**CAPITULO 1. Introducción**

* 1. inferencia y modelos estadísticos

inferencia -> acción y efecto de inferir

inferir -> deducir algo o sacarlo como conclusión de otra cosa

estadística inferencial -> rama de la estadística que busca obtener una conclusión para un subconjunto de individuos o elementos (denominado **población**) a partir de información recolectada de un subconjunto de este (llamado **muestra**)

**modelo estadístico ->** modelo matemático de la regularidad estadística de los posibles resultados de la evolución de un fenómeno aleatorio

distribución de probabilidad construida para pider hacer inferencias o tomar decisiones desde datos

descripción simple de un proceso probabilístico que puede haber dado origen a un conjunto de datos observados

modelo estocastico que contiene parámetros desconocidos que deben ser estimados en base a suposiciones acerca del modelo y los datos

* 1. VARIABLES, PARAMETROS Y ESTADISTICOS

Tabla

Descripción generada automáticamente

Cada columna de la matriz representa una **variable o característica** , mientras que cada fila corresponde a una **unidad de observación** o instancia. Cada fila almacena datos de la misma persona. Usar matrices de datos para almacenar datos es muy conveniente, pues nos ayuda a acceder a los datos y modificarlos mas fácilmente.

Es importante saber que representa cada variable y que rango de valores puede tomar

Texto

Descripción generada automáticamente

No todas las variables pueden tomar el mismo valor, por lo que aparecen los **tipos de variables:**

* Numéricas: pueden tomar valores numéricos y son sensibles a operaciones aritméticas, están se pueden separar en:
  + Continuas: pueden tomar cualquier valor (en un intervalo) del conjunto de los reales. Ejemplo estatura y servicio
  + Discretas: no es posible que tomen cualquier valor (en un intervalo). Ejemplo podrían tomar solo valores enteros positivos como la variable antigüedad.
* Categóricas: solo pueden tomar un valor de entre un conjunto acotado. Cada posible valor se denomina **nivel** . entre las variables categóricas es posible distinguir variables:
  + Nominales: no existe un orden natural entre los niveles. Ejemplo genero y rama
  + Ordinales: existe un orden natural de los niveles. Por ejemplo la idea de jerarquía es evidente al distinguir entre oficiales y suboficiales en la variable escalafón.

Tener diferentes tipos de variables significa que debemos medirlas con distintas escalas:

* Escala nominal: sirve solo para separar un subconjunto de elementos en subclases excluyentes entre si. Los valores no son mas que nombres o estados, por lo que no podemos hacer operaciones aritméticas ni podemos establecer relaciones de orden.
* Escala ordinal o de rangos: al igual que la nominal, permite separar un conjunto de elementos en subclases excluyentes entre sí. Una vez más, los valores son solo nombres o estados, tampoco se pueden hacer operaciones aritméticas. Pero en este caso si se puede establecer una relación de orden, para ello la variable debe tener 3 niveles.
* Escala de intervalo: sirve para datos continuos o discretos con una gran cantidad de niveles. Además de la noción de orden de la escala ordinal, se cumple la distancia entre dos valores cualesquiera de la escala es conocida y constante, por lo que podemos emplear operaciones aritméticas.
* Escala de razón: cumple con todos los atributos de la escala de intervalos, pero además tiene su origen en un cero verdadero.

Dos variables pueden ser:

1. Independientes: no existe asociación o relación entre las variables
2. Dependientes: existe una asociación o relación entre las variables, puede existir:
   1. Asociación positiva: si una variable crece, la otra también lo hace
   2. Asociación negativa: si una variable crece, la otra decrece

En el contexto de estadística decimos:

* Parámetro: cualquier numero que describa una población en forma resumida, como por ejemplo la media poblacional.
* Estadístico: es cualquier cantidad cuyo valor puede ser calculado a partir de datos muestrales. Por ejemplo la media, la mediana o la desviación estándar de un conjunto de datos observados.

Ambos conceptos parecen similares, en realidad existe una gran diferencia. El **parámetro**  describe una población, mientras que el **estadístico** , al ser calculado a partir de una muestra, no es mas que una **estimación puntual**  del parámetro.

* 1. CONOCIENDO R
     1. importación de datos

debemos conocer en primera instancia como importar datos o cargar una matriz de datos denominados *data frame* desde un archivo de texto plano .(txt) o de valores separados por comas (.csv)

* + 1. Importación de paquetes

Si bien el entorno básico de R incluye muchas funcionalidades, existe una enorme variedad de paquetes o colecciones que incorporan otras nuevas ya existentes.

* + 1. construcción de una matriz de datos

se desea construir una matriz de datos que contenga el nombre, la fecha de nacimiento y las calificaciones de estudiantes de 3 evaluaciones para una asignatura.

* + 1. modificación de una matriz de datos

el operador %>% llamado pipe y definido en el paquete **magrittr**, cuya función es entregar un valor o el resultado de una expresión a la siguiente llamada a una función. En términos sencillos, la expresión x%>% f es equivalente a f(x) y su utilidad es que simplifica la lectura de llamadas a funciones anidadas.

* + 1. formulas

**CAPITULO 2. EXPLORACION DE DATOS**

En este capitulo revisaremos las principales estadísticas descriptivas que nos ayudaran a resumir los datos para entenderlos mejor, así como diversos tipos de gráficos que nos permitirán representar los datos de modo que podamos comprenderlos de forma visual.

2.1 Estadísticas descriptivas

Las estadísticas descriptivas son medidas que nos permiten sintetizar y como su nombre lo indica, describir los datos. Estas pueden aplicarse tanto a una muestra como a una población. Cuando una de estas medidas se aplica a la muestra, corresponde a un **estimador puntual** de la misma medida para la población. Al ser una estimación, no es exacta, aunque la precisión tiende a aumentar mientras mayor será el tamaño de la muestra.

Conceto de **distribución,** se considera la distribución de frecuencia que representa cuantas veces aparece cada valor para una variable en un conjunto de datos.

2.1.1 Estadísticas descriptivas para datos numéricos

Una estadística descriptiva mas empleadas en la **MEDIA** conocida como media aritmética o promedio. Denotamos la **media muestral**  donde x corresponde al nombre de la variable y **media población**



Diagrama, Texto, Esquemático

Descripción generada automáticamente

Podemos entender la media como el punto de equilibro de la distribución. Así, la media corresponde a una **medida de tendencia central**



La función mean() devuelve NA (no disponible) si existen valores faltantes en los datos de entrada.

Una medida de tendencia central alternativa a la media es la **mediada**, que es simplemente, el valor central d ellos valores previamente ordenados. Cuando no existe un valor central, vale decir, cuando el tamaño de la muestra es par, la mediana esta dada por el promedio simple de los dos valores centrales. En R la mediana se calcula con la función median()

La **moda** es, simplemente, el valor mas frecuente en el conjunto de datos. No obstante, existe el problema de que pueden haber múltiples modas. Dependiendo de la cantidad de modas se habla de distribuciones unimodales, bimodales y multimodales. En R no existe una función nativa para encontrar la moda, pero el paquete modeest ofrece la función mfv() que entrega el valor mas frecuente de una variable.

Las medidas que hemos estudiado hasta ahora buscan describir el centro del conjunto de datos. No obstante, también es importante conocer su **variabilidad o dispersión** , pues así se puede saber que tan semejantes o diferentes son las observaciones entre si. Esto suele calcularse en base a la **desviación** de las observaciones que se entiende como la distancia entre una observación y la media del conjunto de datos. Las dos principales medidas de dispersión son la **VARIANZA y la DESVIACION ESTANDAR.**

**VARIANZA MUESTRAL:** donde xi son los valores de cada una de las n observaciones

Un reloj de aguja

Descripción generada automáticamente con confianza media

**DESVIACION ESTANDAR de la muestra**: se define como la raíz cuadrada de la varianza, medida que resulta de gran utilidad cuando se necesita saber cuando cercanos son los datos a la media, ya que, se encuentra en la misma escala que la variable.

Un reloj en el medio

Descripción generada automáticamente con confianza media

Al igual que en el caso de la media, podemos usar las formular anteriores para obtener estimaciones puntuales de la varianza y la desviación estándar de la población, denotadas



Las funciones en R para calcular la varianza y la desviación estándar son, respectivamente var() y sd().

Menos empleado esta el **rango** muestra los valores extremos, es decir el mínimo y máximo de una variable. R ofrece la función range() .

A veces es necesario dividir el conjunto de datos en segmentos mas pequeños, por ejemplo 4,10,100 partes con igual cantidad de elementos. Cada fragmento del conjunto de datos dividido de esta forma recibe el nombre de **cuantil.** Algunas divisiones reciben nombre especial:

* Percentiles: divide en 100 subconjuntos
* Deciles : divide en 10 subconjuntos
* Quintiles: divide en 5 subconjuntos
* Cuartiles: divide en 4 subconjuntos

Los cuantiles (al igual que las otras subdivisiones mencionadas) se nombran de forma ascendente según el sentido de crecimiento del conjunto de datos. Así, el percentil 1 contiene a los valores mas pequeños. Hay que destacar que la mediana corresponde al percentil 50. R otorga la función quantile() para calcular cuantiles.

Nueva medida de variabilidad que usaremos, llamada **rango intercuantil o IRQ**

Imagen que contiene Texto

Descripción generada automáticamente

Donde Q1 y Q3 corresponde a los cuartiles 1 y 3 respectivamente. Al igual que la varianza y desviación estándar, mientras mas disperso sea el conjunto de datos, mayor será el valor de IQR En R la función que calcula el estimador es IRQ().

En algunas ocasiones el conjunto de datos contienen lo que se conoce como **valores atípicos o outliers.** Corresponden a observaciones que parecen estar fuera de rango o ser muy extremos con respecto al resto de los datos. La media o la desviación estándar son muy sensibles a estos valores atípicos.

A veces necesitaremos calcular varias medidas de tendencia central y de dispersión descritas en el apartado anterior. Por eso R y sus paquetes nos ofrecen funciones que calculan varios de estos estadísticos con una sola llamada por ejemplo la función summary() que entrega la media, la mediana, el primer y el tercer cuartil, el mínimo y máximo. O también summarise() del paquete dplyr para calcular varias medidas en una sola llamada.

2.1.2 ESTADISTICAS DESCRIPTIVAS PARA DATOS CATEGORICOS

Como primer estadísticos para variables categóricas podemos mencionar la **frecuencia** que corresponde a la cantidad de veces que podemos encontrar cada nivel de la variable en los datos. Otro estadístico importante corresponde a la **proporción** que corresponde a la frecuencia relativa.

La mejor alternativa para este tipo de datos es la tabla de contingencia , también llamada matriz de confusión o tabla de frecuencias donde cada fila representa la cantidad de veces en que ocurre una combinación de variables.

En R se pueden construir de forma sencilla, la función table() y xtabs() funciona equivalentemente , aunque xtabs muestra el nombre de la variable tabulada al imprimir los resultados y table no. Estas tablas no incluyen los totales por filas, pero la función marginSums() permite calcularlos y mostrarlos como un vector a su vez, la función addmargins() permite calcular dichos totales e incorporarlos en la tabla.

2.1.3 trabajando con datos agrupados

Nos veremos en la necesidad de obtener estadísticas descriptivas de una variable separando las observaciones en grupos de acuerdo a una variable categórica. Para ello el paquete **dplyr** ofrece la función group\_by(), que podemos usar en conjunto con summarise().

2.2 REPRESENTACION GRAFICA DE DATOS

Revisaremos diferentes tipos de gráficos que resultan utlies al momento de estudiar un conjunto de datos disponibles, considerando su definición, su utilidad y como se construyen en R. En R utilizaremos el paquete **ggpubr** algunos parámetros importantes son:

Texto

Descripción generada automáticamente

2.2.1 una variable numérica

El HISTROGRAMA resulta muy útil si queremos representar una única variable numérica y la muestra es grande. Podemos decir que este grafico muestra una aproximación a la densidad (o distribución de frecuencias) para la variable, para lo que tenemos que dividir el rango de valores posibles en intervalos (generalmente iguales) y luego contar la cantidad de observaciones de cada intervalo.

Gráfico, Histograma

Descripción generada automáticamente

A medida que avancemos veremos que es de suma importancia conocer la **distribución de frecuencias** de una variable. Podemos ver en la imagen b que la frecuencia es mayor para potencias mas bajas pues las barras de la izquierda son mas altas que las de la derecha. Podría decirse que las observaciones se concentran a la izquierda y que hay una cola que se prolonga hacia la derecha. Cuando esto ocurre decimos que la distribución esta **desviada a la izquierda** o que hay **asimetría negativa.**  El caso contrario podría darse cuando la distribución este desviada a la derecha que presenta asimetría positiva. En el caso del histograma a es mas simétrico, pues las observaciones se aglomeran hacia el centro y hay colas tanto a la izquierda como a la derecha.

El grafico de CAJA. Es muy útil pues su construcción considera 5 estadísticos para representar el conjunto de datos y además facilita la identificación de datos atípicos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Los extremos inferior y superior del rectángulo o caja corresponden al primer y tercer cuartil, mientras la línea horizontal al interior de la caja denota la mediana. Así la caja engloba el 50% central de los datos y su altura corresponde al rango intercuartil. Las barras que se exitenden por sobre y por debajo de la caja , llamadas bigotes, capturan aquellos datos fuera de la caja central y que están situados a no mas de 1.5 veces el IQR. Cualquier observación que este mas allá de la caja y de los bigotes podría tratarse de una observación atípica.

2.2.2 UNA VARIABLE CATEGORICA

Si queremos representar una única variable categórica, lo mas adecuado es usar un GRAFICO DE BARRAS, pues cada barra es tan larga como la proporción de valores en cada nivel de la variable.

Otra variable para representar una única variable categórica es el GRAFICO DE TORTA, que se presenta a continuación.

Gráfico, Gráfico circular

Descripción generada automáticamente

2.2.3 DOS VARIABLES NUMERICAS

Los GRAFICOS DE DISPERSION son adecuados en este caso. Se caracterizan porque muestran información caso a caso, ya que cada punto del grafico corresponde a una observación. Por ejemplo muestra este tipo de grafico para las variables rendimiento y peso

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Los gráficos de dispersión también son muy útiles para identificar si dos o mas variables están relacionadas.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Estos tres gráficos de dispersión diferentes:

El de la izquierda se aprecia que la variable peso y cuarto\_milla son independientes, pues no hay una tendencia definida en la organización de los puntos.

El del centro en cambio podemos ver que la potencia tiende a aumentar a medida que tambien lo hace el peso, por lo que ambas variables están positivamente asociadas.

El de la derecha nos muestra las variables peso y rendimiento presentan asociación negativa, puesto que a medida que la primera aumenta, la segunda disminuye.

Similar al grafico de barras para una variable categórica los GRAFICOS DE BARRAS APILADAS, AGRUPADAS Y ESTANDARIZADAS permiten visualizar la matriz de confusión entre dos variables y encontrar posibles relaciones entre ellas.

Gráfico, Gráfico de barras

Descripción generada automáticamente

El grafico de barras apiladas (izquierda, muestra tres barras cuya altura corresponde a la frecuencia de la cantidad de cambios, al igual que la figura 2.4, pero ahora cada barra esta subdividida en secciones de distinto color para cada tipo de motor.la altura de cada sección esta dada por la frecuencia del tipo de motor para la cantidad de cambios representada en la barra)

El grafico de la derecha, corresponde al grafico de barras estandarizadas, muestra barras de igual altura para cantidad de cambios presentado claramentos los cambios en la proporción de cada tipo de motor por la cantidad de cambios. Se puede apreciar que los automóviles con 3 y 5 cambios tienen mayoritariamente motores en forma de v, ambas en igual proporción, mientras que el uso de motores rectos se da principalmente en automóviles de 4 cambios.

El grafico de barras agrupadas, al centro en la figura 2.8 es equivalente al de la izquierda, pero en lugar de dividir una barra en segmentos, muestra barras contiguas para cada tipo de motor.

Similar al grafico de barras para dos variables, el GRAFICO DE MOSAICO permite representar una tabla de contingencia, para ello divide un área en regiones y el área de cada región es proporcional al porcentaje de observaciones que representa. Usando las variables cambios y motor, el ancho de cada columna es proporcional a la cantidad de automóviles que tienen la correspondiente cantidad de cambios, mientras que la altura de cada barra de las columnas refleja que la proporción de automóviles con un determinado tipo de motor.

CAPITULO 3. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

Se define como **variable aleatoria** una variable o un proceso cuyo resultado sea numérico. Tienen una **distribución de probabilidad**, la cual define la probabilidad de que ocurran los diferentes valores que dicha variable puede tomar

3.1 variables aleatorias

La definición de **variable aleatoria continua** es sencilla, es una variable que puede tomar cualquiera de los infinitos valores posibles dentro de un intervalo.

Una **variable aleatoria discreta**, solo puede tomar un conjunto finito de valores.

Un ejemplo típico de variable aleatoria es el lanzamiento de un dado. Si esta bien balanceado, tendremos igual probabilidad de obtener cualquiera de las caras.

El valor esperado, denotado como E(X) o µ, corresponde al resultado promedio de una variable aleatoria. Para una variable aleatoria discreta, se calcula sumando los valores posibles ponderados por su probabilidad, como muestra la imagen:

Texto, Esquemático

Descripción generada automáticamente

También podemos calcular que tan alejado podría estar el valor obtenido del valor esperado por media de la varianza general, denotado por Var(X) o Ϭ2 , que se calcula como la media de los cuadrados de la diferencia con respecto a la media ponderada según la probabilidad de ocurrencia, como muestra la ecuación una vez más, la desviación estándar corresponde a la raíz cuadrada de la varianza

Un dibujo de una persona

Descripción generada automáticamente con confianza baja

En R, el paquete DiscreteRV permite trabajar con variables aleatorias discretas.

Para ayudarnos a entender mejor la noción de distribución de probabilidad, veamos la imagen, ella nos muestra de izquierda a derecha, las distribuciones de probabilidad para el puntaje total obtenido al lanzar 5,10 y 20 dados respectivamente.

Gráfico, Histograma

Descripción generada automáticamente

Conocer la distribución de probabilidad de una variable discreta nos ayuda hacer estimaciones útiles.

Combinación lineal de las variables X,Y,Z , la formula general de la combinación lineal de n variables esta dada por la ecuación , donde cada Xi corresponde a una variable aleatoria y cada ci es una constante conocida.

Texto

Descripción generada automáticamente

Cuando las variables de una combinación lineal son independientes, podemos calcular el valor esperado y la varianza de la combinación lineal usando las ecuaciones 3.5 y 3.6. una vez más la desviación estándar esta dada por la raíz cuadrada de la varianza

Texto, Carta

Descripción generada automáticamente con confianza media

En R tambien podemos trabajar con combinaciones lineales de variable aleatorias discretas, como muestra el script.

Al examinar con mayor detención los gráficos de la figura 3.1 podemos apreciar que, a medida que se efectúan mas lanzamientos del dado, el histograma se asemeja cada vez mas a una curva continua la cual recibe el nombre de **función de densidad de probabilidad** o simplemente **distribución o densidad**.

Las distribuciones tienen la propiedad de que el área total bajo la curva siempre es 1, lo que resulta muy útil al momento de calcular probabilidades, pues basta con calcular el área bajo la curva del segmento deseado.

Volviendo al ejemplo del desempeño del programa, el tiempo de ejecución es un realidad una variable continua. Así la probabilidad de que el tiempo de ejecución sea mayor a 25 segundos corresponde al área coloreada en el grafico de la figura 3.3 con un valor de 0.0482

Gráfico

Descripción generada automáticamente

3.2 distribuciones continuas

Múltiples funciones de distribución continua que se usan frecuentemente en estadística, las cuales se describen a continuación

**3.2.1 DISTRIBUCION NORMAL**

Tambien conocida como distribución gaussiana, la distribución normal es la más ampliamente empleada en estadística, pues muchas variables se acercan a esta distribución. Se caracteriza por ser unimodal y simétrica, con forma de campana.

La distribución normal se usa para modelar diversos fenómenos y podemos ajustarla mediante dos parámetros:

* µ: la media, que desplaza el centro de la curva a lo largo del eje x
* Ϭ: desviación estándar, que modifica que tan dispersos están los datos con respecto a la media

Así denotamos este tipo de distribución por N(µ,Ϭ), la figura muestra dos ejemplos superpuestos de distribución normal N(µ=0,Ϭ=1) en azul y N(µ=10,Ϭ=6) en rojo

Gráfico, Gráfico de líneas

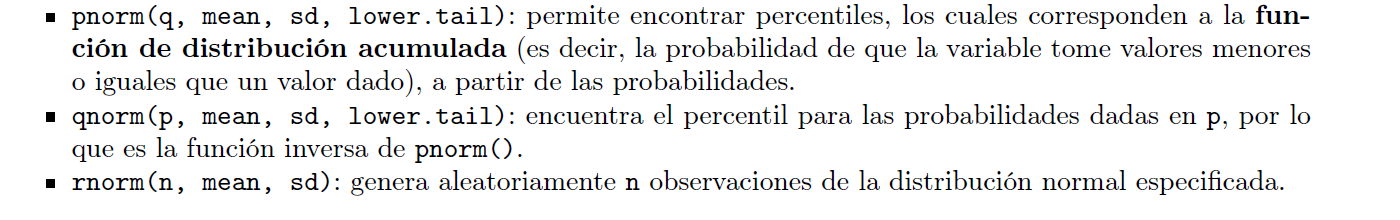
Descripción generada automáticamente

Antes de continuar, fijémonos en la líneas 8 y 16 del script

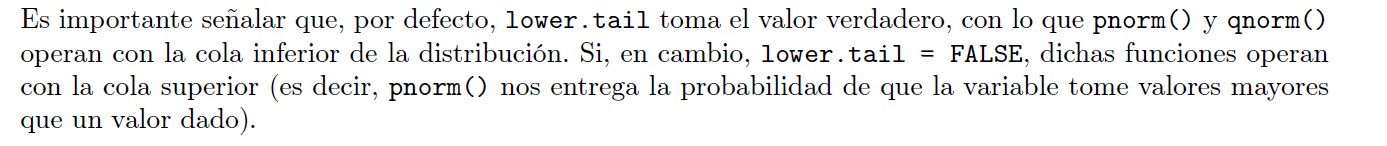




Donde se usa la función **dnorm(x,mean,sd)** esta función calcula la densidad de una distribución normal, además dnorm nos ofrece otras función que resultan de mucha ayuda:







Una regla empírica muy útil al momento de trabajar con distribuciones normales es la llamda regla 68-95-99.7

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Muchas pruebas estadísticas operan bajo el supuesto de que los datos siguen una distribución normal. La normalidad es siempre una aproximación por lo que debemos verificar que el supuesto de una distribución normal sea aceptable.

Una buena opción para ellos es el **grafico cuantil-cuantil** o también llamado **grafico Q-Q.** En este grafico se puede distinguir un grupo de puntos, una recta y una región coloreada. Los puntos corresponden a las observaciones, mientras que la recta representa la distribución normal. En consecuencia mientras mas se asemeje el patron que forman los puntos a la recta, mas parecida será la distribución a la normal. La banda coloreada establece el margen aceptable para suponer normalidad en el conjunto de datos. Así para el conjunto de datos de la figura 3.6 seria imprudente aceptar el supuesto de normalidad.

Gráfico, Gráfico de líneas, Gráfico de dispersión

Descripción generada automáticamente

**3.2.2 DISTRIBUCION Z**

al trabajar con distribuciones, especialmente las simétricas, a menudo usaremos **técnicas de estandarización** para determinar que tan usual o inusual es un determinado valor en una escala única. Así para la distribución normal usamos como estandirazacion la **distribución Z o distribución normal estándar** que no es mas que una distribución normal centrada en 0 y con desviación estándar 1 que se obtiene de la ecuación:

Imagen que contiene objeto, reloj

Descripción generada automáticamente

Al aplicar la ecuación 3.7 a una observación x en una distribución normal obtenemos , entonces, su **valor z,**  que determina cuan por encima o por debajo de la media (en términos de desviación estándar) se encuentra dicha observación x. así observación cuyos valores z sean negativos estarán por debajo de la media, análogamente un valor z positivo indica que la observación esta por sobre la media. Mientras mayor sea el valor absoluto de su valor z, más inusual será la observación.

**3.2.3 DISTRIBUCION CHI-CUADRADO**

tambien conocida como ji-cuadrado o *X2 ,* se usa para caracterizar valores siempre positivos y habitualmente desviados a la derecha. El único parámetro de esta distribución corresponde a los **grados de libertad**, usualmente representada por la letra griega Ѵ, que son una estimación de la cantidad de observaciones empleadas para calcular un estimador. Otra forma de entender esta idea es como la cantidad que pueden cambiar libremente en un conjunto de datos.

Ejemplo: se necesita una muestra de tres elementos cuya media sea 10. Una vez escogidos los primeros 2, solo queda una posibilidad para el tercero de modo que se cumpla la media deseada. **Así solo los dos primeros pueden cambiar libremente, por lo que tienen dos grados de libertar.**

Esta distribución esta relacionada con la ya conocida distribución Z, pues si sumamos los cuadrados de K variables aleatorias independientes que siguen una distribución Z, dicha suma sigue una distribución *X2* con k grados de libertar

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**3.2.4 DISTRIBUCION t de Student**

Ampliamente empleada cuando se trabaja con muestras pequeñas, la distribución t de student o simplemente distribución t, tiene al igual que la distribución X2 los grados de libertad como único parámetro, a medida que los grados de libertar aumentan, esta distribución se asemeja cada vez mas a la normal, aunque sus colas son mas gruesas.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

La distribución t se encuentra relacionada con las distribuciones vistas anteriormente de acuerdo a la ecuación 3.9 donde Z es una distribución normal estándar y X2(V) es una distribución X2 con V grados de libertad

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza media



**3.2.5 Distribución F**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

3.3 DISTRIBUCIONES DISCRETAS

**3.3.1 DISTRIBUCION DE BERNOULLI**

Es aquella en que cada intento individual tiene solo dos resultados posibles: “éxito”, que ocurre con una probabilidad p y se representa habitualmente con un 1 y “fracaso” que ocurre con probabilidad q=1-p y suele representarse por un 0.

Definimos la proporción de la muestra para una distribución de Bernoulli p^ , como la cantidad de éxitos dividida por la cantidad de intentos. Mientras mayor sea la cantidad de intentos, mas cercano será el valor de p^ a la probabilidad real de éxito p.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

3.3.2 DISTRIBUCION GEOMETRICA

Describe la cantidad de intentos que debemos realizar hasta obtener un éxito para variables de Bernoulli **independientes e idénticamente distribuidas**  es decir que no afectan unas a otras y cada una con igual probabilidad de éxito.

La probabilidad de obtener un éxito al n-esimo intento esta dado por la ecuación 3.15 donde podemos ver que las probabilidades en esta distribución decrecen exponencialmente rápido.

Diagrama

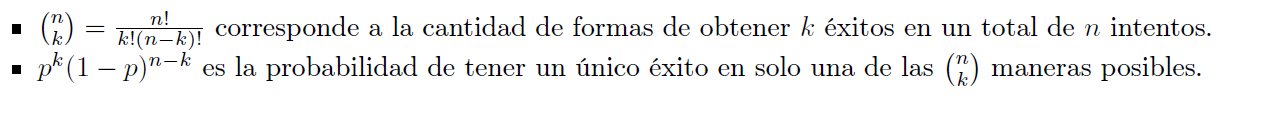
Descripción generada automáticamente con confianza baja

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

3.3.3 DISTRIBUCION BINOMIAL

Describe la probabilidad de obtener exactamente K existos en n intentos independientes de Bernoulli con probabilidad de éxito p, cuya función de probabilidad esta dada por la ecuación 3.19 donde:



Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza media

Gráfico, Histograma

Descripción generada automáticamente

3.3.4 distribucion binomial negativa

Es mas general que la binomial, pues describe la probabilidad de encontrar el k-esimo éxito al n-esimo intento. “en el caso binomial en general se tiene una cantidad fija de intentos y se considera la cantidad de éxitos. en el caso binomial negativo, se examina cuantos intentos se necesitan para observar una cantidad fija de éxitos y se requiere que la ultima observación sea un éxito”

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Texto

Descripción generada automáticamente

3.3.5 DISTRIBUCION DE POISSON

Útil para estimar la cantidad de eventos en una población grande en un lapso de tiempo dado, por ejemplo para estimas la cantidad de contagios de influenza entre los habitantes de Santiago en una semana, la **distribución de Poisson** tiene una función de probabilidad definida por la ecuación 3.25 donde monito es la tasa de o cantidad de eventos que se espera observar en un lapso de tiempo dado y k puede tomar cualquier valor entero no negativo.

Texto

Descripción generada automáticamente con confianza baja

Gráfico, Histograma

Descripción generada automáticamente

Texto

Descripción generada automáticamente

CAPITULO 4. FUNDAMENTOS PARA LA INFERENCIA

Concepto de población, entendido como el conjunto de interés y muestra que es un subconjunto de la población. Tambien se introducen las nociones de parámetro, correspondiente aun valor que resume la población (por ejemplo la media de la población µ) y de estadístico como un calor que resume una muestra media muestral .

La **inferencia estadística** tiene por objeto entender cuan cerca está el estadístico del parámetro real de la población. Conoceremos los principios necesarios para la inferencia estadística.

4.1 Estimadores puntuales

Parámetro y los estadísticos son valores que resumen respectivamente una población y una muestra. En consecuencia, podemos decir que un estadístico corresponde a un **estimador puntual** de un parámetro. El valor de un estimador puntual cambia dependiendo de la muestra que usemos para obtenerlo. Por mas que su valor se acerque al parámetro de la población, difícilmente será igual a este último. El estimador tiende a mejorar a medida que aumentamos el tamaño de la muestra por efecto de la **Ley de los grandes números.** Para ilustrar este fenómeno, consideremos la **media móvil** , que es una secuencia de medias muestrales en que cada una de ellas toma un elemento más de la población que su antecesora.

Imagen que contiene Gráfico

Descripción generada automáticamente

Para determinar que tan adecuado es un estimador, necesitamos saber cuanto cambia de una muestra a otra. Si esta variabilidad es pequeña, es muy probable que la estimación sea buena. Podemos estudiar la variabilidad de la muestra con ayuda de la **distribución muestral,** que representa la distribución de estimadores puntuales obtenidos con **todas** las diferentes muestras de igual tamaño de la población. La figura 4.2 representa las medias para diferentes muestras de una población, aunque solo una selección aleatoria de todas las posibles muestras, incluyendo además una línea vertical roja que señala la media de la población. Podemos destacar que las medias muestrales tienden a aglutinarse en torno a la media poblacional, pues de acuerdo al teorema del limite central la distribución de  se aproxima a la normalidad. Esta aproximación mejora a medida que aumenta el tamaño de la muestra.

Gráfico, Histograma

Descripción generada automáticamente

**4.2 MODELOS ESTADISTICOS**

Ahora que ya conocimos mas conceptos, podemos definir con precisión que es un **modelo estadístico.**  En el capitulo 1 dijimos que un modelo es simplemente una representación y que los modelos estadísticos pueden emplearse para diferentes propósitos entre ellos:

* Describir o resumir datos
* Clasificar objetos o predecir resultados
* Anticipar los resultados de intervenciones (en ocasiones)

Mas formalmente, un modelo estadístico es una descripción de un **proceso probabilístico** con **parámetros desconocidos** que deben ser **estimados** en base a **suposiciones** y un conjunto de datos **observados**. Tiene la forma:

Logotipo

Descripción generada automáticamente con confianza baja

Texto

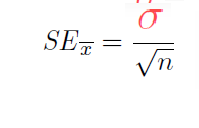
Descripción generada automáticamente

El error de la ecuación se relaciona entonces con la calidad del modelo. Mientras menor sea el error, mejor será el modelo. Por el contrario, un error grande es señal de un modelo fallido, que no describe bien los datos, no ayuda a predecirlos bien, o no ayuda a su correcta clasificación.

La media y la proporción y cualquier estadístico en general, son en si mismos, modelos estadísticos aunque bastante simples.

4.3 Error estándar

En el capitulo 2 conocimos la desviación estándar como la medida que estima la distancia de las observaciones respecto de la media. El **error estándar**, denotado usualmente por  corresponde a la desviación estándar de la distribución de un estimador muestral  de un parámetro . Por ejemplo, el error estándar de la media, es decir la desviación estándar de la distribución de las medias de todas las posibles muestras de n observaciones independientes, se calcula con la ecuación 4.2:



Donde es la desviación estándar de la población y n corresponde al tamaño de la muestra. En esta ecuación queda en evidencia que el error estándar de la media disminuye a medida que el tamaño de la muestra aumenta.

Un método confiable que podemos usar para asegurar que las observaciones sean independientes es realizar un muestreo aleatorio simple que abarque menos del 10% de la población.

Ya mencionamos que la distribución de las medias muestrales tiende a ser cercana a la normal, por lo que en dicho caso es posible usar el **modelo normal**, sustentando en el teorema del limite central. Las condiciones que deben cumplirse para usar este modelo y que, en consecuencia el error estándar sea preciso son:

1. Las observaciones de la muestra son independientes
2. La muestra es grande (en general n >=30)
3. La distribución de la muestra no es significativamente asimétrica. Esto ultimo suele relacionarse con la presencia de valores atípicos. Mientras mayor sea el tamaño de la muestra, mas se puede relajar esta condición.

Si alguna de estas condiciones no se cumple, debemos considerar otras opciones: para muestras pequeñas se deben considerar métodos alternativos y si la distribución de la muestra presenta asimetría significativa entonces tendremos que incrementar el tamaño de la muestra para compensar el efecto de la desviación.

4.4 INTERVALO DE CONFIANZA

Hasta ahora sabemos que un estimador puntual es un único valor (obtenido a partir de una muestra) que como su nombre indica, estima un parámetro de la población. Por ende, dicho valor rara vez es exacto.

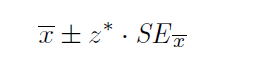
En consecuencia, lo lógico seria establecer un rango de valores plausibles para el parámetro estimado, que llamaremos **intervalo de confianza** y que se construye en torno al estimador puntual. Dado que el error estándar representa la desviación estándar asociada al estimador, tiene sentido que lo usemos de guía en este proceso.

En el capitulo 3 vimos la regla empírica para la distribución normal, la cual señala que en distribuciones normales alrededor de 95% de las veces el estimador puntual se encontrara en un rango de 2 errores estándar del parámetro. Es decir, al considerar un intervalo de confianza de dos errores estándar, tendremos 95% de **confianza de haber capturado el parámetro real.**

Imagen que contiene Gráfico

Descripción generada automáticamente

podemos generalizar la ecuación 4.3 para calcular el intervalo de confianza para la media con cualquier **nivel de confianza** como muestra la ecuación 4.4



El termino Z\* en la ecuación, corresponde usualmente al valor z tal que el área bajo la curva normal estándar comprendida entre -z\* y z\* es igual al nivel de confianza deseado. La expresión

Z\* . SE recibe el nombre de **margen de error.**

Tomemos como ejemplo un **nivel de confianza** (que, por razones que veremos en la sección siguiente denotaremos por ) de 90% es decir 1-alfa = 0.9. Eso significa, entonces que nuestro intervalo de confianza excluye el 5% del área correspondiente a la cola inferior (es decir, el percentil con valor 0.05) e igual porcentaje del área correspondiente a la cola superior (que, como la distribución Z es simétrica, es igual al área anterior ). Puesto que conocemos el percentil, (1-alfa)/2 = 0.05 en R podemos usar la llamada **qnorm(0.05, mean = 0, sd =1, lower.tail = FALSE)** y obtenemos z\* = 1.64 , es importante recordar que esta llamada estamos trabajando con la cola superior para que z\* sea positivo, si hacemos la llamada para la cola inferior, obtenemos z\* = -1.64.

Es importante destacar que debemos ser cuidadosos al interpretar un intervalo de confianza del x%(x=1-alfa). Su significado es , sencillamente, “se tiene X% de certeza que el parámetro de la población se encuentra entre” , es decir que en promedio x5 de los intervalos de confianza que se construyan con el estadístico, con muestras de un tamaño fijo capturaran el verdadero valor del parámetro. Esto NO ES EQUIVALENTE a decir que el valor del parámetro tiene una “probabilidad de x% de estar entre los valores del intervalo calculado, lo que seria incorrecto”. Por otra parte los intervalos de confianza no dicen nada acerca de observaciones individuales, sino que solo hablan del parámetro en cuestión.

**4.5 PRUEBA DE HIPOTESIS**

Nuevo sistema computacional en un banco para gestionar las transacciones. El nuevo sistema N sea puesto a prueba durante un mes con iguales condiciones de hardware en paralelo con el sistema antigua A y el banco tiene un registro del tiempo que tarde cada sistema en efectuar cada transacción. El gerente ha determinado que autorizara la migración al nuevo sistema si únicamente si este es mas rápido que el antiguo al procesar las transacciones. Se sabe que el sistema antiguo tarda µa=530 milisegundos en procesar una transacción, para el sistema nuevo se han registrados n= 1600 transacciones, realizadas en un tiempo promedio de XN= 527.9 [ms] con desviación estándar SN=48[ms].



Una primera aproximación para tomar la decisión puede ser investigar si existe diferencia en los tiempos de ejecución de ambos sistemas, lo que se puede expresarse en torno a dos hipótesis

H0: el nuevo sistema, en promedio, tarde lo mismo que el antiguo en procesar las transacciones es decir µN = µA

HA: los sistemas requieren, en promedio, cantidades de tiempo diferentes para procesar las transacciones es decir µN ≠ µA

La primera hipótesis H0, recibe el nombre de **hipótesis nula**  y suele representar una postura escéptica, que es verdadera, por lo que la **hipótesis nula siempre se formula como una igualdad.**

La segunda HA, llamada **hipótesis alternativa** representa en cambio una nueva perspectiva. Esta primera aproximación corresponde a una **prueba bilateral** o de dos colas, pues la diferencia puede ser en ambos sentidos 

Como en este caso conocemos el valor de µA = 530 [ms], tambien podríamos escribir la formulación matemática de la hipótesis

H0: µN = 530

HA: µN ≠ 530

En este caso 530 recibe el nombre de **valor nulo** , pues representa el valor del parámetro cuando se cumple la hipótesis nula.

Una aproximación mas cercana, seria investigar si el nuevo sistema es efectivamente **mas rápido** que el antiguo. En este caso se habla de una prueba unilateral o de una cola, pues solo interesa saber si el tiempo promedio empleado por el nuevo sistema es menor que el empleado por el sistema antiguo, así quedarían las nuevas hipótesis:

H0: el nuevo sistema tarda, en promedio, tarde lo mismo que el antiguo en procesar las transacciones es decir µN = µA

HA: el nuevo sistema tarda, en promedio, menos que el antiguo en procesar las transacciones es decir µN < µA

Teniendo las hipótesis planteadas, sigue decidir si la hipótesis nula parece o no plausible a través de una **prueba de hipótesis.**  El marco para la prueba de hipótesis es esceptico: no se rechaza la hipótesis nula a menos que haya suficiente evidencia para rechazarla en favor de la hipótesis alternativa. Sin embargo el que no se logre rechazar H0 **no significa aceptarla** como verdadera o como correcta sin mas. Por eso se usa un lenguaje bastante peculiar, señalando que se falla al rechazar H0 o bien que se rechaza H0 en favor de HA.

Si volvemos al caso donde queremos saber si solo existen diferencias en los tiempos entre el sistema antiguo y el nuevo (prueba de hipótesis bilateral) XN = 527.9 es en efecto distinto de µA = 530 . no obstante al ser un estimación puntual como ya hemos estudiado, esta diferencia podría deberse a la muestra escogida, por lo que el parámetro real µN podría ser igual a µA. en consecuencia resulta útil calcular el intervalo de confianza para XN



Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Como el parámetro del sistema antiguo µA=530 cae apenas dentro de este intervalo, se puede suponer que no existe una diferencia significativa entre los tiempos promedio requeridos por ambos sistemas, por lo que no se rechaza H0. Así tenemos un 95% de confianza en que no existe una diferencia entre los tiempos que requieren ambos sistemas para procesar transacciones. Sin embargo , esta decisión es apresurada ya que, el resultado esta cerca del borde de rechazo y en este caso es logico investigar mas (hacer crecer la muestra).

Si volvemos al caso de la hipótesis unilateral (ver si el sistema nueva es mas rápido que el antiguo), mantiendo el nivel de confianza 1-alfa=0.95, en este caso debemos considerar los valores menores a µA=530 para el calculo de z\*. en otras palabras el 5% que descartaremos corresponde únicamente a la cola superior. Así nuestro valor de z\* esta dado por la llamada

**Qnorm(0.05, mean=0, sd=1, lower.tail = FALSE),** obteniéndose z\* = 1.64 por lo que se tiene que la cota superior es:

Imagen que contiene Gráfico

Descripción generada automáticamente

Luego el intervalo de confianza va desde cualquier valor bajo la media observada hasta el valor calculado arriba, por lo que el intervalo con 95% de confianza seria [-infinito;529,874]

Ahora el valor de µA=530 cae apenas fuera del intervalo y podemos decir que existe evidencia de que el nuevo sistema tarda en promedio menos tiempo que el antiguo en procesar las transacciones.

Siempre que se prueban hipótesis podemos cometer un error al momento de decidir si rechazar o no la hipótesis nula, afortunadamente, la estadística ofrece herramientas para cuantificar cuan frecuentes son dichos errores. Existen cuatro posibles escenarios, los cuales se presentan a continuación:

ERROR DE TIPO 1: rechazar H0 cuand0 en realidad es verdadera

ERROR de TIPO 2: no rechazar H0 cuando en realizad HA es verdadera

Texto

Descripción generada automáticamente

Como hemos señalado, la prueba de hipótesis se basa en no rechazar H0 a menos que se tenga evidencia contundente. Por regla general, no se desea cometer el error de rechazar incorrectamente la hipótesis nula (error de tipo I) en mas de 5%de los casos. Esto corresponde al **nivel de significación** de 0.05 denotado por  si unamos un intervalo de confianza de 95% para evaluar una prueba de hipótesis en que la hipótesis nula es verdadera, cometeremos un error de tipo 1 cada vez que el estimador puntual este a 1.96 o mas errores estándar del parámetro de la población. Esto puede ocurrir un 5% de las veces (2.5% en cada cola de la distribución para el caso bilateral). Del mismo modo, un intervalo de confianza del 99% es equivalente a un nivel de significación de .

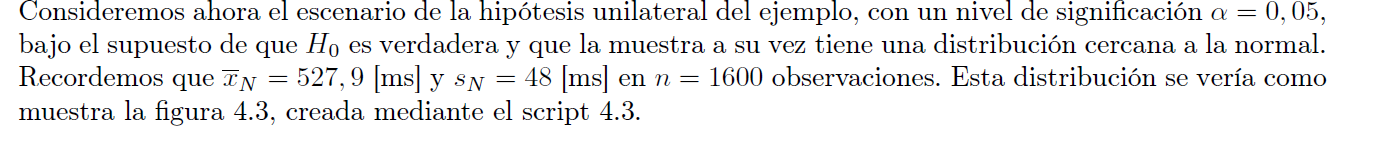


El intervalo de confianza es de mucha ayuda para decidir si rechazar o no H0. No obstante, no aporta información directa acerca de cuan fuerte es la evidencia para la decisión tomada

4.5.1 PRUEBA FORMAL DE HIPOTESIS CON VALORES P

Anterior mente existían dos procedimientos para decidir sobre una prueba de hipótesis, el primero es realizando el intervalo con 1-alfa de confianza de acuerdo al estadístico observado en una muestra y revisar si el valor nulo cae o no dentro del intervalo. El otro procedimiento clásico, consiste en estimar el valor z corresponde la media observada en la distribución normal estandarizada que define el valor nulo y el error estándar: si este estadístico z es mayor que z\* entonces el estadístico cae en una zona de “rechazo” de H0, en caso contrario |z|< z\* se falla en rechazar la hipótesis nula. Estos métodos siguen siendo utilies, su diseño respondia a la existencia de **tablas de probabilidad** en que se tabulaban probabilidades para algunos valores de percentiles de uso común, como 90%,95%,0975% o 0.99%

Con R es posible obtener probabilidades casi exactas para cualquier percentil. Esto hizo que un tercer método para decidir una prueba de hipótesis haya ido ganando popularidad: el uso del **valor p ,** tambien llamado **p-valor** y esta definido como la probabilidad de observar datos al menos tan favorables como la muestra actual para la hipótesis alternatica, si la hipótesis nula es verdadera. De esta forma un p-valor permite cuantificar cuan fuerte es la evidencia en contra de la hipótesis nula (en favor de la hipótesis alternativa)



Gráfico

Descripción generada automáticamente

Es importante hacer una aclaración en relación al valor p. el área bajo la sección de la curva con valores menores o iguales a un estimador se calcula usando para ello el valor z, definido en la ecuación 4.5 como estadístico de prueba.

Interfaz de usuario gráfica

Descripción generada automáticamente

Un estadístico de prueba es un estadístico de resumen que resulta especialmente útil para evaluar hipótesis o calcular el valor p. el valor z se usa cuando el estimado puntual se acerca a la normalidad, aunque existen otros estadísticos de prueba adecuados para otros escenarios.

El valor de p en este caso p=0.040 corresponde al área coloreada en la figura 4.3 y se calcula en la línea 51 del script, esto nos indica que si H0 fuera verdadera y el nuevo sistema tarde en promedio lo mismo que el antiguo en procesar las transacciones, la probabilidad de encontrar una media de a lo mas 527,9 ms para una muestra de 1600 transacciones es de 4% lo que seria bastante poco frecuente.

Cuanto menor sea el valor p, mas fuerte será la evidencia en favor de HA por sobre H0. Y aquí la ventaja de usar este método para decidir: el valor p se puede **comparar directamente con el nivel de significación alfa** y si p es menor que el nivel de significación se considera evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa. En este ejemplo p=0.040 < alfa=0.05, por lo que se falla al rechazar H0 en favor de HA. Como se dijo cuando usamos intervalos de confianza, el valor p esta cerca del valor alfa y convendría ser menos tajante en la decisión y evaluar la posibilidad de ampliar la muestra para conseguir evidencia mas definitiva.

Formular la conclusión en lenguaje llano. Concluimos que los datos sugieren que el nuevo sistema tarda menos que el antiguo en procesar transacciones, pero que es necesario hacer un estudio con mas observaciones para tener un mejor diagnostico.

Volviendo al primer escenario con un alfa de 0.05 , en este caso nos interesan las dos colas de la curva normal. Dado que el modelo es simétrico, el área bajo ambas colas es la misma. El valor p, entonces, ahora es igual a dos veces el área de la cola inferior es decir p=0.080. puesto que p > alfa se falla en rechazar H0. Es decir no hay evidencia suficiente para concluir que existe una diferencia entre los tiempos promedio requeridos por ambos sistemas para procesas las transacciones

Gráfico, Gráfico de líneas

Descripción generada automáticamente

4.5.2 EL EFECTO DEL NIVEL DE SIGNIFICACION

Texto

Descripción generada automáticamente con confianza media

4.6 INFERENCIA PARA OTROS ESTIMADORES

Hasta ahora solo hemos considerado la media como estimador para la inferencia. No obstante, muchos de los conceptos que hemos visto en este capitulo pueden aplicarse, con algunas ligeras modificaciones usando otros estimadores.

4.6.1 estimadores puntuales con distribución cercana a la normal

Existen múltiples de estimadores puntuales, además de la media, cuya distribución muestral es cercana a ala normal si las muestras son los suficientemente grandes, tales como las proporciones y la diferencia de medias. Si bien veremos con detalle la prueba de hipótesis con estos estimadores puntuales en capítulos posteriores, es importante contar con algunas orientaciones generales.

El estimador puntual  debe ser insesgado. Esto significa que la distribución muestral de teta tongo tiene su centro en el valor del parámetro teta que estima. Debe proveer una estimación cercana al parámetro real.

Texto

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente con confianza media

4.6.2 ESTIMADORES CON OTRAS DISTRIBUCIONES

Existen métodos de construcción de intervalos de confianza y prueba de hipótesis adecuados para aquellos casos en que el estimador puntual o el estadístico de prueba no son cercanos a la normal (por ejemplo, si la muestra es pequeña, se tiene una mala estimación del error estándar o el estimador puntual tiene una distribución distinta a la normal).la selección del métodos alternativos debe hacerse siempre teniendo en cuenta la distribución muestral del estimador puntual o del estadístico de prueba.

Una consideración importante es que **siempre debemos verificar el cumplimiento de las condiciones requeridas por una herramienta estadística**, pues de lo contrario las condiciones pueden ser erradas y carecerán de validez

CAPITULO 5. INFERENCIA CON MEDIAS MUESTRALES

Modelo normal, es decir la distribución muestral de la media sigue aproximadamente una distribución normal, supuesto que en general se cumple si la muestra tiene a lo menos 30 observaciones

Conoceremos las primeras pruebas estadísticas, las cuales nos permiten inferir acerca de una o dos medias muestrales.

**5.1 PRUEBA Z**

prueba adecuada para inferir acerca de las medias con una muestra. Para usarla se deben cumplir ciertas condiciones, muchas están asociadas al modelo normal que conocimos en el capitulo anterior:

* + La muestra debe tener al menos 30 observaciones, si tiene menos de 30 se debe conocer la varianza de la población
  + Las observaciones deben ser independientes, es decir que la elección de una observación para la muestra no influye en la selección de otras
  + La población de donde se obtuvo la muestra sigue aproximadamente una distribución normal

Esta prueba resulta adecuada si queremos **asegurar o descartar** que la media de la población tiene un cierto **valor hipotético.**

Ejemplo, queremos saber si, en promedio, las utilidades mensuales de una pequeña empresa son de 20 millones de pesos y que el gerente general, Esteban, nos ha informado que la desviación estándar para las utilidades es de 2,32 millones de pesos. El señor Esteban nos ha proporcionado una muestra , obtenida mediante muestreo aleatorio simple, con las utilidades (en millones de pesos) reportadas para 20 meses, que muestra la tabla.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

El señor Esteban nos ha dicho que debemos ser muy exigentes con respecto a nuestras conclusiones, por lo que se debe usar un nivel de significación alfa = 0.01 (es decir, un nivel de confianza de 99%)

Formular las hipótesis:

H0: la media de las utilidades mensuales de la empresa µ es de 20 millones de pesos, es decir

µ = 20 [M$]

HA: las utilidades mensuales de la empresa son, en promedio, distintas de 20 millones de pesos, es decir: µ≠ 20[M$]

Verificar el cumplimiento de las condiciones para usar la prueba Z. en cuanto a la primera condición el enunciado nos indica que si bien la muestra tiene 20 observaciones, la desviación de la población es conocida, por lo que se verifica su cumplimiento. Tambien podemos comprobar en el enunciado que las observaciones son independientes entre si, pues fueron obtenidas mediante muestreo aleatorio simple. Si bien no estamos seguros que esta muestra considere menos del 10% de las observaciones podemos ocuparla razonablemente.

Se realiza el grafico Q-Q y nos indica que no se observan valores atípicos. Otra forma de comprobar esta condición es mediante la prueba de Shapiro-Wilk que se puede realizar en R mediante la función shapiro.test(X), donde x es un vector con las observaciones de la muestra:

hipótesis nula de esta prueba es que la muestra fue extraída desde una distribución normal

hipótesis alternativa la muestra fue extraída desde una distribución diferente a la normal

Al ejecutar el script, podemos ver que el valor p obtenido es p = 0.244 > alfa = 0.01 (p valor muy superior a nuestro nivel de significación) por lo que podemos asegurar con relativa confianza que la población de donde proviene la muestra sigue una distribución normal.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Luego de comprobar que se cumplan todas las condiciones, podemos hacer una prueba Z para una muestra. Se comienza calculando el estadístico de prueba:

Tabla

Descripción generada automáticamente

Con este resultado calculamos el valor p. debemos recordar que las funciones en R nos entregan la probabilidad asociada al área correspondiente a una sola cola de la distribución, por lo que debemos multiplicar el resultado por 2 para considerar ambas colas si, como en este caso, se trata de una prueba bilateral. Al hacer la llamada **2 \* pnorm(2.6147, lower.tail = FALSE)**  se obtiene un p valor = 0.009 < 0.01 = alfa, con lo que se rechaza la hipótesis nula en favor de la hipótesis alternativa, debemos ser cuidadosos porque nuestro p valor es cercano al nivel de significación establecido, por lo que se debería evaluar los resultados con una muestra mas grande. Concluimos que los datos sugieren que en promedio las utilidades mensuales de la empresa difieren de los 20 millones establecidos.

Texto

Descripción generada automáticamente



5.2 PRUEBA T DE STUDENT

En la practica rara vez podemos conocer la desviación estándar de la población y a menudo nos encontraremos con muestras pequeñas, por lo que la prueba Z no es muy utilizada.

En el caso de la media, el teorema del limite central se cumple para datos normales, es decir, independientemente del tamaño de la muestra, la media muestral tendrá una distribución cercana a la normal siempre que la observaciones sean independientes y provengan de una distribución cercana a ala normal. Cuando el conjunto de datos es muy pequeño es muy difícil comprobar el cumplimientos de las condiciones.

En el capitulo 3 conocimos la distribución t de Student o simplemente distribución t. se destaca esta distribución, siempre centrada en 0 y definida únicamente por los grados de libertad (Ѵ) como parámetro, es su semejanza con la distribución normal pese a que sus colar son algo mas gruesas. Este grosor adicional de las colas tiene como consecuencia que para la distribución t, es mas probable que una observación este a mas de dos desviaciones estándares de la media que en el caso de la distribución normal. Este fenómeno permite que la estimación del error estándar sea mas certera que al usar la distribución normal cuando el conjunto de datos es pequeño.

La prueba t de Student basada en la distribución t, es en consecuencia la alternativa más ampliamente empleada para inferir acerca de una o dos medias muestrales.

**5.2.1 PRUEBA T PARA UNA MUESTRA**

Aunque la prueba t no opera bajo el supuesto de normalidad, aun así, requiere verificar algunas condiciones para poder usarla:

1. Las observaciones son independientes entre si
2. Las observaciones provienen de una distribución cercana a la normal

las condiciones son casi las mimas que para la prueba Z, excepto por el hecho de que no limitan el tamaño de la muestra para ver que sea mayor a 30. La ventaja evidente de eliminar esta restricción es que la distribución t permite su uso para muestras pequeñas igualmente adecuada para muestras grandes. Esto se debe a que la distribución t es regulada por los grados de libertad y a medida que aumentan, mas se parece a una distribución normal. Este parámetro, al trabajar con medias de muestras de tamaño n, siempre estará dado por Ѵ=n-1

veamos el siguiente ejemplo:

un ingeniero desea determinar si el tiempo promedia que tarda una implementación dada de un algoritmo en resolver un problema, sabiendo que el algoritmo siempre se ejecuta en las mismas condiciones(misma maquina, igual disponibilidad) es inferior a 500 milisegundos. Para ello ha seleccionado aleatoriamente 15 instancias del problema y registrado el tiempo en milisegundos para cada una de ellas.

Tabla

Descripción generada automáticamente

Primer paso, formular las hipótesis:

H0: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es igual a 500 milisegundos

HA: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es inferior a 500 milisegundos

Recordar que µ0 es el valor nulo, por lo que en este caso µ0=500 [ms]. Matemáticamente las hipótesis puede formularse como:

H0: µ=µ0, esto es µ = 500

HA: µ < µ0, es decir µ <500

Ahora hay que verificar que se cumplan las condiciones necesarias para usar la distribución t:

* Como las muestras fueron elegidas al azar, se puede asumir que son independientes
* El grafico de la figura 5.3 muestra que es valido suponer una distribución cercana a la normal. Si bien los puntos de la muestra no forman una recta, no se observan valores atípicos que se alejen de la región aceptable.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

La media muestral X = 434.2921 y la desviación estándar s = 38.0963



El estadístico de prueba es el estadístico T, el cual sigue una distribución t con Ѵ = n-1 con grados de libertar y esta dado por la ecuación 5.1 donde la expresión(s / raíz n) corresponde al error estándar de la media (cuando no se conoce la desviación estándar de la población Ϭ)

Patrón de fondo

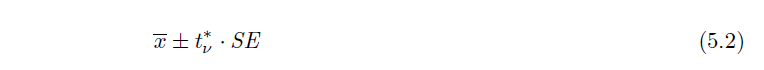
Descripción generada automáticamente con confianza baja

Word

Descripción generada automáticamente con confianza media

A partir de este resultado, obtenemos el valor p con ayuda de la función pt(), obteniéndose p=5.219\*10^-6 o simplemente como dicta la convención p < 0.01

La formula para construir el intervalo de confianza es ligeramente diferente al caso normal, como muestra la ecuación 5.2. para este ejemplo se considera un nivel de confianza de 97.5% es decir un nivel de significación alfa=0.025



Texto

Descripción generada automáticamente con confianza media

Una vez mas R, permite realizar esta prueba de manera mas rápida gracias a la función **t.test(x,alternative,mu,conf.level) donde:**

Texto

Descripción generada automáticamente

a partir de los resultados podemos observar que el valor o obtenido es muy pequeño, dando a entender que, si se cumple el supuesto de que verdadera media es µ =500 (hipótesis nula), seria muy improbable obtener una media muestra de X=434.2921. además el valor p es mucho menor que el nivel de significación , por lo que la evidencia a favor de HA es muy fuerte, en consecuencia se rechaza H0 en favor de HA. Se puede afirmar con 97,5% de confianza que el tiempo promedia que tarda el algoritmo en resolver una instancia del problema es inferior a 500 milisegundos.



**5.2.2 PRUEBA T PARA DOS MUESTRAS PAREADAS**

Para esta prueba supongamos ahora que el ingeniero tiene dos algoritmos diferentes A y B, que en teoría deberían tardar lo mismo en resolver un problema. Para ello probo algoritmos con 35 instancias del problema elegidas al azar de igual tamaño y registro sus tiempos además de calcular la diferencia en los tiempos de ejecución como muestra la tabla. El ingeniero desea comprobar si efectivamente el rendimiento de ambos algoritmos es equivalente.

Tabla

Descripción generada automáticamente

En este ejemplo, tenemos dos tiempos de ejecución diferentes para cada instancia del problema uno con cada algoritmo. En consecuencia, los datos están **PAREADOS**. Es decir cada observación del conjunto tiene una correspondencia o conexión especial con exactamente una observación del otro. Para examinar datos pareados usamos la diferencia entre cada par de observaciones, para lo cual podemos usar la técnica de la distribución t vista anteriormente.



La media de las diferencia es Xdif=-12,08591 y la desviación estándar es sdif = 36.08183

Paso 1. Definimos las hipótesis:

H0: la media de las diferencias en los tiempos de ejecución es igual a 0

HA: la media de las diferencias en los tiempos de ejecución es distinta de 0.

Matemáticamente se expresan:

H0: µdif = 0

HS: µdif ≠ 0

Paso 2. Verificar condiciones

Como las instancias fueron escogidas al azar, se supone que las observaciones son independientes pues el conjunto de instancias posibles es infinito y las 35 seleccionadas no superan el 10% de la población. Además al aplicar la prueba de normalidad de Shapiro-Wilk se obtiene p = 0.357 con lo que podemos concluir que la diferencia en los tiempos de ejecucion se acerca razonablemente a una distribución normal. En consecuencia podemos proceder con la prueba t de Student. El ingeniero no necesita ser riguroso, por lo que usaremos un nivel de confianza del 95%.

La función t.test() de R, permite efectuar la prueba de dos maneras diferentes (con idéntico resultado), como muestra el script. La primera de ellas es aplicar la prueba t directamente a las diferencias, tal como en la sección anterior (es decir una prueba t para cada muestra). La segunda consiste en entregar a la función ambas muestras por separado e indicarle que están pareadas, en esta caso la llamada tiene la forma **t.test(x,y,paired, alternative,mu,conf.level) donde los argumentos son:**

Texto

Descripción generada automáticamente

5.2.3 PRUEBA t PARA DOS MUESTRAS INDEPENDIENTES

La prueba t se usa para comparar las medias de dos poblaciones en que las observaciones con que se encuentra no tienen relación con ninguna de las otras observaciones, ni influyen en su selección, ni en la misma ni en la otra muestra. La inferencia se hace sobre la diferencia de las medias µ1 - µ2 = d0, donde d0 es un valor hipotético fijo para la diferencia. Usualmente d0=0, en cuyo caso las muestras podrían venir de dos poblaciones distintas con igual media o desde la misma población. Para ello, la prueba usa como estimador puntual la diferencia de las medias muestrales (X1 – X2), así el estadístico T en este caso toma la forma de la ecuación 5.3



Patrón de fondo

Descripción generada automáticamente con confianza baja

Al usar la distribución t de Student para la diferencia de medias, se deben cumplir los siguientes requisitos:

1. Cada muestra cumple las condiciones para usar la distribución t
2. Las muestras son independientes

Un ejemplo para este caso: el doctor Matta desea determinar si una nueva vacuna A es mas efectiva que una vacuna B a fin de inmunizar a la población mundial contra el covid. Para ello recluto 28 voluntarios en diferentes países, 15 de los cuales seleccionados al azar recibieron la vacuna A y los 13 restantes la vacuna B, la tabla muestra para cada voluntario la concentración de anticuerpos al cabo de un mes de recibir la vacuna.

Tabla

Descripción generada automáticamente

Paso 1: formular hipótesis

H0: no hay diferencia entre la efectividad promedia de ambas vacunas

HA: la vacuna A es en promedio mas efectiva que la B

En lenguaje matemático

H0: µA - µB = 0 -> µA = µB

HA: µA - µB > 0 -> µA > µB

Paso 2. Verificar condiciones

Ambas muestras son independientes entre si, pues son diferentes voluntarios y fueron designados aleatoriamente a cada grupo. Además, se asume que las observaciones son independientes, pues cada muestra es significativamente menor a la población total a vacunar. En cuanto a la normalidad para cada muestra, al aplicar a cada una la prueba de Shapiro-Wilk se obtiene respectivamente p=0.428 y p=0.445 en ambos casos p valor es bastante alto,por lo que podemos concluir que ambas muestras provienen de poblaciones que se distribuyen de forma aproximadamente normal. Ahora que hemos comprobado que cumplen las condiciones, podemos llevar a cabo la prueba t pata dos muestras independientes.

Ahora bien, como las muestras son algo pequeñas, serie prufente proceder con algo de cautela. Además en este escenario un error de tipo I (rechazar H0 cuando es verdadera) implicaría reducir innecesariamente la cantidad de vacunas disponibles y retrasar el proceso de vacunación, poniendo en riesgo a todos los habitantes del planeta. Un error de tipo II, en cambio podría causar que se continue el uso instinto de ambas vacunas retrasando ligeramente el efecto inmune en la población. En consecuencia, el error de tipo I es mas grave, por lo que el nivel de significación debiese ser aun mas exigente, en consecuencia optamos por usar alfa = 0.01

Al aplicar la prueba t, obtenemos que la diferencia entre las medias es 6.683 [mg/ml] y que el intervalo de confianza es [2.2739,infinito]. Además, el valor p es p <0.001, muy inferior al nivel de signidicacion alfa=0.01. esto significa que la evidencia en favor de HA es muy fuerte, por lo rechazamos la hipótesis nula. En consecuencia, podemos concluir con 99% de confianza que la vacuna A es, en promedio, mejor que la vacuna B (produce una mayor concentración media de anticuerpos en las personas vacunadas con ella que la producida por la vacuna B)

Si leímos atentamente, habremos notado que no hemos definido el error estándar para cuando tenemos dos muestras. En este caso, SE construye a partir del error estándar de cada muestra, como se aprecia en la ecuación 5.4. este escenario,la determinación de los grados de libertad es mas compleja por lo que se recomienda usar programas estadísticos o en su defecto escoger el menor valor entre n1-1 y n2-1

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Se puede lograr un mejor ajuste de la distribución t si se sabe con certeza que las desviaciones estándares de ambas poblaciones son casi iguales. En este caso, se puede usar una **varianza agrupada (**s2p, del ingles pooled variance) que reemplaza tanto a s21 como a s22. Esta varianza agrupada se calcula como muestra la ecuación 5.5 y en este caso se consideran n1+n2-2 grados de libertad.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

**CAPITULO 6. PODER ESTADISTICO**

en el capitulo 4 estudiamos el procedimiento para someter hipótesis a prueba, junto con los errores de decisión que podríamos cometer:

* **Error de tipo I: rechazar H0 en favor de HA cuando H0 es verdadera**
* **Error de tipo II: no rechazar H0 en favor de HA cuando HA es verdadera**

Allí conocimos el nivel de significación α, como herramienta para representar y de alguna manera controlar la probabilidad de cometer un error de tipo 1, con lo que la preocupación se centra en controlar la ocurrencia de esta clase de errores, desviando la atención de cometer errores de tipo 2. Esto se debe a que la hipótesis nula representa el **status quo**, es decir, mantener las cosas y creencias tal como están y por ende, cuando no se rechaza H0 no suele tomarse ninguna acción. En contraste, la hipótesis alternativa describe un cambio de condiciones, por lo que rechazar H0 en favor de HA usualmente conlleva un esfuerzo, mayor costoso, para adaptarse o aprovechar las nuevas condiciones.

El valor de α debe ser acorde con las consecuencias de cometer errores tanto de tipo 1 como 2, pero no sabemos como se relaciona el nivel de significación con los errores de tipo II.

Así como el nivel de significación α corresponde a la probabilidad de cometer errores de tipo I, definimos β como la probabilidad de cometer errores de tipo II. .. α y β están relacionados **para un tamaño fijo de la muestra: al reducir β, α aumenta y viceversa** . este fenómeno queda en evidencia con mayor fuerza mientras mas pequeña sea la muestra. En la practica resulta conocer la probabilidad de **NO** cometer errores de tipo II. Esto nos lleva a un nuevo concepto: **el poder estadístico** de una prueba de hipótesis, dado por **1- β** , que se define como la **probabilidad de correctamente rechazar H0 cuando es falsa.**

Otra forma de entender la noción de poder de una prueba es que tan propensa es esta para distinguir un efecto real de una simple casualidad, lo que nos lleva a la noción de **tamaño del efecto,** que corresponde a una cuantificación de la diferencia entre dos grupos o del valor observado con respecto al valor nulo.

En el capitulo 5 conocimos la prueba t para inferir acerca de dos medias. En este contexto, el tamaño del efecto corresponde a que tan grande es la diferencia real entre ambas.

6.1 PODER, NIVEL DE SIGNIFICACION Y TAMAÑO DE LA MUESTRA

El poder corresponde a la probabilidad de **no** cometer un error de tipo II y que esta muy relacionado con el tamaño de la muestra. También existe una relación entre poder y nivel de significación.

Imagen que contiene Texto

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

De manera similar, la figura 6.2 considera las mismas muestras y los mismos niveles de significación que la figura 6.1, pero ahora para una prueba t unilateral. 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 seria posible detectar una diferencia en el sentido opuesto aunque fuese muy grande (pues no hay región de rechazo en dicho sentido). El script empleado para la construcción de la figura 6.2 es idéntico al script 6.1, excepto porque el argumento alternative toma como valor “one.sided” en las llamadas power.t.test()

La figura 6.3 muestra las curvas de poder para una prueba t unilateral y otra bilateral, ambas para una muestra de tamaño 6, desviación estándar s=1 y α=0.05. en ella se evidencia claramente que la ventaja de las pruebas unilaterales: cuando el tamaño de del efecto aumenta en el sentido de la hipótesis alternativa, el poder es mayor que para una prueba bilateral.

Es deseable que las pruebas que se empleen para docimar hipótesis tengan un alto poder y , si hay mas de una prueba disponible, se debe escoger la mas poderosa. Los cálculos del poder suelen ser altamente complejos. La teoría permite en muchos casos conocer la prueba con mayor poder posible ante cualquier hipótesis alternativa, nivel de significación y tamaño de la muestra (siempre que se cumplan las condiciones de base). Estas pruebas reciben el nombre de **uniformemente mas poderosas** y tal es el caso de la prueba t de Student.

6.2 TAMAÑO DEL EFECTO

El problema qur podríamos tener al considerar el tamaño del efecto en la misma escala de la variable estudiada, como hemos hecho hasta ahora, es que esta escala varia de variable en variable. Para poder hacer comparaciones con mayor libertad, existen diferentes **medias estandarizadas de efecto** que podemos usar. En esta sección conoceremos la llamada **d de Cohen** una media estándar ampliamente empleada para el tamaño del efecto con esta prueba.

En términos generales, se considera que d=0.2 es un efecto pequeño, d=0.5 un efecto mediano y d=0.8 un efecto grande.

En el caso de la prueba de t de una muestra, la d de Cohen se calcula como la muestra la ecuación 6.1 donde:



* X : media muestral
* µ0: media teórica para el contraste (valor nulo)
* s: desviación estándar de la muestra con n-1 grados de libertad

Patrón de fondo

Descripción generada automáticamente con confianza baja

Para la prueba t de diferencia de dos medias (tambien llamda prueba t para dos muestras independientes o simplemente prueba t independiente), si el tamaño de la muestra es mayor a 50 elementos, se calcula como muestra la ecuación 6.2 y para muestras pequeñas se aplica un factor de corrección, como indica la ecuación 6.3, donde:

* X1 y X2 : medias muestrales de cada grupo



* N1 y n2 son los tamaños de ambas muestras
* Sp: desviación estándar agrupada, dada por la ecuación 6.4

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

En el caso de la variante de Welch para la prueba t independiente, la formula para el calculo de la d de Cohen es ligeramente diferente, como puede apreciarse en la ecuación 6.5

Patrón de fondo

Descripción generada automáticamente con confianza baja

Por ultimo, las ecuaciones 6.6 y 6.7 muestran como se calcula la d de Cohen en el caso de la prueba t con muestras pareadas grandes (n>50) y pequeñas, respectivamente, donde D corresponden a las diferencias entre las observaciones pareadas

Patrón de fondo

Descripción generada automáticamente con confianza baja

Patrón de fondo

Descripción generada automáticamente con confianza baja

6.3 PODER, TAMAÑO DEL EFECTO Y TAMAÑO DE LA MUESTRA

Mencionamos en paginas anteriores que el poder puede tambien entenderse como que tan propensa es una prueba estadística para distinguir un efecto real de una simple casualidad y que podemos cuantificar este efecto.

Una gran ventaja del poder estadístico es que nos sirve para determinar el tamaño adecuado de la muestra para detectar un cierto tamaño del efecto. La figura 6.4 elaborada con el script 6.2 muestra el aumento del poder estadístico a medida que el tamaño de la muestra aumenta (para un tamaño del efecto y nivel de significación fijos). En ella se aprecia que a medida que el tamaño de la muestra crece, el poder estadístico también crece asintóticamente a 1, valor que equivale a tener la certeza de rechazar la hipótesis nula si esta es falsa.

Gráfico

Descripción generada automáticamente

6.4 CALCULO TEORICO DEL PODER

El poder es la probabilidad de correctamente rechazar H0, cuando es falsa, lo que equivale a la probabilidad de distinguir un efecto real de una mera casualidad. Ahora veremos algunos ejemplos.

Lola ha diseñado dos nuevos algoritmos (A y B) que resuelven un mismo problema. Lola desea saber si existe diferencia entre los tiempos de ejecución de ambos algoritmos. Para ello ha decidido realizar una prueba t con muestras pareadas, con un nivel de significación α=0.05, usando para ellos 36 instancias del problema de tamaño fijo que se ejecutan bajo iguales condiciones. Lola ya sabe que la diferencia en el tiempo de ejecución sigue una distribución normal con desviación estándar Ϭ = 12 milisegundos. Lola ha formulado las siguientes hipótesis:

H0: µ(Ai -Bi) = 0, es decir que la media de las diferencias en el tiempo de ejecución necesitado por los algoritmos Ay B para cada posible instancia i , es cero.

HA: µ(Ai -Bi) ≠0

La figura 6.5 muestra como seria la distribución de la muestra (media de las diferencias en los tiempos de ejecución) si la hipótesis nula H0 fuera cierta, con las áreas correspondientes a la región de rechazo de H0 coloreadas

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Supongamos por un moento que, en realidad el algoritmo B en promedio es 4 milisegundos mas rápido que el algoritmo A. en este caso tendríamos que la media de las diferencias es de -4 ms, correspondiente al tamaño del efecto. En este caso, su distribución seria como muestra la figura 6.6 en color azul. Al superponer esta nueva curva a la que ya teníamos bajo el supuesto de que la hipótesis nula fuera verdadera, vemos que el área bajo la curva real que se situaría dentro de la región de rechazo de la curva teórica es aquella coloreada en azul. Esta área corresponde al poder de la prueba t , que en este caso es 0.515 de acuerdo al análisis teórico. Puesto que el poder corresponde a la probabilidad de NO cometer un error de tipo II, de acuerdo al resultado obtenido se tiene que β=0.484, lola no seria capaz de detectar una diferencia de -4 casi la mitad de las veces.

Gráfico, Histograma

Descripción generada automáticamente

6.5 calculo del poder en R

En R podemos usar funciones para calcular el poder. Como primera alternativa, R trae incorporada la función power.t.test(n,delta,sd,sig.level, power, typw, alternative (empleadas en los script) donde:

Texto

Descripción generada automáticamente

Esta función entrega como resultado un objeto con diversos elementos (que podemos indexar del mismo modo que las columnas de una matriz de datos), entre los que se incluyen los 5 primeros argumentos definidos para la función.

Si revisamos con detenimiento los argumentos de la función power.t.test() veremos que recibe el poder como uno de sus argumentos, esto no tiene sentido. Como ya hemos visto existe una relación entre: poder, tamaño de la muestra, tamaño del efecto y nivel de significación. A esta combinación de elementos debemos añadir también la desviación estándar, aunque no estudiaremos las matemáticas subyacentes.

En realidad para usar la función power.t.test() siempre debemos señalar el tipo de prueba t con el que estamos trabajando y si la hipótesis alternativa es de una o dos colas. Esta función nos permite calcular cualquiera de los demás argumentos (tamaño de la muestra, tamaño del efecto, desviación estándar, nivel de significación o poder estadístico) para la prueba en cuestión a partir de los 4 argumentos restantes. Así al argumento que queremos calcular se le asigna el valor NULL en la llamada.

Recordemos que el ejemplo anterior de la sección anterior. Lola desea usar una prueba t bilateral para dos muestras pareadas a fin de determinar si hay diferencia entre los tiempos de ejecución promedio de ambos algoritmos. Para ellos ha considerado n=36 y α=0.05, sabiendo que sd=12. Las líneas 4 a la 14 muestran como calcular el poder para este ejemplo si se desea detectar un tamaño del efecto de 4 ms, obteniéndose como resultado que el poder es de 0.494 (y β=1-poder = 0.506) ligeramente diferente al obtenido en forma teórica debido a los errores de redondeo.

¿Cuántas instancias debería usar Lola para lograr un poder de 0.9, manteniendo α=0.05, sd = 12ms y g=4? Las líneas 17-28 del script muestran como hacer este calculo obteniéndose como resultado n=97. Como el tamaño de la muestra siempre debe ser entero positivo, la línea 27 aproxima el resultado al entero superior.

Otra alternativa es usar la función pwr.t.test(n,d,sig.level,power,type,alternative) donde:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Debemos fijarnos en que, si bien esta función opera de manera similar a power.t.test(), en este caso la desviación estándar y la diferencia son reemplazadas por el tamaño del efecto que podemos cuantificar, como ya vimos, mediante la d de Cohen. Sin embargo, debemos tener cuidado, pues la función pwr.t.test() solo es adecuada para una muestra, dos muestras pareadas o cuando ambas muestras tienen igual tamaño. En el caso de la prueba t para dos muestras independientes con diferentes tamaños, debemos usar en cambio la función pwr.t2n.test(n1,n2,d,sig.level,power,alternative).

CAPITULO 7. INFERENCIA CON PROPORCIONES MUESTRALES

En el capitulo 5 conocimos las pruebas Z y t de Student para contrastar hipótesis con una y dos medias. Ahora estudiaremos el método de Wald y Wilson para inferir acerca de una y dos proporciones.

**7.1 método de WALD**

En el capitulo 3 vimos que cuando queremos responder preguntas del tipo ¿Qué proporción de la ciudadanía apoya al gobierno actual?, estamos hablando de una variable aleatoria que sigue una distribución binomial. En general, no conocemos la **probabilidad de éxito *p***de la población, por lo que tenemos que usar el estimador puntual (correspondiente a la proporción de éxito de la muestra), denotado por p . Este estimador se distribuye de manera cercana a la normal cuando se cumplen las siguientes condiciones:



1. Las observaciones de la muestra son independientes
2. Se cumpke la **condición de éxito-fracaso** , que establece que se espera al observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos. Matematicamente np >=10 y n(1-p)>=10

Así la distribución muestral de p cumple con las condiciones anteriores, se dice que es cercana a la normalidad con media µ = p, desviación estándar Ϭ=raíz(p(1-p)) y error estándar dado por la ecuación 7.1



Patrón de fondo

Descripción generada automáticamente con confianza baja

**7.1.1 METODO DE WALD PARA UNA PROPORCION**

El metodo de **Wald** permite construir instervalos de confianza y contrastar hipótesis bajo el supuesto de normalidad para una proporción. Consideremos el siguiente ejemplo: Aquiles desea conocer que porporcion de las ejecuciones de un algoritmo de ordenamiento para instancias con 100.000 elementos (bajo iguales condiciones) tardan menos de 25 segundos. Para ello registro los tiempos de ejecución para 150 instancias generales de manera aleatoria, encontrando que 64% de dichas instancias fueron resueltas en un tiempo menor al señalado.

Si bien no conocemos la probabilidad real de éxito para la población, sabemos que p = 0.64. así, si se cumplen las condiciones para que la distribución de p sea cercana a la normal, podemos construir un intervalo de confianza para la verdadera proporción muestral.



En el enunciado nos indican que las instancias del problema fueron escogidas de manera aleatoria y sabemos que estas representan menos del 10% del total de instancias posibles, con lo que se verifica independencia de las observaciones. Por otra parte, nos dicen que la proporción de éxito es p = 0.64 , por lo que esperamos encontrar 0.64 \* 150 = 96 instancias que tardan menos de 25 segundos y (1-0.64)\*150 = 54 fracasos (instancias que tardan 25 segundos o mas), con lo que se cumple la condición de éxito-fracaso. En consecuencia, podemos asumir que la distribución muestral de p sigue aproximadamente a la normal.

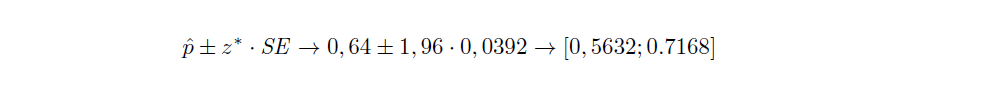


Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Con ello, construimos el intervalo de confianza para un nivel de significación α=0.05, usando la ecuación general 4.6 con p como estimador puntual:





Esta intervalo significa que tenemos 95% de confianza que la proporción de instancias (de 100.000 elementos) del problema que el algoritmo ordena en menos de 25 segundos se encuentra entre 56,32% y 71,6%

Desde luego, tambien podemos usar el modelo normal en el contexto de la prueba de hipótesis para una proporción. Para ello, se deben cumplir las condiciones de independencia y éxito-fracaso que ta verificamos para construir el intervalo de confianza, pero en este caso tenemos que verificar la segunda condición con el valor nulo , denotado por p0 . Una vez verificadas ambas condiciones, el error estándar y el estadístico Z que permiten calcular el p-valor se calculan usando las siguientes ecuaciones

Imagen que contiene Aplicación

Descripción generada automáticamente

Volviendo al ejemplo, Aquiles afirma que mas del 70% de las instancias de tamaño 100.000 se ejecutan en menos de 25 segundos. Sin embargo, su jefe no esta seguro, por lo que decide comprobarlo mediante una prueba de hipótesis con un nivel de significación α=0.05 (recordar que n=150 y p = 0.64)



H0: el 70% de las instancias se ejecutan en menos de 25 segundos

HA: mas del 70% de las instancias se ejecutan en menos de 25 segundos

Denotando como p a la proporción de todas las instancias de tamaño 100.000 que se ejcutan en menos de 25 segundos y considerando el valor hipotético p0=0.7 para este parámetro

H0: p = p0

HA: p > p0

Anteriormente comprobamos la independencia. Considerando que el valor nulo fuese verdadero esperaríamos encontrar 0.7\*150=105 exitos y (1-0.7)\*150=45 fracasos, ambos valores mayores que 10, por lo que la condición de éxito-fracaso se verifica

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

El valor p asociado, calculado con R mediante la llamada a la función pnorm(-1.6042, lower.tail = FALSE) es p = 0.9456. en consecuencia, la evidencia no es suficiente para rechazar la hipótesis nula, por lo que se concluye con un 95% de confianza que no es cierto que el algoritmo se ejecute en menos de 25 segundos para mas del 70% de las instancias de tamaño 100.000

R no ofrece esta prueba, como función. Sin embargo, podemos hacerla como muestra el script.

7.1.2 METODO DE WALD PARA DOS PROPORCIONES

Tambien se puede utilizar el metodo de Wald para estudiar **diferencia entre las proporciones de dos poblaciones,** considerando para ello como estimador puntual la diferencia p1-p2.



Verificar condiciones:

1. Cada proporción, por separado, sigue el modelo normal
2. Las dos muestras son independientes una de la otra

El error estándar para la diferencia entre dos proporciones muestrales esta dado por la ecuación 7.4, donde p1 y p2 corresponden a las proporciones de las poblaciones y n1 y n2 a los tamaños de las muestras. La construcción del intervalo de confianza se realiza, una vez mas con la ecuación general 4.6

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ejemplo: una universidad desea determinar si la tasa de reprobación de estudiantes que rinden la asignatura de programación por primera vez es igual para hombres y mujeres. Para ello, se examina la situación final de los estudiantes que rindieron la asignatura durante el segundo semestre de 2017. Para una muestra de 48 hombres (de un total de 632), se encontró que 26 de ellos reprobaron la asignatura. De manera similar, para una muestra de 42 mujeres (de un total de 507) se encontraron 20 reprobaciones, con ambas muestras tomadas de manera aleatoria.

Paso 1. Verificar condiciones de normalidad para cada una de las muestras

En ambos casos, las observaciones son independientes entre si, pues provienen de personas diferentes que representan a menos del 10% de la población. Además los datos entregados evidencian que en ambos casos se cumple la condición éxito-fracaso. Adicionalmente ambas muestras son independientes entre si, pues ambas categorías se excluyen entre si mutuamente. Se verifica la normalidad para la diferencia de proporciones.

Sean p1 y p2 las proporciones de éxito muestrales (considerando en este contexto la reprobación como éxito para hombres y mujeres, respectivamente):



Texto, Carta

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente



En consecuencia , podemos afirmar con 95% de confianza que la diferencia en la tasa de reprobación de la asignatura de programación para hombres y mujeres varia entre -14,11% y 27,21%

Tambien se puede realizar prueba de hipótesis en este escenario:

H0: no hay diferencia en la tasa de reprobación de hombres y mujeres

HA: las tasas de reprobación son diferentes para hombres y mujeres

Texto

Descripción generada automáticamente con confianza baja

Ya comprobamos el supuesto de normalidad cuando construimos el intervalo de confianza. Sin embargo **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 para ello la **proporción agrupada** , dada por la ecuación 7.5 donde p1n1 y p2n2 representan la cantidad de éxitos en la primera y segunda muestra.

Imagen que contiene Texto

Descripción generada automáticamente

En consecuencia, en el caso de los hombres encontrar pn1>24 exitos (reprobaciones) y (1-p)n1 >23 fracasos. Del mismo modo, para las mujeres esperamos pn2>21 exitos y (1-p)n2 >20 fracasos, con lo que se verifican las condiciones para emplear el modelo normal.

El error estándar se calcula como ya mencionamos usando la proporción agrupada:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

En consecuencia, el valor p correspondiente es p=0.5351. puesto que el valor p es mayor a alfa=0.05 se falla en rechazar la hipótesis nula. Así, podemos decir con 95% de confianza que no existe evidencia suficiente para concluir que hay diferencia en la tasa de reprobación de hombre y mujeres para el primer curso de programación.

Cuando contrastamos hipótesis para la **diferencia entre dos hipótesis con un valor nulo distinto de 0**, el procedimiento es ligeramente diferente. En este caso la comprobación de éxito-fracaso se realiza de manera independiente para ambas muestras y el error estándar se calcula como ya se estudio para los intervalos de confianza mediante la ecuación 7.4

Otro ejemplo: supongamos que la universidad ha decidido replicar el estudio realizado para el curso de programación, esta vez para una asignatura de física . las autoridades están convencidad de que la tasa de reporbacion es 10% mayor para los hombres y que incluso la diferencia podría ser mayor. Desean comprobar con un nivel de confianza de 95% y para ello, seleccionaron aleatoreamente a 89 de los 1023 hombres y a 61 de las 620 mujeres de la cohorte correspondiente al primer semestre de 2019. En las muestras se encuentran, respectivamente 45 y 21 reprobaciones

Las hipótesis son :

H0: la tasa de reprobación de los hombres es exactamente 10% mas alta que de las mujeres

HA: la tasa de reprobación de los hombres es mas de 10% mas alta que de las mujeres

Imagen que contiene Texto

Descripción generada automáticamente

Comprobando condiciones: las observaciones de cada muestra son independientes entre si pues corresponden a menos del 10% de la población y fueron escogidos aleatoriamente. A su vez, los datos proporcionados indican que se cumple la condición de éxito-fracaso para cada muestra. Como ambas muestras pertenecen a grupos diferentes de estudiantes, son independientes entre si, en consecuencia se cumplen las condiciones para operar bajo el modelo normal.

En el caso de los hombres la tasa de éxito se estima como:



Análogamente para las mujeres tenemos



Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Escala de tiempo

Descripción generada automáticamente con confianza media

En consecuencia, se falla en rechazar H0 en favor de HA, por lo que concluimos , con 95% de confianza, que la tasa de reprobación de los hombres es 10% superior a la de las mujeres para el curso de física.

7.2 METODO DE WILSON

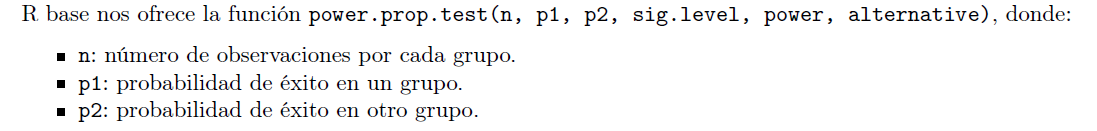
El metodo de wald, trata en la sección anterior, es el metodo que tradicionalmente se ha usado y el que aparece en la mayoría de los libros clásicos de inferencia estadística. Pero es criticado porque hace importantes simplificaciones matemáticas en su procedimiento y ya hay evidencia empírica que ha demostrado sus limitaciones. Así nace el metodo de Wilson mas robusto, las formulas para estimar la proporción en la muestra y el error estándar son diferentes.

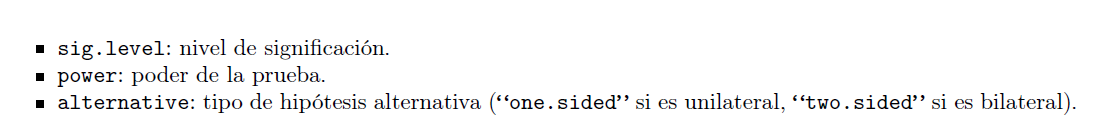
Texto

Descripción generada automáticamente

7.3 PODER Y PRUEBAS DE PROPORCIONES

En capitulo anterior conocimos el poder estadístico y vimos que esta relacionado con el nivel de significación, el tamaño de la muestra y el tamaño del efecto que queremos detectar





Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

CAPITULO 8. INFERENCIA NO PARAMETRICA CON PROPORCIONES

Que significa “no parametrica”, en el capitulo 5 conocimos las prueba Z y de t Student. Ambas formulan hipótesis relativas al parámetro µ de una distribución normal (o la diferencia µ1 - µ2 de dos distribuciones normales) así estas pruebas y otras que se verán mas adelante hacen una fuerte suposicion acerca de la distribución que subyace a las poblaciones estudiadas, lo que permite inferir sobre los parámetros de estas distribuciones. Lo mismo ocurre con la pruebas de Wald y Wilson estudiadas en el capitulo 7 las cuales contrastan hipótesis en torno a un cierto valor para el paramtro p de una población binomial (o la diferencia de los parámetros p1 – p2 de dos de estas poblaciones)

En este capitulo conoceremos pruebas para inferir acerca de proporciones cuyas hipótesis nula y alternativa **no mencionan parámetro alguno.** Es más, **ninguna de ellas hace alguna suposición sobra la distribución de la población** de donde proviene la muestra analizada. Es por esta razón que a estas pruebas se les denomina **no paramétricas o libres de distribución.**

las pruebas no paramétricas son ofrecen una ventaja evidente: **son menos restrictivas** que las pruebas paramétricas, porque imponen menos supuestos a las poblaciones para poder trabajar con ellas. Asegurar que una población sigue una distribución normal o binamial, por ejemplo, puede ser una tarea difícil y en la practica, no es infrecuente encontrarse con conjuntos de datos que no parecen seguir alguna de estas distribuciones. Pero si son tan ventajosas estas pruebas ¿por qué no usarlas siempre?

Por dos razones:

1. Las pruebas no paramétricas **nos entregan menos información**, como veremos en este cap el caso de las proporciones estas pruebas se limitan a trabajar con hipótesis del tipo “las poblaciones muestran las mimas proporciones” versus “las poblaciones muestran proporciones distintas”, pero **no se indica cuales serian esas proporciones**  en la realidad, ni siquiera si es mayor en una o en la otra.
2. Cuando si se cumplen las condiciones para aplicar una prueba paramétrica, las versiones no paramétricas presentan menor poder estadístico y en consecuencia , suelen necesitar muestras de mayor tamaño para detectar diferencias significativas que pudieran existir entre las poblaciones comparadas.

En este capitulo estudiaremos pruebas no paramétricas para estudiar la relación entre dos variables categóricas.

**8.1 PRUEBA CHI-CUADRADO DE PEARSON**

Conocida como prueba *X2*  de asociación, la prueba chi-cuadrado de Pearson sirve para inferir con proporciones cuando disponemos de dos variables categóricas (Sexo, mujer=1 y hombre=0, Medio de pago , efectivo, credito) y una de ellas es dicotómica (es decir que tiene solo dos niveles). En este caso podemos registrar las frecuencias observadas para las posibles combinaciones de ambas variables mediante una tabla de contingencia o matriz de confusión. Como ya estudiamos en el capitulo 2. En adelante nos referiremos a cada combinación como un grupo.

Verificar condiciones para Chi-cuadrado:



1. Las observaciones son independientes entre si
2. Debe haber a lo menos 5 observaciones esperadas en cada grupo

Texto

Descripción generada automáticamente

8.1.1 Prueba chi-cuadrado de homogeneidad

Esta prueba resulta adecuada si queremos determinar si **dos poblaciones** (la variable dicotómica) presentan **las mimas proporciones en los diferentes niveles de una variable categórica.**

**Ejemplo:** se hace una encuesta a 300 programadores del todo el país, escogidos al azar y se les ha preguntado cual es su lenguaje de programación favorito. La tabla muestra las preferencias para cada lenguaje separadas en programadores y programadoras ¿son similares las preferencias de programación entre hombres y mujeres?

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Si fuera cierto que ambas poblaciones tienen las mimas frecuencias, esperaríamos encontrar proporciones similares en las muestras, pese a la variabilidad. En consecuencia, necesitamos determinar si las diferencias entre las cantidades observadas y las esperadas son lo suficientemente grandes como para proporcionar evidencia convincente de que las preferencias son disimiles. La tabla 8.2 muestra las frecuencias esperadas para cada lenguaje bajo este supuesto, calculadad mediante la ecuación 8.1 donde:

Imagen que contiene Texto

Descripción generada automáticamente

Tabla

Descripción generada automáticamente



Ahora como ya sabemos como determinar la cantidad de observaciones esperadas en cada grupo, podemos verificar que, para cada caso este valor es mayor que 5. Adicionalmente, es razonable suponer que la muestra representa menos el 10% de los programadores del país y sabemos que fue seleccionada de manera aleatoria por lo que podemos proceder con la prueba X2 de homogeneidad.

Formulación de hipótesis:

H0: programadores hombres y mujeres tienen las mismas preferencias en lenguaje de programación (ambas poblaciones muestran las mismas proporciones para cada lenguaje estudiado)

HA: programadores hombres y mujeres tienen preferencias distintas en lenguajes de programación favorito.

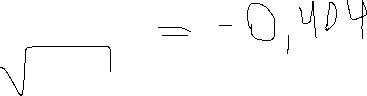
Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Imagen que contiene Aplicación

Descripción generada automáticamente

Así para los programadores varones en C se tiene:



Se repite el procedimiento para ca grupo, se obtienen los valores Z que se presentan en la siguiente tabla

Tabla

Descripción generada automáticamente

Pero necesitamos transformar estos estadísticos para cada grupo en un único estadístico de prueba. Para ello se considera la suma de sus cuadrados, pues así todos los valores son positivos y las diferencias significativas se incrementan aun mas (como el caso de la varianza). Así se define el estadístico de prueba *X2*, definido en la ecuación 8.2 donde m y n son respectivamente la cantidad de filas y la cantidad de columnas de la tabla de frecuencias sin considerar los totales.

Texto

Descripción generada automáticamente

Como estamos sumando m\*n valores Z al cuadrado, **el estadístico** X2 **sigue una distribución chi-cuadrado con Ѵ=(m-1)\*(n-1) grados de libertad. En el ejemplo Ѵ=(2-1)\*(5-1)=4**

El valor p para la prueba chi-cuadrado esta dado por el área bajo la curva de la distribución chi-cuadrado con valores mayores al obtenido para el estadístico de prueba. En este caso, gracias a la llamada en R **pchisq(1.611, df=4, lower.tail=FALSE)**  obtenemos que p=0.807. suponiendo un nivel de significación de α=0.05 , p > α , por lo que se falla al rechazar la hipótesis nula. Es decir no hay evidencia suficientemente fuerte que sugiera con 95% de confianza, que programadores hombres y mujeres prefieren lenguajes de programación distintos.

En R podemos realizar la prueba chi-cuadrado de homogeneidad como muestra el script 8.1,usando para ello la función **chisq.test(x) , donde x corresponde a la matriz de confusión. Al ejecutar el script, debemos tener en cuenta que el valor p obtenido en R es ligeramente diferentes debido a los redondeos aplicados en la tabla 8.2 y al resolver la ecuación 8.2**

**8.1.2 PRUEBA CHI-CUADRADO DE BONDAD DE AJUSTE**

Esta prueba **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.

Ejemplo: una empresa de 660 programadores, el gerente selecciono un subconjunto de 55 programadores, supuestamente de forma aleatoria para cursos de capacitación. El sindicato lo acusa de haber seleccionado estas personas a su conveniencia de los intereses mezquinos de la gerencia, impidiendo que el grupo sea representativo . ante el inminente riesgo de movilizaciones el gerente necesita demostrar que el grupo seleccionado es una muestra representativa de sus programadores.

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Paso 1. Verificar condiciones

1. La muestra representa meno del 10% de la población y fue elegida de manera aleatoria, las observaciones son independientes entre si.
2. La segunda opción es mas compleja, comenzamos por calcular la proporción de programadores de la nomina (población) especialista en lenguaje, para el caso de C tenemos

Texto

Descripción generada automáticamente con confianza media

En este ejemplo: las hipótesis a contrastar son:

H0: las proporciones de especialistas en cada lenguaje son las mismas para la nomina y la muestra

HA: las proporciones de especialistas en cada lenguaje son diferentes en la nomina que en la muestra.

Se puede proceder de la misma forma que para la prueba de bondad de ajuste, para este ejmplo p = 0.461 por lo que se falla al rechazar la hipótesis nula con un nivel de significación α = 0.05. en consecuencia, podemos concluir con 95% de confianza que la muestra seleccionada es, en efecto, representativa de la nomina de programadores de la empresa, por lo que la acusación del sindicato no tiene fundamentos.

**8.1.3 PRUEBA CHI-CUADRADO DE INDEPENDENCIA**

esta prueba permite **determinar si dos variables categóricas, de una misma población, son estadísticamente independientes o si, por el contrario, están relacionadas**

ejemplo: un micologo desea determinar si existe relación entre la forma del sombrero de los hongos y si estos son comestibles o no . se recolecta una muestra de 8120 hongos y se obtiene la siguiente tabla de contingencia.

Tabla

Descripción generada automáticamente

Verificar condiciones:

1. Podemos suponer que la muestra fue obtenida de forma aleatoria ya que se trata de un estudio publicado en una revista científica y representa a menos del 10% de la población mundial de hongos. Se verifica la condición de independencia de las observaciones de las muestras
2. Ahora debemos determinar cuantas observaciones esperaríamos tener en cada grupo si las variables fueran independientes. En este caso la frecuencia esperada para cada celda esta dada por la ecuación:

Texto

Descripción generada automáticamente con confianza baja

Definición de hipótesis:

H0: las variables clase y forma del sombrero son independientes

HA: las variables clase y forma del sombrero están relacionadas.

Al ejecutar en R el script, obtenemos que el valor para el estadístico de prueba es *X2*=485,64 con Ѵ=4 grados de libertad y con un valor p < 2.2\*10-16 . aun para un nivel de significación muy exigente como α=0.01, el valor p obtenido nos permite rechazar la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, concluimos con un 99% de confianza que las variables clase y forma del sombrero están relacionadas (son dependientes).

**8.2 PRUEBAS PARA MUESTRAS PEQUEÑAS**

La prueba chi-cuadrado nos pide que las observaciones esperadas para cada grupo sean a lo menos 5. Sin embargo, hay escenarios donde esta condición no se cumple, por lo que debemos recurrir a una alternativa

**8.2.1 PRUEBA EXACTA DE FISHER**

**La prueba exacta de Fisher**  es una alternativa a la prueba de *X2* de independencia en el caso de que **ambas variables sean dicotómicas.**  Así las hipótesis a contrastar son:

H0: las variables son independientes

HA: las variables están relacionadas

En este escenario, las frecuencias de la muestra pueden resumirse en una tabla de contingencia de 2x2.

Si se asumen independencia entre ambas variables y los totales por filas y columnas son fijos, la **probabilidad exacta de observar el conjunto de frecuencias de la tabla 8.8 esta dada por la ecuación 8.4**

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Lleva el nombre de exacta porque internamente construye todas las tablas posibles con los mismos totales marginales que recibe como entrada y para cada una de ellas, determina la probabilidad exacta de observarla. El valor p corresponde en este caso a la suma de las probabilidades de todas las tablas con probabilidad menos o igual a la tabla dada.

Ejemplo: estudio desea determinar si dos vacunas Argh y Grrr, son igualmente efectivas para inmunizar a la población ante una mordida de vampiro. Para ellos los investigadores reclutaron 17 voluntarios de todo el mundo, de los cuales 6 recibieron la vacuna Argh y 11 Grrr. Al pasar 3 meses los voluntarios fueron mordidos por un vampiro y observaron qur ninguno vacunado con Argh resulto afectado mientras que 5 con Grr se convirtieron en vampiros.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Diagrama, Gráfico de cajas y bigotes

Descripción generada automáticamente con confianza media

Considerando un nivel de significación α=0.05, se falla al rechazar la hipótesis nula. En consecuencia, se concluye con un 95% de confianza que no hay una asociación estadísticamente significativa entre la cantidad de nuevos vampiros y la vacuna recibida.

En R podemos usar la función **Fisher.test(x,conf.level),** donde x corresponde a la tabla de contingencia y conf.leves el nivel de confianza.

**8.2.2 PRUEBA DE mcNemar**

esta prueba resulta apropiada cuando una misma característica , con respuesta dicotimica, se mide en dos ocasiones diferentes para los mismos sujetos (muestras pareadas) y queremos determinar si se produce un cambio significativo en las mediciones. Una vez mas podemos registras las frecuencias en una matriz de confusión como la que vimos en 8.8. en ella podemos ver que las celdas a y d corresponden a instancias en que no hay cambios. La celda b de dicha tabla representa a las intanscias que cambian de presenta a ausente y la celda c, a instancias que cambian de ausente a presente.

Hipótesis para la prueba de McNemar:

H0: **no**  hay cambios significativos en las respuestas

HA: **si** hay cambios significativos en las respuestas

Texto

Descripción generada automáticamente

Para ilustrar el funcionamiento de la prueba de mcNemar, suponga que un cientista de datos ha construido dos modelos para predecir, a partir de las notas obtenidas en cursos previos. Si sus estudiantes aprobaran o no la asignatura de aprendizaje automático. Al probar sus modelos los 25 estudiantes del semestre anterior, observo que predijeron el resultado final de cada estudiante como muestra la tabla 8.11 y se resume en la matriz de confusión tabla 8.12

El cientista de datos desea saber si existe diferencia entre el desempeño de ambos algoritmos por lo que decide emplear la prueba de mcNemar. Al calcular el estadístico de prueba (con el factor de corrección) obtiene:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamenteel valor p esta dado por el área bajo la cola superior de la distribución chi-cuadrado, que en R puede calcularse como **pchisq(0.083,1,lower.tail=FALSE)** obteniendo que p =0.773. en consecuencia, se falla al rechazar la hipótesis nula (para un alfa de 0.05) y se concluye que no hay diferencia en el desempeño de ambos clasificadores.

8.3 PRUEBA Q DE COCHRAN

La **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 mas de dos observaciones pareadas (cuando ambas variables son dicotómicas, esta prueba es equivalente a la de mcNemar).

Ejemplo: Elsa, tiene como tarea determinar si existe una diferencia significativa en el desempeño de tres metaheurísticas que buscan resolver el problema del vendedor viajero. Para ello, el profesor le ha proporcionado los datos representados en la tabla 8.13 , donde la columna instancia identifica cada instancia del problema empleada para evaluar las metaheurísticas y las restantes columnas indican si la metaheurística encontró en cuestión encontró (1) o no (0) la solución optima para dicha instancia.

Tabla

Descripción generada automáticamente con confianza baja

Las hipótesis constrastadas por la prueba Q de Cochran son:

H0: la proporción de éxitos es la misma para todos los grupos

HA: la proporción de éxitos es distinta para al menos un grupo

Se deben cumplir las condiciones:

1. La variable de respuesta es dicotómica
2. La variable independiente es categórica
3. Las observaciones son independientes entre si
4. El tamaño de la muestra es lo suficientemente grande. Sugiere n\*k>=24, donde n es el tamaño de la muestra (la cantidad de instancias, para el ejemplo) y k la cantidad de niveles de la variable independite

El estadístico de prueba se calcula como:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Al ejecutar el script, obtenemos el resultado que se muestra en la figura 8.1. tenemos que el valor p es p=0.028 menor al nivel de significación α = 0.05, por lo que rechazamos la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, Elsa concluye con 95% de confianza que al menos una de las metaheurísticas tiene un desempeño diferente a las demás.

Debemos mencionar que la hipótesis nula de la prueba de Q de Cochran no es especifica, sino que comprueba la igualdad de todas las proporciones. Esta clase de hipótesis suele llamarse **OMNIBUS** (en ocasiones también colectiva o global). Así se dice que la prueba Q de Cochran es una prueba ómnibus porque tiene esta clase de hipótesis nula, con la dificultad de que solo detecta si existe al menos un bloque con una proporción de “éxito” diferente. Desde luego, existen métodos para responder a esta ultima pregunta llamados **pruebas post-hoc** o tambien **a posteriori**. Reciben este nombre porque se realizan una vez que se ha concluido gracias a la prueba ómnibus que existen diferencias significativas.

Algo importante que debemos recordar: **solo haremos un procedimiento post-hoc si la prueba ómnibus rechzada la hipótesis nula en favor de la hipótesis alternativa**. Además el procedimiento post.hoc realizado debe considerar el mismo nivel de significación que la prueba ómnibus.

El procedimiento post-hoc consiste en efectuar pruebas de mcNemar entre cada par de bloques. R nos permite realizar esto mediante la función **pairwiseMcnemar(formula, data,method)**  del paquete rcompanion, donde formula y data son las mismas que para la prueba Q de Cochran y method nos permite determinar el método para ajustar los valores p de las comparaciones, pero ¿por qué queramos ajustar los valores?

Como explican Geoman y Solari, cuando contrastamos hipótesis acotamos la probabilidad de cometer errores de tipo 1 por medio del nivel de significación α. Sin embargo, cuando hacemos múltiples contrastes de hipótesis simultáneamente, cada uno de ellos tendra una probabilidad α de cometer error de tipo 1. Esto se traduce en un **incremento de la probabilidad de cometer este tipo de errores** a medida que aumenta la cantidad de hipótesis contrastadas y en consecuencia una reducción del poder estadístico.

Muchos factores de corrección tienen por objeto distribuir el nivel de significación empleado para la prueba ómnibus en cada par de bloques. El método mas sencillo para ajustar los valores p es **la corrección de Bomferroni.** como explica la ayuda en R, esta corrección simplemente multiplica el valor p obtenido en cada prueba por la cantidad de pruebas realizadas. En general, no se recomienda el uso del método de Bomferroni, especialmente si el numero de grupos es alto, pues es considerado muy conservador, lo que significa que mantiene la probabilidad de cometer un error de tipo 1 mas baja que el nivel de significación establecido (por ende es mas propensa a cometer errores de tipo 2)

Otra alternativa es la **correccion de Holm** , con mayor poder estadistico que la de Bomferroni. esta correccion comienza por efectuar las pruebas entre pares de bloques y luego ordena los valores p en forma creciente. A continuación se calcula el factor de Holm, HB, para cada par de bloques

Texto

Descripción generada automáticamente con confianza baja

Luego se compara el valor p con su respectivo valor de Holm y si el valor p es menor, se considera que existe una diferencia significativa. R implementa esta correccion de manera diferente, de modo que el valor p ajustado pueda ser comparado con el nivel de significación original.

Imagen que contiene Texto

Descripción generada automáticamente