

Numéricas: pueden tomar muchos valores numéricos, y son sensibles a operaciones aritméticas.

- Continuas: pueden tomar cualquier valor (en un intervalo) del conjunto de los reales. Por ejemplo, las variables estatura y servicio descritas en la tabla 1.2.
- Discretas: no es posible que tomen cualquier valor (en un intervalo). Por ejemplo, podrían tomar únicamente valores enteros no negativos, como la variable antigüedad de la matriz de datos ffaa.

Categorías: solo pueden tomar un valor de entre un conjunto acotado. Cada posible valor se denomina nivel.

- Nominales: no existe un orden natural entre los niveles. Ejemplos de variables nominales son género y rama de la matriz de datos ffaa.
- Ordinales: existe un orden natural entre los niveles

Escala nominal: sirve solo para separar un conjunto de elementos en subclases excluyentes entre sí. Los valores no son más que nombres o estados, por lo que no podemos hacer operaciones aritméticas ni podemos establecer relaciones de orden.

Escala ordinal o de rangos: esta escala, 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, por lo que tampoco podemos hacer operaciones aritméticas. Pero en este caso sí podemos establecer una relación de orden, aunque para ello es necesario que la variable tenga a lo menos tres niveles. A modo de ejemplo, si queremos una variable para medir el nivel de estudios de las personas en un grupo demográfico, podríamos considerar una escala ordinal con los niveles “ninguna”, “básica completa”, “media”, “superior” y “postgrado”.

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 que la distancia entre dos valores cualesquiera de la escala es conocida y constante, por lo que podemos emplear operaciones aritméticas. Aunque el punto cero y la unidad de medida son arbitrarios, la razón entre dos intervalos es independiente de ambos elementos. Tomemos, por ejemplo, la escala Celsius de temperatura. El cero está dado por el punto de congelación del agua. La medida o tamaño se calcula en base a los puntos de congelación y ebullición del agua. Sin embargo, a pesar de estos parámetros arbitrarios, el cambio en la cantidad de calor es el mismo si aumentamos la temperatura de 10 a 15 grados Celsius, o de 25 a 30. Si miramos ahora la escala Fahrenheit de temperatura, los puntos fijos son diferentes a los empleados por la escala Celsius, por lo que el cero no significa lo mismo. Sin embargo, existe una transformación lineal que nos permite transformar una medida en una escala a su equivalente en otra escala.

Escala de razón: cumple con todos los atributos de la escala de intervalos, pero además tiene su origen en un cero verdadero. Ejemplos de tales escalas son, por ejemplo, las que permiten medir la masa o la distancia. En una escala de razón, la diferencia entre dos puntos es independiente de la unidad de medida. Por ejemplo, si medimos la masa de dos objetos, la razón es constante independientemente de si empleamos kilogramos, libras u onzas (a diferencia de lo que ocurre con la temperatura usando las escalas Celsius y Fahrenheit).

La estadística usa los datos para responder diversas preguntas, muchas de las cuales se orientan a encontrar relaciones entre variables. Así, 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:

Asociación positiva: si una variable crece, la otra también lo hace.

Asociación negativa: si una variable crece, la otra decrece.

- (a) **Texto plano delimitado por tabulaciones.**
datos2 <- read.delim (file . choose ())
- (b) **Valores separados por comas (inglés).**
datos3 <- read.csv("C:\\Inferencia \\ ejemplo1 -csv -eng.csv ")
- (c) **Valores separados por punto y comas (español).**
datos4 <- read.csv2 (" ejemplo1 -csv -esp.csv ")

Variables aleatorias

```
# Calcular el valor esperado .
10 esperado <- E(X)
11 cat (" Valor esperado :", esperado , "\n")
12
13 # Calcular la varianza .
14 varianza <- V(X)
15 cat (" Varianza :", varianza , "\n")
16
17 # Calcular la desviación estándar .
18 desviacion <- SD(X)
19 cat (" Desviación estándar :", desviacion , "\n")
```

Es importante señalar que, por defecto, lower.tail toma el valor verdadero, con lo que pnorm() y qnorm() operan con la cola inferior de la distribución. Si, en cambio, lower.tail = FALSE, dichas funciones operan con la cola superior (es decir, pnorm() nos entrega la probabilidad de que la variable tome valores mayores que un valor dado).

Distribución Z

Su valor z, que determina cuán por encima o por debajo de la media (en términos de la desviación estándar) se encuentra dicha observación x. Así, observaciones cuyos valores z sean negativos estarán por debajo de la media. Análogamente, un valor Z positivo indica que la observación está por sobre la media. Mientras mayor sea el valor absoluto de su valor z (|z|), más inusual será la observación.

$$Z = \frac{x - \mu}{\sigma}$$

Distribución chi-cuadrado

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.

dchisq gives the density
pchisq gives the distribution function
qchisq gives the quantile function
rchisq generates random deviates.

Distribución t de Student

El único parámetro de esta distribución corresponde a los grados de libertad.

dt gives the density

Distribución F

Usada para comparar varianzas.

Se requiere dos distribuciones χ^2 y dos grados de libertad respectivamente.

df gives the density

Distribución de Bernoulli

Una variable aleatoria 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.

Se requiere que los datos sean independientes.

dbern gives the density

Distribución geométrica

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

La probabilidad de obtener un éxito al n -ésimo intento.

dgeom gives the density

Distribución binomial

Describe la probabilidad de tener exactamente k éxitos en n intentos independientes de Bernoulli con probabilidad de éxito p .

dbinom gives the density

Antes de decidir usar la distribución binomial, tenemos que verificar cuatro condiciones:

1. Los intentos son independientes.
2. La cantidad de intentos (n) es fija.
3. El resultado de cada intento puede ser clasificado como éxito o fracaso.
4. La probabilidad de éxito (p) es la misma para cada intento.

Distribución binomial negativa

Describe la probabilidad de encontrar el k -ésimo éxito al n -ésimo 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 cuántos intentos se necesitan para observar una cantidad fija de éxitos y se requiere que la última observación sea un éxito.

dnbinom gives the density

Como adelanta la comparación anterior, antes de decidir usar la distribución binomial negativa tenemos que verificar cuatro condiciones:

1. Los intentos son independientes.
2. El resultado de cada intento puede ser clasificado como éxito o fracaso.
3. La probabilidad de éxito (p) es la misma para cada intento.
4. El último intento debe ser un éxito.

Distribución de Poisson

Útil para estimar la cantidad de eventos en una población grande en un lapso de tiempo dado, por ejemplo, la cantidad de contagios de influenza entre los habitantes de Santiago en una semana.

dpois gives the density

Prueba bilateral

H_0 : El nuevo sistema, en promedio, tarda lo mismo que el antiguo en procesar las transacciones, es decir: $\mu_{UN} = \mu_{UA}$.

H_A : Los sistemas requieren, en promedio, cantidades de tiempo diferentes para procesar las transacciones, es decir: $\mu_{UN} \neq \mu_{UA}$

Prueba unilateral

H_0 : El nuevo sistema tarda, en promedio, lo mismo que el antiguo en procesar las transacciones, es decir: $\mu_{UN} = \mu_{UA}$.

H_A : El nuevo sistema tarda, en promedio, menos que el antiguo en procesar las transacciones, es decir: $\mu_{UN} < \mu_{UA}$

$$Z^* = q_{norm}$$

El error tipo I corresponde a rechazar H_0 cuando en realidad es verdadera, mientras que el error tipo II corresponde a no rechazarla cuando en realidad H_A es verdadera.

Prueba formal de hipótesis con valores p

Cuanto menor sea el valor p, más fuerte será la evidencia en favor de H_A por sobre H_0 . Y aquí la ventaja de usar este método para decidir: el valor p se puede comparar directamente con el nivel de significación α , 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, $p = 0.040 < \alpha = 0.05$, por lo que se rechaza H_0 en favor de H_A . Pero como se dijo cuando usamos intervalos de confianza, el valor p está cerca del valor α y convendría ser menos tajante en la decisión y evaluar la posibilidad de ampliar la muestra para conseguir evidencia más definitiva.

Como $p > \alpha$, se falla en rechazar la hipótesis nula en favor de la hipótesis alternativa, por lo que se concluye con un 95% de confianza. Se acepta H_0 .

Como $p < \alpha$, se rechaza la hipótesis nula en favor de la hipótesis alternativa, por lo que se concluye con un 95% de confianza. Se acepta H_A .

El efecto del nivel de significación

Si resulta costoso o peligroso cometer un error de este tipo, debemos requerir evidencia más fuerte para rechazar la hipótesis nula (es decir, reducir la probabilidad de que esto ocurra), lo que podemos lograr usando un valor más pequeño para el nivel de significación, por ejemplo, $\alpha = 0.01$. Sin embargo, esto necesariamente aumentará la probabilidad de cometer un error de tipo II. Si, por el contrario, el costo o el peligro de cometer un error de tipo II (no rechazar H_0 cuando en realidad H_A es verdadera) es mayor, debemos escoger un nivel de significación más elevado (por ejemplo, $\alpha = 0.10$).

Estimadores puntuales con distribución cercana a la normal

además de la media, cuya distribución muestral es cercana a la normal si las muestras son lo suficientemente grandes, tales como las proporciones y la diferencia de medias.
el estimador puntual $\hat{\theta}$ debe ser insesgado.

Prueba de hipótesis usando el modelo normal:

1. Formular la hipótesis nula (H_0) y alternativa (H_A) en lenguaje llano y luego en notación matemática.
2. Identificar un estimador puntual (estadístico) adecuado e insesgado para el parámetro de interés.
3. Verificar las condiciones para garantizar que la estimación del error estándar sea razonable y que la distribución muestral del estimador puntual siga aproximadamente una distribución normal.
4. Calcular el error estándar. Luego, graficar la distribución muestral del estadístico bajo el supuesto de que H_0 es verdadera y sombrear las áreas que representan el valor p.
5. Usando el gráfico y el modelo normal, calcular el valor p para evaluar las hipótesis y escribir la conclusión en lenguaje llano.

PRUEBA Z

Condiciones:

- La muestra debe tener al menos 30 observaciones. Si la muestra tiene menos de 30 observaciones, 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 las otras.
- La población de donde se obtuvo la muestra sigue aproximadamente una distribución normal. Con la prueba Q-Q o shapiro.test(x).

$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{26,066 - 20}{2,32} = 2.6147$$

2 * pnorm(2.6147, lower.tail = FALSE), al ser bilateral se multiplica el valor por dos.

z.test(x,mu, stdev, alternative, conf.level)

```
prueba <- z.test (media , mu = valor_nulo , alternative = "two.sided ",stdev = desv_est , conf.level = 1- alfa )
```

PRUEBA T DE STUDENT

$$\bar{x} \pm t_v^* \cdot SE$$

intervalo de confianza

SE= desv estándar/ raíz n

1. Prueba t para una muestra

Condiciones:

No tiene un mínimo de observaciones.

Las observaciones son independientes entre sí.

Las observaciones provienen de una distribución cercana a la normal.

t.test(x,alternative, mu, conf.level)

2. Prueba t para dos muestras pareadas

```
prueba_2 <- t.test (x = t_A,  
                    y = t_B,  
                    paired = TRUE ,  
                    alternative = "two.sided ",  
                    mu = valor_nulo ,  
                    conf.level = 1 - alfa )
```

3. Prueba t para dos muestras independientes

```
prueba <- t.test (x = vacuna_A,  
                  y = vacuna_B,  
                  paired = FALSE ,  
                  alternative = "greater ",  
                  mu = 0,  
                  conf.level = 1 - alfa )
```

PODER ESTADÍSTICO

Error tipo I: rechazar H_0 en favor de H_A cuando H_0 es en realidad verdadera.

Error tipo II: no rechazar H_0 en favor de H_A cuando H_A es en realidad verdadera.

Poder estadístico = 1-beta que es la probabilidad de correctamente rechazar H_0 cuando es falsa.

- El poder de la prueba aumenta mientras mayor es el tamaño del efecto (en este caso, la distancia entre el valor nulo y la media de la muestra).
- A medida que el tamaño del efecto disminuye (es decir, el estimador se acerca al valor nulo), el poder se aproxima al nivel de significación.
- Usar un valor de α más exigente (menor), manteniendo constante el tamaño de la muestra, hace que la curva de poder sea más baja para cualquier tamaño del efecto (lo que verifica la relación entre α y β).
- Usar una muestra más grande aumenta el poder de la prueba para cualquier tamaño del efecto distinto de 0.

TAMAÑO DEL EFECTO

$$d = \frac{\bar{x} - \mu_0}{s}$$

d de Cohen

PODER, TAMAÑO DEL EFECTO Y TAMAÑO DE LA MUESTRA

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.

CÁLCULO TEÓRICO DEL PODER

CÁLCULO DEL PODER EN R

`power.t.test(n, delta, sd, sig.level, power, type, alternative)`

`lower.tail=` por la izquierda

INFERENCIA CON PROPORCIONES MUESTRALES

Métodos de Wald y de Wilson para inferir acerca de una y dos proporciones.

MÉTODO DE WALD

Condiciones:

1. Las observaciones de la muestra son independientes.
2. Se cumple la condición de éxito-fracaso, que establece que se espera observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos.

En consecuencia, podemos asumir que la distribución muestral de \hat{p} sigue aproximadamente a la normal.

Método de Wald para una proporción

```
error_est_hip <- sqrt((valor_nulo * (1 - valor_nulo)) / n)
```

```
Z <- (p_exito - valor_nulo) / error_est_hip
```

```
p <- pnorm(Z, lower.tail = FALSE)
```

Método de Wald para dos proporciones

Condiciones:

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

```
p <- pnorm(Z, lower.tail = FALSE)
```

MÉTODO DE WILSON

Método de Wilson para una proporción

```
prop.test(exitos, n = n, p = valor_nulo, alternative = "greater", conf.level = 1 - alfa)
```

Esta función tiene la limitante de que, al trabajar con dos proporciones, no permite establecer un valor nulo distinto de cero para la diferencia.

Método de Wilson para dos proporciones

prueba <- prop.test (exitos , n = n, alternative = "two.sided ", conf.level = 1 - alfa)

PODER Y PRUEBAS DE PROPORCIONES

power.prop.test(n, p1, p2, sig.level, power, alternative)

- pwr.p.test(h, n, sig.level, power, alternative): para pruebas con una única proporción.
- pwr.2p.test(h, n, sig.level, power, alternative): para pruebas con dos proporciones donde ambas muestras son de igual tamaño.
- pwr.2p2n.test(h, n1, n2, sig.level, power, alternative): para pruebas con dos proporciones y muestras de diferente tamaño.

El tamaño del efecto puede calcularse como muestra la ecuación 7.6, implementada en R en la función ES.h(p1, p2) del paquete pwr.

Otra función que nos puede ser de ayuda es bsamsize(p1, p2, fraction, alpha, power), del paquete Hmisc. En el caso de una prueba de Wilson con dos muestras, calcula los tamaños de cada grupo.

INFERENCIA NO PARAMÉTRICA CON PROPORCIONES

En este capítulo conoceremos algunas pruebas para inferir acerca de proporciones cuyas hipótesis nula y alternativa no mencionan parámetro alguno. Es más, ninguna de ellas hace alguna suposición sobre la distribución de la población desde donde proviene la muestra analizada. Es por esta razón que a estas pruebas (y a otras que se abordan en capítulos posteriores) se les denomina no paramétricas o libres de distribución.

son menos restrictivas

Pero... si las pruebas no paramétricas parecen tan ventajosas, ¿por qué no usarlas siempre?

Por dos grandes razones:

- Las pruebas no paramétricas nos entregan menos información. Como veremos en este capítulo para el caso de las proporciones, estas pruebas se limitan a trabajar con hipótesis del tipo “las poblaciones muestran las mismas proporciones” versus “las poblaciones muestran proporciones distintas”, pero ninguna indica cuáles serían esas proporciones en realidad, ni siquiera si es mayor en una o en la otra.
- Cuando sí 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.

PRUEBA CHI-CUADRADO DE PEARSON

Condiciones:

1. Las observaciones deben ser independientes entre sí.
2. Debe haber a lo menos 5 observaciones esperadas en cada grupo.

Prueba chi-cuadrado de homogeneidad

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

pchisq(1.611, df = 4, lower.tail = FALSE)

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.

prueba <- chisq.test (tabla , correct = FALSE)

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.

PRUEBAS PARA MUESTRAS PEQUEÑAS

Hemos visto que la prueba chiz 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 alguna

alternativa.

Prueba exacta de Fisher

La prueba exacta de Fisher es una alternativa a la prueba χ^2 de independencia en el caso de que ambas variables sean dicotómicas. Así, las hipótesis a contrastar son:

H₀: las variables son independientes.

H_A: las variables están relacionadas.

`fisher.test(x, conf.level)`

Prueba de McNemar

Esta prueba resulta apropiada cuando una misma característica, con respuesta dicotómica, se mide en dos ocasiones diferentes para los mismos sujetos (muestras pareadas) y queremos determinar si se produce o no un cambio significativo entre ambas mediciones.

Las hipótesis asociadas a la prueba de McNemar son:

H₀: no hay cambios significativos en las respuestas.

H_A: sí hay cambios significativos en las respuestas.

`pchisq(0.083, 1, lower.tail = FALSE)`

`prueba <- mcnemar.test(tabla)`

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 más de dos observaciones pareadas.

Las hipótesis contrastadas por la prueba Q de Cochran son:

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

H_A: la proporción de “éxitos” es distinta para al menos un grupo.

Como ya debemos suponer, esta prueba también requiere que se cumplan algunas condiciones:

1. La variable de respuesta es dicotómica.
2. La variable independiente es categórica.
3. Las observaciones son independientes entre sí.
4. El tamaño de la muestra es lo suficientemente grande. Glen (2016a) sugiere que $n \cdot k \geq 24$, donde n es el tamaño de la muestra (la cantidad de instancias, para el ejemplo) y k , la cantidad de niveles en la variable independiente.

`cochran.qtest(formula, data, alpha = 0.05)`

En este punto, debemos mencionar que la hipótesis nula de la prueba Q de Cochran no es específica, sino que comprueba la igualdad de todas las proporciones. Esta clase de hipótesis nula suele llamarse ómnibus (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. Sin embargo, de ser afirmativa la respuesta, no nos dice qué grupos presentan diferencias (Lane, s.f.). Desde luego, existen métodos para responder a esta última pregunta, llamados pruebas post-hoc, o también a posteriori. Reciben este nombre porque se realizan una vez que se ha concluido gracias a la prueba ómnibus que existen diferencias significativas.

`pairwiseMcnemar(formula, data, method) method= Bonferroni o Holm`