Capitulo 1:

Población: conjunto de individuos o elementos

Muestra: subconjunto de una población

Modelo estadístico (Kaplan):representación con un propósito particular.

-Simplifica de la realidad en la que aparecen algunas de sus propiedades, permite estudiar de forma simple y comprensible una porción de la realidad, resumen de manera conveniente a juicio de sus creadores , se buscan los aspectos más relevantes del fenómeno estudiado y sus relaciones (seleccionar y observar)

Útiles para describir , clasificar y anticipar el resultado de intervenciones.

Respecto a los datos:

Suelen almacenarse en matrices o tablas (matrices de datos): facilita modificación del conjunto de datos y el acceso a cada dato.

Columnas representan variables (o característica)- Filas corresponden unidades de observación (o instancia)

Importante reconocer: a qué corresponde cada característica, unidades de medición y valor posibles a tomar por cada dato (rangos y tipos)

Tipos de variables:

Numéricas: -**Continuas**: valores en intervalos (números reales)(ej: estatura) - **Discretas**: (valores enteros no negativos)(ej: año de nacimiento)

Categóricas: - **Nominales:** no poseen orden natural entre los niveles (ej: género, rama) - **Ordinales:** orden natural (ej: jerarquías)

Respecto a relaciones entre variables:

Independientes: no existe asociación o relación con otras varibles

Dependientes: **Asociación positiva:** si una crece la otra también lo hará - **Asociación negativa:** si una crece la otra decrece.

Parámetro: número que describe una población en forma resumida (ej:promedio).

Estadístico: cualquier cantidad cuyo valor puede ser calculado a partir de datos muestrales

Respecto a R:

La primera fila se usa para los nombres de las columnas o variables.

La primera columna contiene los nombres de las observaciones, que deben ser únicos.

Los nombres de las columnas deben respetar las convenciones de R:

• No está permitido el uso de espacios ni símbolos especiales (?, $, \*, +, #, (, ), -, /, }, {, |, >, <,

etc.). Solo se admite el uso de puntos (.) y guiones bajos (\_).

• Los nombres de variables no pueden comenzar con un dígito.

• Los nombres de las columnas deben ser únicos.

R es sensible a las mayúsculas.

No puede haber filas en blanco.

No debe tener comentarios.

Los valores faltantes deben ser denotados mediante NA

Para columnas con fechas, se usa el formato mm/dd/aaaa.

Debe tener uno de los siguientes formatos, ejemplificados en la figura 2.1:

• Extensión .txt con tabulaciones como delimitador y punto decimal para valores reales.

• Extensión .csv en formato inglés: con punto y coma (;) como delimitador y punto decimal para

valores reales.

• Extensión .csv en formato español: con comas (,) como delimitador y coma decimal para valores

reales.

DE LOS EJEMPLOS ESTUDIADOS

Script 2.1: Sentencias para importar un conjunto de datos

1 # Cargar un conjunto de datos disponible en R.

2 datos 1 <- mtcars

3

4 # Importar desde un archivo de texto plano delimitado por tabuladores .

5 datos 2 <- read . delim ( file . choose ())

7 # Importar desde un archivo de valores separados por coma

8 # en formato ingl és.

datos 3 <- read .csv("C:\\ Inferencia \\ ejemplo 1-csv - eng.csv")

10

11 # Configurar carpeta de trabajo

12 setwd ("C:\\ Inferencia ")

13

14 # Importar desde un archivo de valores separados por coma

15 # en formato espa ñol.

16 datos 4 <- read .csv2(" ejemplo 1-csv -esp.csv")

17

18 # Mostrar las primeras 6 filas del conjunto de datos

19 # almacenado en datos 1.

20 head ( datos 1)

21

22 # Mostrar las ú ltimas 6 filas del conjunto de datos

23 # almacenado en datos 1.

24 tail ( datos 1)

Script 2.2: Sentencia para instalar un paquete de R.

1 # Instalar un paquete .

2 install . packages (" ggpubr ")

3

4 # Primera forma de importar un paquete .

5 library ( ggpubr )

6

7 # Segunda forma de importar un paquete .

8 require ( ggplot 2)

9

10 # Importar un paquete , instal á ndolo de ser necesario .

11 if (! require ( dplyr )){

12 install . packages (" dplyr ", dependencies = TRUE )

13 require ( dplyr )

14 }

Script 2.3: Construir un dataframe.

1 # Crear un vector de strings y guardarlo en la variable nombre .

2 nombre <- c(" Alan Brito Delgado ",

3 " Zacar ías Labarca del Río",

4 " Elsa Payo Maduro ")

5

6 # Crear un vector de fechas y guardarlo en la variable

7 # fecha \_ nacimiento .

8 fecha \_ nacimiento <- as. Date (c("2008-1-25", "2006-10-4", "2008-3-27"))

9

10 # Crear tres vectores de reales entre 1.0 y 7.0 y guardarlos

11 # en prueba \_i, respectivamente .

12 prueba \_1 <- c(5.5, 3.4, 4.5)

13 prueba \_2 <- c(3.2, 4.7, 4.1)

14 prueba \_3 <- c(4.8, 4.3, 5.1)

15

16 # Construir un data frame a partir de los vectores anteriores y

17 # guardarlo en la variable dataframe .

18 dataframe <- data . frame (nombre ,

19 fecha \_ nacimiento ,

20 prueba \_1,

21 prueba \_2,

22 prueba \_3,

23 stringsAsFactors = FALSE )

Script 3.1: Cálculo de la media en R.

1 dataframe <- mtcars

2

3 # Calcular la media para la variable mpg.

4 mean ( dataframe $mpg )

5

6 # Calcular la media para todas las variables .

7 sapply ( dataframe , mean )

8

9 # Calcular la media para la segunda y quinta columnas

10 # ( variables cyl y drat ).

11 sapply ( dataframe [c(2, 5)], mean )

12

13 # Calcular la media para las 7 primeras columnas

14 # ( variables mpg , cyl , disp , hp , drat , wt y qsec ).

15 sapply ( dataframe [1:7], mean )

16

17 # Calcular la media para la variable mpg omitiendo

18 # valores faltantes .

19 mean ( dataframe $mpg , na.rm = TRUE )

# Cá lculo de percentiles para la variable mpg.

4 # Cuartiles .

5 quantile ( dataframe $mpg)

6

7 # Quintiles .

8 quantile ( dataframe $mpg , seq(0, 1, 0.2))

9

10 # Deciles .

11 quantile ( dataframe $mpg , seq(0, 1, 0.1))

12

13 # Percentiles .

14 quantile ( dataframe $mpg , seq(0, 1, 0.01))

GRAFICOS

GRAFICO DE DISPERSION

library ( ggpubr )

3 dataframe <- mtcars

5 ggscatter ( dataframe ,

6 x = "mpg",

7 y = "wt",

8 color = " red",

9 title = " Rendimiento v/s peso ",

10 xlab = " Millas por galón",

11 ylab = " Peso "

12 )

HISTOGRAMA

5 gghistogram ( dataframe ,

6 x = "mpg",

7 bins = 10,

8 add = " mean ",

9 title =" Rendimiento ",

10 xlab =" Millas por galón",

11 ylab =" Cantidad ",

12 color = " blue ",

13 fill = " blue ")

14

15 gghistogram ( dataframe ,

16 x = "hp",

17 bins = 10,

18 add = " mean ",

19 title =" Potencia ",

20 xlab =" Caballos de fuerza ",

21 ylab =" Cantidad ",

22 color = " red",

23 fill = " yellow ")

GRAFICO CAJA

5 ggboxplot ( dataframe $hp ,

6 title =" Potencia [hp]",

7 color = " red",

8 fill = " pink "

9 ) + rremove ("x. ticks ")

10 + rremove ("x. text ")

11 + rremove (" axis . title ")

GRAFICO FUNCION ACUMULATIVA

5 # Grá fico ECDF para la variable mpg .

6 ggecdf ( dataframe ,

7 x = "mpg",

8 color = " blue ")

GRAFICO CUANTIL.CUANTIL

5 # Grá fico Q-Q para la variable mpg.

6 ggqqplot ( dataframe ,

7 x = "mpg",

8 color = " red")

GRAFICO DE BARRAS

5 # Crear la tabla de frecuencias y convertirla a data frame .

6 contingencia <- as. data . frame ( xtabs (~ gear , data = dataframe ))

8 # Graficar .

9 ggbarplot ( contingencia ,

10 x = " gear ",

11 y = " Freq ",

12 fill = c(" brown ", " purple ", " orange "),

13 title = " Cantidad de marchas de los autom ó viles ",

14 xlab = " Cantidad de marchas ",

15 ylab = " Frecuencia ")

GRAFICO DE TORTA

5 # Crear la tabla de frecuencias y convertirla a data frame .

6 contingencia <- as. data . frame ( xtabs (~ gear , data = dataframe ))

8 # Graficar .

9 ggpie ( contingencia ,

10 x = " Freq ",

11 label = " gear ",

12 fill = c(" brown ", " purple ", " orange "),

13 title = " Cantidad de marchas de los autom ó viles ")

Gráficos de barras para dos variables.

1 library ( ggpubr )

3 dataframe <- mtcars

5 # Crear tabla de contingencia para las variables vs y gear ,

6 # y guardarla como data frame .

7 tabla <- xtabs (~ vs + gear , data = dataframe )

8 contingencia <- as. data . frame ( tabla )

TABLAS:

Tabla de contingencia para UNA VARIABLE

1 dataframe <- mtcars

3 # Crear tabla de contingencia para la variable gear .

4 contingencia <- table ( dataframe $ gear )

5 contingencia

6

7 # Otra forma de crear la misma tabla .

8 contingencia <- xtabs (~ gear , data = dataframe )

9 contingencia

10

11 # Calcular totales por fila y mostrarlos por separado .

12 marginSums ( contingencia )

13

14 # Calcular totales por fila y agregarlos a la tabla .

15 con \_ totales <- addmargins ( contingencia , 1)

16

17 # Convertir a tabla de proporciones

18 proporciones <- prop . table ( contingencia )

19 proporciones <- addmargins ( proporciones , 1)

20 proporciones

21

22 # Convertir a tabla de porcentajes con 2 decimales .

23 porcentajes <- round ( prop . table ( contingencia ), 4) \* 100

24 porcentajes <- addmargins ( porcentajes )

25 porcentajes

Tablas de contingencia y proporciones para DOS VARIABLES

1 dataframe <- mtcars

3 # Crear tabla de contingencia para las variables am y gear .

4 contingencia <- table ( dataframe $am , dataframe $ gear )

5 contingencia

7 # Otra forma de crear la misma tabla .

8 contingencia <- xtabs (~ am + gear , data = dataframe )

9 contingencia

11 # Proporciones por fila .

12 proporciones \_ fila <- prop . table ( contingencia , margin =1)

13 proporciones \_ fila <- addmargins ( proporciones \_fila , margin =2)

14 proporciones \_ fila

16 # Proporciones por columna .

17 proporciones \_ columna <- prop . table ( contingencia , margin =2)

18 proporciones \_ columna <- addmargins ( proporciones \_ columna , margin =1)

19 proporciones \_ columna

21 # Proporciones totales .

22 proporciones <- prop . table ( contingencia )

23 proporciones <- addmargins ( proporciones )

24 proporciones

Tabla de contingencia para TRES VARIABLES

1 dataframe <- mtcars

3 # Crear tabla de contingencia para las variables am ,

4 # gear y vs.

5 ftable ( dataframe $am , dataframe $gear , dataframe $vs)

7 # Otra forma de crear la misma tabla .

8 xtabs (~ gear + am + vs , data = dataframe )

De los ejercicios propuestos y respuestas del foro:

9. ¿Cómo se puede generar la secuencia 0.00, 0.25, 0.50, ..., 2.75, 3.00 en R?

Para esto se debe utilizar la función de R “seq(a,b,r)” la cual genera una lista de números que empieza en a y termina en b, de la forma a, a+r, a+2r, … por lo tanto “r” seria el intervalo o diferencia entre los números que serán generados. De esta forma para obtener la secuencia anteriormente solicitada se deberá usar :

seq(0, 3, by=0.25)

función distribución acumulativa empírica:

La función de distribución acumulada (CDF) calcula la probabilidad acumulada de un valor dado de x. Utilice la CDF para determinar la probabilidad de que una observación aleatoria que se toma de la población sea menor que o igual a cierto valor. También puede usar esta información para determinar la probabilidad de que una observación sea mayor que cierto valor o se encuentre entre dos valores.

La función de distribución empírica (FED) o cdf empírica es una función de paso que salta por 1/N a la ocurrencia de cada observación:

FDE(x)=1N∑i=1NI{x≤xi}

Donde

{A} es el indicador de la función de un evento

I{x≤xi}={10 if x≤xi if x>xi

Por definición, la función FDE calcula la distribución acumulativa del número aleatorio subyacente. El FDE estima la verdadera función de densidad acumulativa subyacente de los puntos en la muestra; Se garantiza virtualmente que converge con la distribución verdadera a medida que el tamaño de la muestra se hace lo suficientemente grande.

5. ¿Qué tipo de información buscada llevaría a utilizar un gráfico de dispersión?

Los gráficos de dispersión se caracterizan porque despliegan información caso a caso, pues cada punto del

gráfico corresponde a una observación. Se usan para averiguar la intensidad de la relación entre dos o más variables numéricas permitiendo responder preguntas sobre los datos, por ejemplo: ¿cuál es la relación entre dos variables? ¿Cómo está distribuido? ¿Dónde están los valores atípicos? Así como confirmar relaciones anticipadas entre dos conjuntos asociados de datos. En cuanto a las relaciones, podemos encontrar variables independientes, con asociación positiva (si una variable aumenta, la otra tiene tendencia al aumento) y asociación negativa (si una variable aumenta la otra tiene tendencia a la disminución).

14.¿Qué significa que un estadístico tenga un valor p de 0,025?

El valor p se define como la probabilidad de observar datos al menos tan favorables como la muestra actual para la hipótesis alternativa, si esta es verdadera", es una manera de cuantificar cuán fuerte es la evidencia en contra de la hipótesis nula (y en favor de la hipótesis alternativa). La determinación del valor p nos permite determinar si debemos rechazar o no rechazar una hipótesis reivindicada. Se establece el nivel de significancia, que sirve como el nivel de corte, para determinar si una hipótesis debe ser rechazada o no.

Suponiendo una significancia de α=0,05 en caso de tener un valor p= 0,025 implica una probabilidad del 0.025 y dado que cuanto menor sea el valor p, más fuerte será la evidencia en favor de HA por sobre H0. Un valor p menor que el nivel de significación se considera evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa.

De esta forma la hipótesis nula(H0) del estadístico sería rechazada a favor de la hipótesis alternativa(HA) (para 0,025 < 0,05), mientras se mantenga el supuesto de un nivel de significancia mayor a este valor p.

15.¿ Si una hipótesis nula es rechazada a un nivel de significación de 0,01, ¿será rechazada a un nivel de

significación 0,05? Explique.

En caso de que la hipótesis nula (H0) sea rechazada a un nivel de significación de 0.01, también seria rechazada ante un nivel de significación de 0.05, dado que el nivel de significación es un punto de corte para determinar si una hipótesis (H0) debe ser rechazada en caso de que el valor “p” sea menor a este nivel de significación establecido. Suponiendo un valor p = x , anteriormente al ser rechazado H0 se puede asumir que x < 0.01 (valor p menor a nivel de significación) , luego 0.01 < 0.05 y por lo tanto , x < 0.05 , lo que nos indica que , dado un valor p menos al nivel de significación , la hipótesis nula será rechazada.

Que significa que un estimador puntual tenga que ser insesgado

Se denomina sesgo de un estimador a la diferencia entre la esperanza (o valor esperado) del estimador y el verdadero valor del parámetro a estimar. Es deseable que un estimador sea insesgado o centrado, es decir, que su sesgo sea nulo por ser su esperanza igual al parámetro que se desea estimar.

Esto significa que la distribución muestral tiene su centro en el parámetro que ésta estima. En otras palabras, un estimador insesgado (como la media) tiende a proveer una estimación cercana al parámetro real.

Una primera propiedad deseable para un estimador es que el centro de la distribución de los valores que puede tomar coincida con el valor del parámetro que queremos aproximar.

Respecto al último párrafo de la sección 5.5.2, ¿Qué características deberían poseer o no lo datos para decidir si es correcto el realizar una prueba unilateral sobre estos? Y así no incurrir en el aumento de las probabilidades de cometer errores de tipo 1.

Respecto a lo dicho en la página 2, donde se demuestra el cumplimiento de las condiciones para la utilización de la distribución “t” en el ejemplo. Cuando se menciona la “región aceptable”, ¿De qué manera esta región puede ser delimitada y así establecida? ¿ Existe un método en particular o esta se genera a través de la observación del gráfico?

Sobre el tamaño del efecto

Esto lleva a la noción de tamaño del efecto, que corresponde a una

cuantificación de la diferencia entre dos grupos o la diferencia real entre dos medias.

A medida que el tamaño del efecto disminuye (es decir, el estimador se acerca al valor nulo), el poder

se aproxima al nivel de significación.

Usar un valor de \_ más exigente (menor) manteniendo constante el tamaño de la muestra hace que la

curva de poder sea más baja para cualquier tamaño del efecto (lo que verifica la relación entre \_ y \_).

Usar una muestra más grande aumenta el poder de la prueba para cualquier tamaño del efecto distinto

de 0.

En ella se evidencia claramente la ventaja de las pruebas unilaterales: cuando el tamaño del efecto aumenta en el sentido de la hipótesis alternativa, el poder es mayor que para una prueba bilateral.

En ella se aprecia que la gran desventaja de las pruebas

unilaterales es que el poder tiende a cero a medida que el tamaño del efecto aumenta en sentido contrario a la hipótesis alternativa, por lo que no sería posible detectar una diferencia en el sentido opuesto aunque fuese muy grande (pues no hay una región de rechazo en dicho sentido).

El poder de la prueba aumenta mientras mayor es el tamaño del efecto.

El tamaño del efecto es el nombre asignado a una familia de índices cuyo objetivo es medir la magnitud del efecto estudiado, en nuestro ejemplo, la magnitud (estandarizada) de la diferencia entre la media del grupo control y la del grupo experimental.

Importancia

Una medida que ayuda a cuantificar y comprender los resultados de una prueba de hipótesis es el tamaño del efecto, este complementa la significación. El tamaño del efecto es la magnitud del resultado, que nos permite ofrecer una estimación del alcance de nuestros hallazgos. En estadística, el tamaño del efecto se refiere a una forma de cuantificar el tamaño de la diferencia entre dos grupos. Es relativamente sencillo de calcular y comprender, además puede ser aplicado a cualquier resultado medido en las ciencias sociales. Este es especialmente valioso al cuantificar la efectividad de una intervención, relativa a alguna comparación.

4. ¿Por qué se necesita conocer el tamaño del efecto?

El tamaño de efecto permite comprender y cuantificar los resultados de una prueba de hipótesis (en complementación con la significación α), es la medida que representa la magnitud del resultado y así estimar el alcance de los hallazgos realizados. Siendo este una forma de cuantificar el tamaño de la diferencia entre dos grupos o la diferencia real entre dos medidas tiene un gran valor dado que mide la efectividad de una intervención , referente a alguna comparación en diferentes situaciones. Junto con esto, si bien el p-valor podrá informar respecto a la dirección de un efecto, la estimación del tamaño podrá entregar la medida de este efecto (tamaño).

Ni el valor de p ni la significación estadística miden el tamaño de un efecto o la importancia de un resultado.

1. ¿En qué condiciones la distribución muestral de una proporción tiene comportamiento aproximadamente

normal? ¿Cómo se calcula la desviación estándar de esta distribución muestral?

por lo que se cumple la condición de éxito-fracaso. En consecuencia, la distribución muestral de ^p

se acerca a la normal.

1. ¿En qué condiciones la distribución muestral de una proporción tiene comportamiento aproximadamente normal? ¿Cómo se calcula la desviación estándar de esta distribución muestral?

Considerando el estimador puntual correspondiente a la proporción de éxito de la muestra, denotado por ṗ . Este estimador se distribuye de manera cercana a la normal si se cumplen las siguientes condiciones:

1. Las observaciones de la muestra son independientes.

2. Se cumple la denominada condición de éxito-fracaso, que establece que se espera observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos. Matemáticamente, np ≥10 y n(1 - p) ≥ 10.

Si la distribución muestral de ṗ cumple con las condiciones anteriores, entonces es cercana a la normalidad con media μ= p, **desviación estándar σ =√(p(1-p))**  y error estándar dado por la ecuación:    Seṗ =√(((1-p))/n)

-------------------------------------------

Considerando el estimador puntual correspondiente a la proporción de éxito de la muestra, denotado por ṗ . Este estimador se distribuye de manera cercana a la normal si se cumplen las siguientes condiciones:



1. Las observaciones de la muestra son independientes.

2. Se cumple la denominada condición de éxito-fracaso, que establece que se espera observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos. Matemáticamente,

np ≥10 y n(1 - p) ≥ 10.

Si la distribución muestral de ṗ cumple con las condiciones anteriores, entonces es cercana a la normalidad con media μ= p, desviación estándar σ y error estándar dado por la ecuación :



En primer lugar, como señalas, las condiciones que se deben verificar para emplear ambos modelos son las mismas, por lo tanto, para saber cuál de las dos distribuciones se deben utilizar, debemos tener en consideración las características de ambas.

En el caso de la distribución normal , esta si bien solo trabaja con muestras grandes (mayores a 30), nos da valores “p” exactos, a diferencia del caso de la distribución t, la cual si bien trabaja con muestras de tamaños pequeños ( menos de 30 datos, no limitado por este número necesariamente) , razón por la cual posee una gran ventaja, necesita de mayor evidencia para rechazar la hipótesis nula de investigación , dado que utiliza únicamente los grados de libertad. Cabe señalar que a mayor numero de grados de libertad, más próxima será la curva a la generada por la distribución normal (que posee valor p exactos).

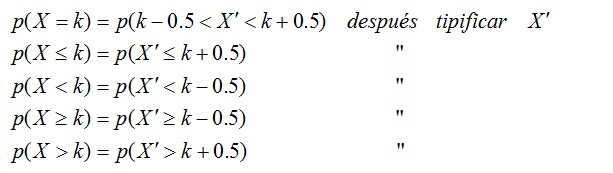
Por esta razón , se consideraría recomendable el aplicar la distribución normal, en el caso de que el tamaño de la muestra sea mayor (>30), aún cuando es posible realizarlo a través de la distribución t , la cual se aplicará con tamaños de muestras con menos de 30 datos.

En estadística un factor de corrección se entiende con una constante que ajusta un valor resultante de un cálculo determinado, son comúnmente usados para adecuar indicadores que sufren alteraciones debido a variables no consideradas inicialmente durante la recolección de la información o selección de la muestra. El factor de corrección de continuidad es el ajuste de media unidad de medida para mejorar la exactitud cuándo a una distribución discreta se le aplica una distribución continua.

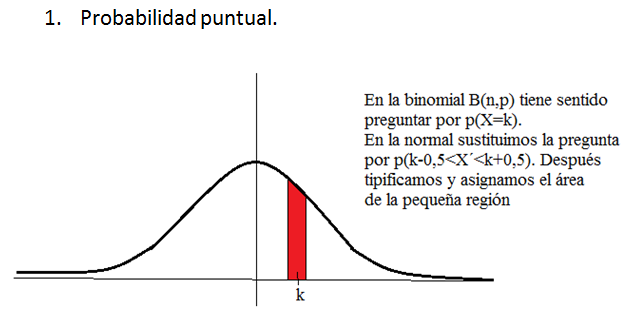
Para aplica el factor de continuidad de Yates, se obtiene la formula mostrada:

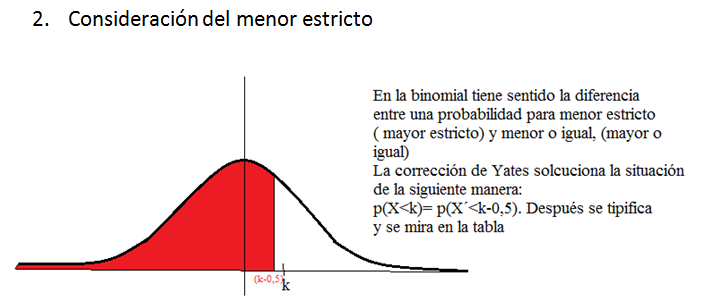


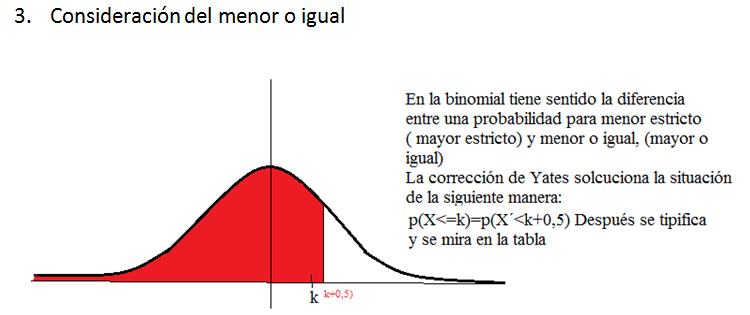
Obteniendo los siguientes casos

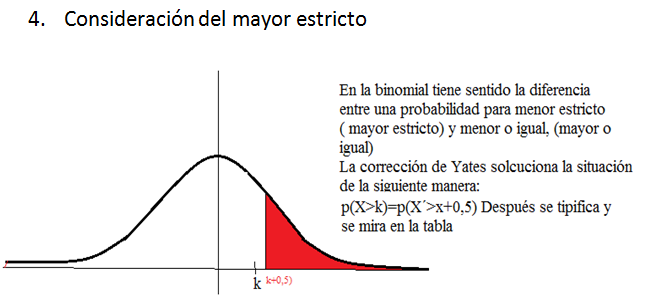


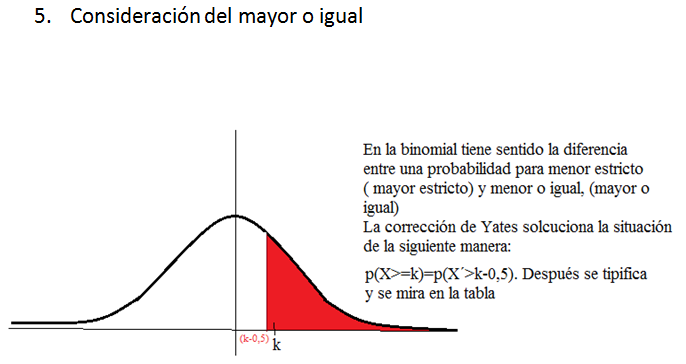
Donde:











10-

4. ¿Qué significa que un procedimiento ANOVA sea de una vía?

El análisis de varianza (ANOVA) de una vía se utiliza para determinar si existen diferencias estadísticamente significativas entre las medias de tres o más grupos.

Los procedimientos ANOVA para muestras independientes y muestras correlacionadas corresponden a análisis de varianza de una vía, pues solo consideran una única variable independiente (de tipo categórica, un factor) cuyos niveles definen los grupos que se están comparando.

Una vía significa que tenemos una única variable explicativa o predictor, también llamada variable independiente. Esta variable debe tener tres o más niveles o categorías.

Por ejemplo, si queremos analizar el pH de distintas muestras de jabón de bebé, la marca del jabón es nuestra variable independiente .

6. ¿Qué sería ANOVA de dos vías?

La prueba ANOVA bidireccional es una prueba estadística que se utiliza para determinar el efecto de dos variables precursoras nominales sobre una variable de rendimiento continua.

Un ANOVA de dos factores prueba el efecto de dos variables independientes sobre una variable dependiente. Una prueba ANOVA de dos vías analiza el efecto de las variables independientes sobre el resultado esperado, así como su relación con el resultado en sí. Se consideraría que los factores aleatorios no tienen un impacto estadístico en un conjunto de datos, mientras que se consideraría que los factores sistémicos tienen importancia estadística. Permite examinar simultáneamente los efectos de dos variables independientes e, incluso, determinar si ambas interactúan.