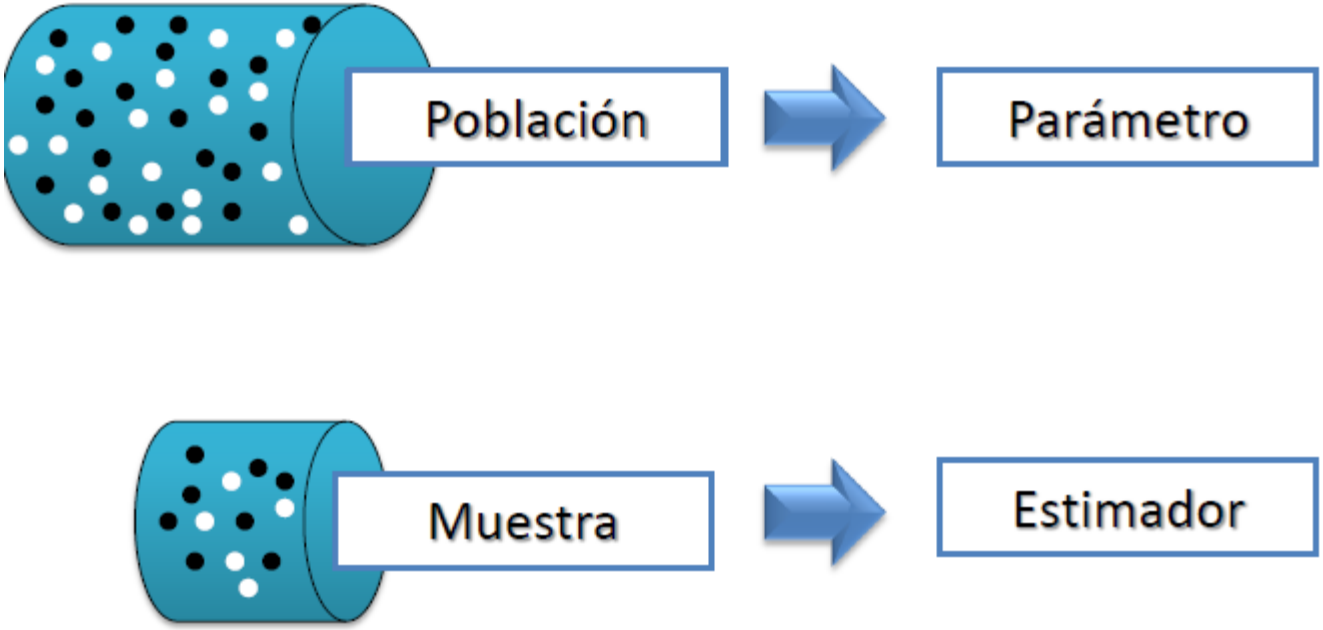
Característica medible sobre la muestra.



Ejemplo:

- Edad promedio de una muestra de los estudiantes de ingeniería.
- Diámetro promedio de una muestra de tornillos fabricados en una empresa.

Definición de terminos



Definición de términos

Característica	Muestra (Estadístico)	Población (Parámetro)
Variable Cuantitativa		
Media	$\overline{x}$	μ
Desviación típica	s	σ
Varianza	s^2	σ^2
Variable Categórica		
Porcentaje	$\widehat{P}$	P

Teorema central de limite (TCL)

- El teorema central de limite es uno de los más importantes en estadística.
- Sugiere que la distribución de promedios de variables aleatorias, se aproximan a una distribución normal a medida que el tamaño de la muestra incrementa.
- El CLT se aplica en una infinita variedad de configuraciones.
- Sí $X_1, \dots, X_n$ es una colección de v.a. iid variables aleatoria con media μ y varianza σ^2.
- Dado $\bar{X}_n$ es su media muestral.
- Entonces $\bar{X}_n$ tiene una distribución normal, a medida que n se vuelve grande.

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}} \sim N(0,1)$$

$$n \geq 30??$$

Distribuciones muestrales

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\widehat{P} \sim N(P, \sqrt{\frac{P \cdot (100 - P)}{n}})$$

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

$$\boxed{F = \frac{\widetilde{s}_1^{\,2}}{\widetilde{s}_2^{\,2}}}$$

Estimadores

Estimador puntual: Estadístico que estima el valor de un parámetro.

Intervalo de Confianza: Forma de estimar un parámetro en la cual se calcula un intervalo que indica con cierta seguridad un rango donde puede estar el parámetro.

Un estudio pretende estimar el porcentaje de hipertensos que hay entre las personas mayores de 65 años en la Comunidad Valenciana.

Estimador puntual: (167/350) x100=47.71%
Intervalo de confianza: **[42.48, 52.94]**

Intervalo de confianza

$$(L_1, L_2)_{(1-\alpha)\%}$$

La construcción depende de:

- Un nivel de confianza definido.
- Una distribution de referencia.
- El tamaño de la muestra.
- La variabilidad.

Intervalo de confianza para la varianza

Varianza

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right]$$

Media

$$\bar{X} \pm t_{n-1,1-\alpha/2}\, S/\sqrt{n}$$

Comparación de medias

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2}\, S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

- Varianzas desiguales.
- Pocos datos.
- Muchos datos
- Requiere que ambas distribuciones sean normales.
- Datos pareados.

Pruebas de hipótesis

Hipótesis: Se refiere a una prueba formal para tomar decisiones usando datos.

- **Hipótesis nula Ho**. Afirmación acerca del valor de un parámetro poblacional que se considera válida para desarrollar el procedimiento de prueba.

- **Hipótesis Alternativa Ha.-** Afirmación que se aceptará si los datos muéstrales proporcionan evidencia de que la hipótesis nula es falsa

$$H_0: \mu = 20$$
$$H_1: \mu \neq 20$$

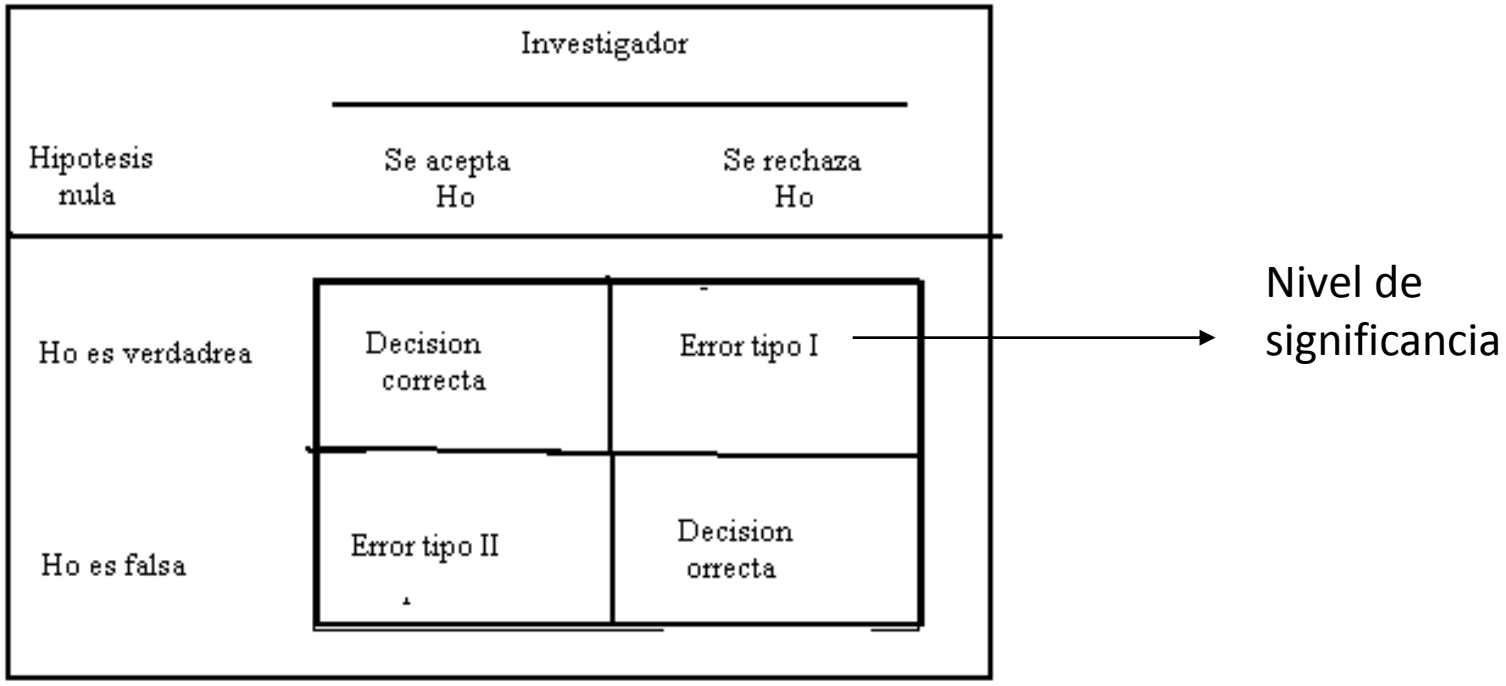
$$H_0: \sigma = 8$$
$$H_1: \sigma \neq 8$$

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

$$H_0: \mu = 20$$
$$H_1: \mu > 20$$

Pruebas de hipótesis

- La hipótesis alternative normalmente se representa de la forma <, > ó ≠.



Pruebas de hipótesis

Una empresa está interesada en lanzar un nuevo producto al mercado. Tras realizar una campaña publicitaria, se toma la muestra de 1 000 habitantes, de los cuales, 25 no conocían el producto. A un nivel de significación del 1% ¿apoya el estudio las siguientes hipótesis?

$$H_0: P = 0.03$$
$$H_1: P > 0.03$$

Más del 0.03 de la población no conoce el nuevo producto.

Valor P

- Es una medida común de test de significancia.
- Se ha vuelto un tema controversial en investigación según la interpretación.

http://warnercnr.colostate.edu/~anderson/thompson1.html

- Definición: Se refiere a la probabilidad de obtener datos como los obtenidos en la muestra, asumiendo que la hipótesis nula es cierta.

Enfoque:

Definir una distribución de probabilidad hipotética.

Calcular el estimador.

Un nivel de significancia o confianza.

Hablar sobre supuestos.

Prueba de comparación de varias poblaciones

Análisis de varianza

$H_0: \mu_1 = \mu_2 = \cdots = \mu_p$ La media de las muestras son iguales.

$H_A: \mu_1 \neq \mu_2$ La media de dos muestras son significativamente distintas.

Pruabas pos anovas

Tukey HSD

Prueba chi cuadrado

Permite determinar si dos variables cualitativas están o no asociadas.

H_0: No hay asociación entre las variables

H_a: Sí hay asociación entre las variables, es decir, el bajo peso y el fumar durante la gestación están asociados.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Prueba de normalidad

- Shapiro.wilk
- qqnorm-qqline
- Anderson darling

H_0: Los datos siguen una distribución normal.
H_A: Los datos no siguen un distribución normal.

Pruebas con Rangos

- Prueba de Wilcoxon
- Prueba U de man witney

H_0: $Me_1 = Me_2$ La mediana de dos muestras son iguales.
H_A: $Me_1 \neq Me_2$ La mediana de dos muestras son significativamente distintas.

Pruebas con Rangos

- Prueba de Kruskal-Wallis
- Prueba de friedman.

H_0: $Me_1 = Me_2$ = … = Me_p La mediana de las muestras son iguales.

H_A: $Me_i \neq Me_j$ Al menos dos medianas de dos muestras son significativamente distintas.

Bibliografía

- http://www.tuveras.com/estadistica/estadistica02.htm
- http://es.scribd.com/doc/70141495/Curso-breve-de-Estadistica