

## IME

MATRIZ: CADA COLUMNA REPRESENTA UNA VARIABLE O CARACTÉRISTICA

CADA FILA CORRESPONDE A UNA UNIDAD DE OBSERVACIÓN O INSTANCIA

## Tipos de Variables

NUMÉRICAS: VALORES NUMÉRICOS Y SON SENSIBLES A OPERACIONES ARITMÉTICAS

- CONTINUAS: TOMAN CUALQUIER VALOR (EN UN INTERVALO) DEL CONJUNTO DE LOS REALES

- DISCRETAS: NO ES POSIBLE QUE TOMEN CUALQUIER VALOR (EN UN INTERVALO). POR EJ PODRIAN TOMAR SOLO VALORES POSITIVOS (EJ "ANTIGÜEDAD")

CATEGÓRICAS: TOMAN UN VALOR ENTRE UN CONJUNTO ACOTADO. CADA VALOR SE CONOCÉ COMO NIVEL

- NOMINALES: NO EXISTE ORDEN NATURAL ENTRE LOS NIVELES (EJ "GÉNERO")

- ORDINALES: EXISTE UN ORDEN NATURAL ENTRE LOS NIVELES

## Tipos de Escalas

- NOMINAL: LOS VALORES SOLO SON NOMBRES O ESTADOS, NO SE PUEDEN HACER OPERACIONES ARITMÉTICAS, NO SE PUEDE ESTABLECER RELACIONES DE ORDEN

- ORDINAL & DE RANGOS: MISMO QUE LA NOMINAL, PERO SI SE PUEDEN ESTABLECER RELACIONES DE ORDEN (LA VARIABLE DEBE TENER MÍNIMO 3 NIVELES)

- **INTERVALO:** PARA DATOS CONTINOS o DISCRETOS CON UNA GRAN CANTIDAD DE NIVELLES.  
NOTIÓN DE ORIGEN, DISTANCIA ENTRE 2 VALORES ES CONOCIDA Y CONSTANTE, POR LO QUE SE PUEDEN REALIZAR OPERACIONES ARITMÉTICAS. PUNTO CERO Y UNIDAD DE MEDIDA ARBITRARIOS
- **RATÓN:** MISMO QUE ESCALA DE INTERVALOS, PERO TIENE SU ORIGEN EN UN CERO VERDADERO (POR EJ. MEDIR LA MASSA o DISTANCIA). LA DIFERENCIA ENTRE DOS PUNTOS ES INDEPENDIENTE DE LA UNIDAD DE MEDIDA

## Relaciones entre variables

Independientes: No existe asociación o relación entre las variables

Dependientes: Existe una asociación o relación entre las variables

- Asociación positiva: Si una variable crece, la otra también lo hace

- Asociación negativa: Si una variable crece, la otra decrece

PARÁMETRO: CUALQUIER NÚMERO que DESCRIBA UNA POBLACIÓN EN FORMA RESUMIDA (ej "MEDIA POBLACIONAL")

ESTADÍSTICO: CUALQUIER CANTIDAD cuyo valor puede ser CALCULADO A PARTIR DE DATOS MUESTRALES (ej "MEDIA, MEDIANA, DESVIACIÓN ESTÁNDAR")

## CAP 2

### ESTADÍSTICAS DESCRIPTIVAS

Permiten sintetizar y describir los DATOS. Pueden aplicarse tanto a UNA muestra como a UNA población.

Cuando se aplican a una muestra, se conoce como ESTIMADOR PUNTUAL de la misma medida para la población.

Como es una estimación no es EXACTA, pero la precisión TIENDE A AUMENTAR mientras MAYOR sea el TAMAÑO de la muestra.

DISTRIBUCIÓN DE FRECUENCIA: Representa CUANTAS VECES aparece CADA VALOR para una VARIABLE en UN CONJUNTO DE DATOS

### ESTADÍSTICAS DESCRIPTIVAS PARA DATOS NUMÉRICOS

MEDIA; media ARITMÉTICA; promedio. ~~Media Aritmética~~

MEDIA MUESTRAL  $\rightarrow \bar{x}$

MEDIA POBLACIONAL  $\rightarrow \mu$

Podemos entender la MEDIA como el PUNTO DE EQUILIBRIO de la DISTRIBUCIÓN. CORRESPONDE A UNA MEDIDA DE TENDENCIA CENTRAL.

MEDIANA: valor central de los valores PREVIAMENTE ORDENADOS

MODA: Es el VALOR MÁS FRECUENTE en el CONJUNTO DE DATOS

Dependiendo de la CANTIDAD de MODAS, se habla de DISTRIBUCIONES UNIMODALES, BIMODALES y MULTIMODALES

Es importante conocer la variabilidad o dispersión del conjunto de datos, esto nos permite saber qué tan semejantes (o diferentes) son las observaciones entre sí.

Desviación de las observaciones: Distancia entre una observación y la media del conjunto de datos

Medidas de dispersión: Varianza y desviación estandar

Al igual que con la media, se pueden obtener estimaciones puntuales de la varianza y desviación estandar de la población, denotadas por  $\sigma^2$  y  $\sigma$ , respectivamente

Rango: Muestra de los valores extremos, es decir, el mínimo y el máximo de una variable

Formas de dividir un conjunto de datos

Cada fragmento del conjunto de datos dividido se conoce como cuantil

Percentiles: Dividen el conjunto de datos en 100 subconjuntos de igual tamaño

Deciles: Dividen el conjunto de datos en 10 " " " "

Quintiles: Dividen el conjunto de datos en 5 " " " "

Cuartiles: Dividen el conjunto de datos en 4 " " " "

IQR (Rango Intercuartil): Mientras más disperso sea el conjunto de datos, mayor será el valor del IQR

Valores atípicos o outliers: Observaciones que parecen estar fuera de rango. La mediana es una buena medida de tendencia central y el IQR una buena medida de dispersión

## ESTADÍSTICA DESCRIPTIVA PARA DATOS CATEGÓRICOS

Frecuencia: cantidad de veces que podemos encontrar cada nivel de la variable en los datos

Proporción: corresponde a la frecuencia relativa, frecuencia de un nivel de la variable dividida por la cantidad total de observaciones

Tabla de contingencia - matriz de confusión - tabla de frecuencias: cada fila representa la cantidad de veces en que ocurre una combinación de variables

Tabla de frecuencias relativas: utiliza porcentajes o proporciones

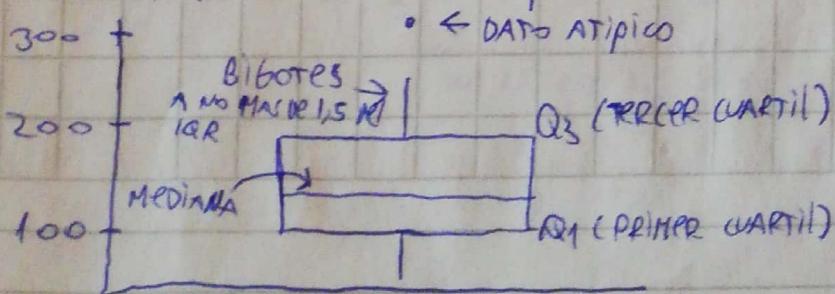
## GRÁFICOS

- 1 variable numérica: Histograma, muestra una aproximación a la densidad (o distribución de frecuencias) para la variable

Asimetría negativa: se da cuando las barras de la izq son más grandes que los de la derecha, es decir, la distribución es desviada a la izquierda

Asimetría positiva: distribución desviada a la derecha

Gráfico de cajas: considera 5 estadísticos para su construcción y facilita la identificación de datos atípicos



- 1 VARIABLE CATEGÓRICA : GRÁFICO DE BARRAS , CADA BARRA ES TAN LARGA COMO LA PROPORCIÓN DE VALORES PRESENTES EN CADA NIVEL DE LA VARIABLE

### GRÁFICO DE TORZA

- 2 VARIABLES NUMÉRICAS : GRÁFICOS DE DISPERSIÓN , MUESTRAN INFORMACIÓN CASO A CASO , YA QUE CADA PUNTO DEL GRÁFICO CORRESPONDE A UNA OBSERVACIÓN

TAMBÍEN SIRVEN PARA IDENTIFICAR SI DOS (O MÁS) VARIABLES ESTÁN RELACIONADAS

- 2 VARIABLES CATEGÓRICAS : GRÁFICOS DE BARRAS APILADAS , AGRUPADAS Y ESTANDARIZADAS , PERMITEN VISUALIZAR LA MATRIZ DE CONTIGÜIDAD ENTRE 2 VARIABLES Y ENCONTRAR POSIBLES RELACIONES

GRÁFICO DE MOSAICO : TAMBÍEN PERMITE REPRESENTAR LA TABLA DE CONTINGÜENCIA

- 1 VARIABLE NUMÉRICA Y 1 CATEGÓRICA : GRÁFICO DE CAJAS POR GRUPOS

SÍ LA CANTIDAD DE OBSERVACIONES ES PEQUEÑA , EL GRÁFICO DE TIROS ES SIMILAR AL GRÁFICO DE DISPERSIÓN

## Resumen Gráficos

Una variable numérica:

Histograma →

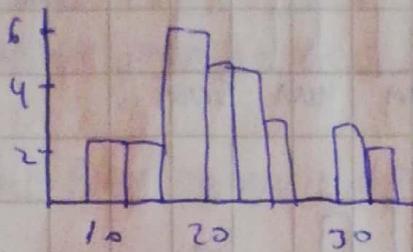
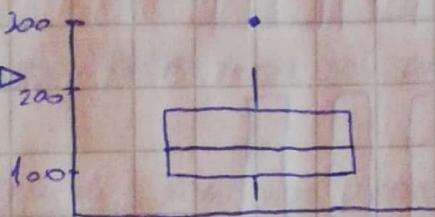


Grafico de cajas →



Una variable categórica:

Grafico de barras →

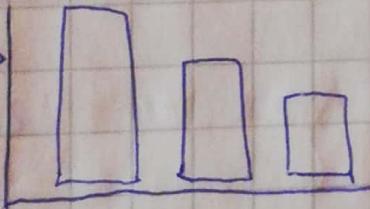
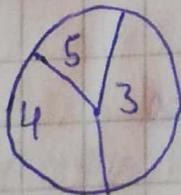
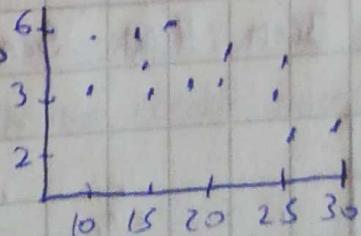


Grafico de tortas →



Dos variables numéricas:

Grafico de dispersión →

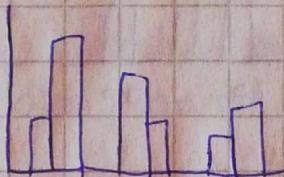


## Dos variables categóricas:

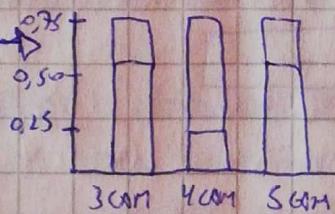
- BARRAS APILADAS →



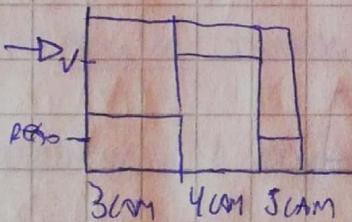
BARRAS AGRUPADAS →



BARRAS ESTANDARIZADAS →



MOSAICO →



## Una variable numérica y otra categórica

CAJAS POR grupo →

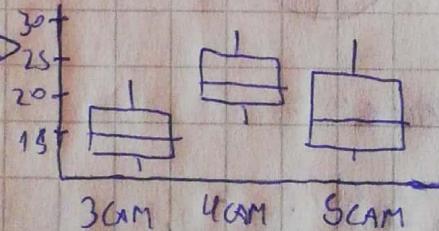
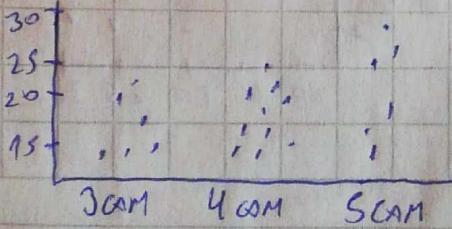


GRAFICO DE TIPOS →



## CAP 3

VARIABLES ALEATORIAS: se define como una variable a proceso cuyo resultado sea numérico

DISTRIBUCIÓN DE PROBABILIDAD: define la probabilidad de que ocurren los diferentes valores que dicha variable puede tomar

VARIABLE ALEATORIA CONTINUA: Puede tomar cualquiera de los infinitos valores posibles dentro de un intervalo

VARIABLE ALEATORIA DISCRETA: Puede tomar un conjunto finito de valores

VALOR ESPERADO -  $E(X) - \mu$ : Resultado promedio de una variable aleatoria

VARIANZA GENERAL -  $Var(X) - \sigma^2$ : Permite calcular que tan alejado podría estar el valor obtenido del valor esperado

### DISTRIBUCIONES CONTINUAS

DISTRIBUCIÓN NORMAL O DISTRIBUCIÓN GAUSSIANA: UNIMODAL, SIMÉTRICA, FORMA DE CAMPAÑA

$$N(\mu, \sigma)$$

$\mu$ : LA MEDIA, DESPLAZA EL CENTRO DE LA CURVA A LO LARGO DEL EJE X

$\sigma$ : LA DESVIACIÓN ESTÁNDAR, MODIFICA QUE TAN DISPERSOS ESTAN LOS DATOS CON RESPECTO A LA MEDIA

DISTRIBUCIÓN NORMAL ESTÁNDAR

→ DISTRIBUCIÓN Z: RENOMBRADA DE ESTANDARIZACIÓN PARA DETERMINAR QUE TAN USUAL O INUSUAL ES UN DETERMINADO VALOR EN UNA ESCALA ÚNICA. ESTA CENTRADA EN 0 Y TIENE  $\sigma = 1$

Distribución chi-cuadrado +  $\chi^2$ : Se usa para caracterizar valores siempre positivos y habitualmente desviados a la derecha.

Grados de libertad (v): Estimación de la cantidad de observaciones empleadas para calcular un estimador. Cantidad de valores que pueden cambiar libremente en un conjunto de datos.

$$\mu = v, \sigma = \sqrt{2v}$$

Distribución t de Student: Para muestras pequeñas. A mayor grados de libertad más se asemeja a la normal.

MEDIA PARA  $v > 1 \rightarrow \mu = 0$ . LA DESVIACIÓN ESTÁNDAR PARA  $v > 2 \rightarrow \sigma = \sqrt{\frac{v}{v-2}}$

Distribución F: Para comparar varianzas

$$\frac{\chi_1^2(v_1)}{v_1} \quad ; \quad \frac{\chi_2^2(v_2)}{v_2}$$

$$v_2 > 2 \rightarrow \mu = \frac{v_2}{v_2 - 2} ; \quad v_2 > 4 \rightarrow \sigma = \sqrt{\frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}}$$

## DISTRIBUCIONES DISCRETAS

DISTRIBUCIÓN DE BERNOULLI: EN CADA INTENTO INDIVIDUAL TIENE SÓLO 2 RESULTADOS POSIBLES, ÉXITO Y FRACASO.

- Proporción de la muestra ( $\hat{p}$ ): CANTIDAD DE ~~ÉXITOS~~ EXITOS / CANTIDAD DE INTENTOS.  
A MAYOR CANTIDAD DE INTENTOS, MÁS CERCANO SERÁ  $\hat{p}$  A LA PROB. REAL DE ÉXITO  $p$ .

DISTRIBUCIÓN GEOMÉTRICA: CANTIDAD DE INTENTOS QUE DEBEMOS REALIZAR HASTA OBTENER UN ÉXITO, PARA VARIABLES DE BERNOULLI INDEPENDIENTES E IDENTICAMENTE DISTRIBUIDAS.

DISTRIBUCIÓN BINOMIAL: PROBABILIDAD DE TENER EXACTAMENTE  $K$  EXITOS EN  $N$  INTENTOS INDEPENDIENTES DE BERNOULLI CON PROB. DE ÉXITO  $p$ .  
CONDICIONES:

- INTENTOS SON INDEPENDIENTES
- LA CANTIDAD DE INTENTOS ( $n$ ) ES FIJA
- EL RESULTADO DE CADA INTENTO PUEDE SER CLASIFICADO COMO ÉXITO O FRACASO
- LA PROB. DE ÉXITO ( $p$ ) ES LA MISMA PARA CADA INTENTO

DISTRIBUCIÓN BINOMIAL NEGATIVA: PROBABILIDAD DE ENCONTRAR EL  $K$ -ÉSIMO ÉXITO AL  $N$ -ÉSIMO INTENTO. SE EXAMINA CUANTOS INTENTOS SE NECESITAN PARA OBTENER UNA CANTIDAD FIJA DE ÉXITOS Y SE REQUIERE QUE LA ÚLTIMA OBS. SEA UN ÉXITO.

CONDICIONES:

- INTENTOS SON INDEPENDIENTES
- RESULTADOS DE CADA INTENTO ÉXITO → FRACASO
- PROB. DE ÉXITO ( $p$ ) ES LA MISMA PARA CADA INTENTO
- ÚLTIMO INTENTO DEBE SER UN ÉXITO

Distribución de Poisson: cantidad de eventos en una población grande en un lapso de tiempo dado

## CAP 4

INFERENCIA ESTADÍSTICA: cuan cerca está el estadístico del parámetro real de la población

Estadístico: estimador puntual de un parámetro

Distribución muestral: distribución de estimadores puntuales obtenidos con todas las diferentes muestras de igual tamaño de una misma población

Modelo estadístico: descripción de un proceso probabilístico con parámetros desconocidos que deben ser estimados en base a suposiciones y un conjunto de datos observados

ESTIMADOR PUNTUAL: es un único valor (obtenido a partir de una muestra) que estima un parámetro de la población

ERROR ESTÁNDAR: corresponde a la desviación estándar de la distribución de un estimador muestral  $\hat{\theta}$  de un parámetro  $\theta$ .  $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

Intervalo de confianza:  $\bar{x} \pm \underbrace{z^* \cdot SE_{\bar{x}}}_{\text{MARGEN DE ERROR}}$

si estamos con un Nivel de confianza 90% y es prueba bilateral  $1-\alpha=0,9$   
 $\alpha=0,1$

para obtener  $z^* \Rightarrow \frac{\alpha}{2}=0,05 \quad qnorm(0.05, mean=0, sd=1, lower.tail=false)$

## Pruebas de hipótesis

Hipótesis Nula: Suelo representar una postura escéptica, siempre se formula con una igualdad

Hipótesis alternativa: Nueva perspectiva

Valor nulo: Representa el valor del parámetro cuando se cumple la hipótesis Nula

Error de tipo I: Correspondiente a rechazar  $H_0$  cuando en realidad es verdadera

Error de tipo II: Correspondiente a no rechazar  $H_0$  cuando en realidad  $H_A$  es verdadera

\* Si usamos un intervalo de confianza de 95% para evaluar una prueba de hipótesis en que la hipótesis Nula es verdadera, cometeremos un error de tipo I cada vez que el estimador puntual ( $\hat{z}^*$ ) esté a 1,96 o más errores estándar del parámetro de la población.

Prueba formal de hipótesis con valores p:  $\hat{z} < z \rightarrow$  zona de rechazo de  $H_0$   
 $|z| < \hat{z}^* \rightarrow$  se falla en rechazar  $H_0$

Valor p & p-value: Probabilidad de observar datos al menos tan favorables como la muestra actual para la hipótesis alternativa, si la hipótesis Nula es verdadera.

Un p-valor permite cuantificar cuan fuerte es la evidencia en contra de la hipótesis Nula (y en favor de la hipótesis alternativa)

Prueba unilateral: Se usan cuando se desea verificar un incremento o un decremento

El área bajo la curva normal con valores  $\leq -z$  a un lado se calcula usando el valor  $z$

Cuanto menor sea el valor  $p$ , más fuerte sera la evidencia en favor de  $H_A$  por sobre  $H_0$ .  $p < \alpha \rightarrow$  evidencia suficiente para rechazar  $H_0$

Prueba bilateral:

$\alpha = 0,050 \quad p > \alpha \rightarrow$  se falla en rechazar  $H_0$  (se acepta  $H_0$ )  
 $\alpha = 0,05 \quad p < \alpha \rightarrow$  se rechaza  $H_0$  en favor de  $H_A$

Efecto de nivel de significación: representa la proporción de veces en que se comete un error de tipo I,

\* Para reducir la prob. de rechazar  $H_0 \rightarrow$  Menor nivel de significancia

\* Si el costo de cometer un error tipo II es mayor  $\rightarrow$  Mayor nivel de significancia

## CAP 5

DISTRIBUCIÓN MUESTRAL DE LA MEDIA sigue APROXIMADAMENTE UNA DISTRIBUCIÓN NORMAL, SI LA MUESTRA TIENE AL MENOS 30 OBSERVACIONES

VALOR  
NÚMERO  
↑

PRUEBA Z: INFERIR ACERCA DE MEDIAS CON UNA O DOS MUESTRAS  $Z = \frac{\bar{X} - \mu}{\sigma}$

- LA MUESTRA DEBE TENER MÍNIMO 30 OBSERVACIONES, SI LA MUESTRA TIENE MENOS SE DEBE CONOCER LA VARIANZA DE LA POBLACIÓN
- LAS OBS. DEBEN SER INDEPENDIENTES, LA ELECCIÓN DE UNA OBSERVACIÓN PARA LA MUESTRA NO INFUYE EN LA SELECCIÓN DE LAS OTRAS
- LA POBLACIÓN Sigue APROXIMADAMENTE UNA DISTRIBUCIÓN NORMAL

## Shapiro TEST (P-value)

- $H_0$ : MUESTRA VIENE DE DISTRIBUCIÓN NORMAL
- $H_A$ : MUESTRA VIENE DE DISTRIBUCIÓN DIFERENTE A LA NORMAL

$P > \alpha \Rightarrow$  Se acepta  $H_0$ , sigue UNA DISTRIBUCIÓN NORMAL

$P < \alpha \Rightarrow$  Se acepta  $H_A$ , NO Sigue UNA DISTRIBUCIÓN NORMAL

## Prueba T de Student

- UNA MUESTRA: - OBS SON INDEPENDIENTES

- OBS PROVIENEN DE UNA DISTRIBUCIÓN CERCA A LA NORMAL

- DOS MUESTRAS PAREJAS

- DOS MUESTRAS INDEPENDIENTES: - CADA MUESTRA cumple LAS CONDICIONES PARA USAR LA DISTRIBUCIÓN t  
- LAS MUESTRAS SON INDEPENDIENTES ENTRE SI

## CAP 6

$\alpha$  (Nivel de significación): Probabilidad de cometer un error de tipo I

$\beta$ : Probabilidad de cometer errores de tipo II

Poder estadístico ( $1-\beta$ ): Probabilidad de correctamente RECHAZAR  $H_0$  CUANDO ES FALSA (Prob. DE NO COMETER UN ERROR DE TIPO II)

Tamaño del efecto: CUANTIFICACIÓN DE LA DIFERENCIA ENTRE DOS GRUPOS, + DEL VALOR OBSERVADO CON RESPECTO AL VALOR NULO

D de Cohen:  $d = 0,2$  EFECTO PEQUEÑO (IMPERCEPTIBLE A SIMPLE VISTA)

$d = 0,5$  EFECTO MEDIANO (PERCEPCIONABLE A SIMPLE VISTA)  $\rightarrow$  PROBABILMENTE

$d = 0,8$  EFECTO GRANDE (DEFINITIVAMENTE PERCEPCIONABLE A SIMPLE VISTA)

Prueba T de una muestra:  $d = \frac{\bar{x} - \mu_0}{s}$

Prueba T de dos medias independientes:  $n > 50 \rightarrow d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$

$$\cdot n < 50 \rightarrow d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \cdot \frac{n_1 + n_2 - 3}{n_1 + n_2 - 2,75}$$

$$s_p = \sqrt{\frac{\sum (x - \bar{x}_1)^2 + \sum (x - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$$\cdot \text{VARIANTE DE WELCH}: d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

Prueba T Dos muestras pareadas:  $n > 50 \rightarrow d = \frac{\bar{X}_D}{S_D}$

$$n < 50 \rightarrow d = \frac{\bar{X}_D}{S_D} \cdot \frac{n_1 - 2}{n_1 - 1,25}$$

## CAP 7

Método de Wald:  $\hat{p}$  (estimador puntual correspondiente a la proporción de éxito de la muestra)

Este estimador se distribuye de manera cercana a la normal si se cumple:

- obs de la muestra son independientes
- se cumple la condición de éxito-fallido,  $np \geq 10$  y  $n(1-p) \geq 10$

Si se cumple se dice que  $\hat{p}$  es cercano a la normal con media  $\mu = p$ , desviación estandar  $\sigma = \sqrt{p(1-p)}$  y error estandar  $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Una Proporción: - intervalo de confianza  $\hat{p} \pm z^* \cdot SE$

USANDO el Modelo NORMAL, la segunda condición se verifica usando el valor nulo, denotado por  $p_0$ . ERROR ESTANDAR  $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$  y estadístico  $Z = \frac{\hat{p} - p_0}{SE}$

Dos Proporciones: ESTIMADOR PUNTUAL  $\hat{p}_1 - \hat{p}_2$

- CADA proporción, por separado, sigue el Modelo NORMAL
- LAS DOS MUESTRAS SON INDEPENDIENTES UNA DE LA OTRA

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

intervalo de confianza:  $\hat{p}_1 - \hat{p}_2 \pm z^* SE_{\hat{p}_1 - \hat{p}_2}$

Valor nulo = 0

CUANDO LA HIPÓTESIS NULA SUPONE QUE NO HAY DIFERENCIA ENTRE LAS PROPORCIONES, LA VERIFICACIÓN DE LA CONDICIÓN DE ÉXITO-FRACASO Y LA ESTIMACIÓN DEL ERROR ESTÁNDAR SE REALIZAN USANDO LA PROPORCIÓN AGREGADA

CANTIDAD DE ÉXITOS EN LAS MUESTRAS  $\rightarrow \hat{P}_1 n_1$  Y  $\hat{P}_2 n_2$

$$\hat{P} = \frac{\text{NRO DE ÉXITOS}}{\text{NRO DE CASOS}} = \frac{\hat{P}_1 n_1 + \hat{P}_2 n_2}{n_1 + n_2}$$

$$SE = \sqrt{\frac{\hat{P}(1-\hat{P})}{n_1} + \frac{\hat{P}(1-\hat{P})}{n_2}}$$

$$\text{ESTIMADOR PUNTAZ} = \hat{P}_1 - \hat{P}_2$$

$$Z = \frac{\text{ESTIMADOR PUNTAZ} - \text{VALOR NULO}}{SE}$$

Valor nulo  $\neq 0$

- CONDICIÓN DE ÉXITO-FRACASO SE REALIZA DE MANERA INDEPENDIENTE PARA AMBAS MUESTRAS

$$- SE = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

MÉTODO DE WILSON ; PROP.TEST(). LIMITANTE AL TRABATAR CON DOS PROPORCIONES NO PERMITE ESTABLECER UN VALOR NULO DISTINTO DE 0 PARA LA DIFERENCIA

## CAP 8

Pruebas no paramétricas o libres de distribución: No mencionan un parámetro para las hipótesis nula y alternativa. Además en ninguna de ellos se hace alguna suposición sobre la distribución de la población.

- Entregan menos información
- Hipótesis del tipo "Las poblaciones muestran las mismas proporciones" versus "Las poblaciones muestran proporciones distintas". Ninguna indica cuáles serían esas proporciones.
- Presentan menor poder estadístico y suelen necesitar muestras de mayor tamaño para detectar diferencias significativas

Prueba chi-cuadrado de Pearson (Prueba  $\chi^2$  de Asociación):

- Inferir con proporciones de dos variables categóricas y una de ellas debe ser dicotómica (tiene solo 2 niveles)
- obs independientes entre sí
- debe haber a lo menos 5 obs esperadas en cada grupo

Prueba chi-cuadrado de homogeneidad: Es adecuada si queremos determinar si dos poblaciones (variable dicotómica) presentan las mismas proporciones en los diferentes niveles de una variable categórica.

Prueba chi-cuadrado de bondad de ajuste: Permite comprobar si una distribución de frecuencias observada se asemeja a una distribución esperada. Usualmente se emplea para comprobar si una muestra es representativa de la población.

Prueba chi-cuadrado de independencia: Permite determinar si dos variables categóricas, de una misma población, son estadísticamente independientes, o por el contrario si están relacionadas.

### Pruebas para muestras pequeñas

Prueba exacta de Fisher: Alternativa a la prueba  $\chi^2$  de independencia en el caso que ambas variables sean dicotómicas

Prueba de McNemar: Resulta apropiada cuando una misma característica, con respuesta dicotómica, se mide en dos ocasiones diferentes para los mismos sujetos (muestras pareadas) y queremos determinar si se produce un cambio significativo entre ambas mediciones.

- Prueba Q de Cochran: Es una extensión de la prueba de McNemar, adecuada cuando la variable de respuesta es dicotómica y la variable independiente tiene más de dos observaciones pareadas (cuando ambas variables son dicotómicas)

- Variable de respuesta es dicotómica
  - La variable independiente es categórica
  - Obs son independientes entre sí
  - El tamaño de la muestra es lo suficientemente grande,  $n \cdot k \geq 24$
- $n \rightarrow$  Tamaño de muestra (cantidad de instancias)  
 $k \rightarrow$  cantidad de niveles en la variable independiente

Ómnibus: Tipo de hipótesis nula que comprueba la igualdad de todas las proporciones

Pruebas post-hoc - a posteriori: Se realizan una vez que se ha concluido gracias a la prueba omnibus que existen diferencias significativas

Este procedimiento se hace solo si la prueba omnibus rechaza la hipótesis nula en favor de la hipótesis alternativa

$P < HB$  (factor de holm)  $\rightarrow$  existe una diferencia significativa