Curso de métodos estadísticos dirigido al estudio de tests paramétricos y no paramétricos

Manuel Oviedo de la Fuente (manuel.oviedo@usc.es)
Beatriz Pateiro López (beatriz.pateiro@usc.es)
Universidade de Santiago de Compostela

CITIUS - Centro Singular de Investigación de Tecnoloxías da Información

Contenidos

Contenidos: Análisis exploratorio de datos

▶ Índice del curso

Conceptos básicos

Las tareas vinculadas a la Estadística se clasifican en tres grandes disciplinas:

- **Estadística Descriptiva**: Se ocupa de recoger, clasificar y resumir la información contenida en la muestra.
- **Cálculo de Probabilidades**: Es una parte de la matemática teórica que estudia las leyes que rigen los mecanismos aleatorios.
- **Inferencia Estadística**: Pretende extraer conclusiones para la población a partir del resultado observado en la muestra.

Conceptos básicos

En general, no se puede conocer la totalidad del universo o población a estudiar (o su análisis resulta inviable en términos de tiempo o económicos). Por lo tanto, los parámetros de interés de la población son, en general, desconocidos.

La **Estadística descriptiva** tiene como objetivo sintetizar la información contenida en un conjunto de datos ofreciendo un resumen numérico o gráfico del estado de las cosas.

- **Población**: Colectivo de individuos sobre los que se quiere extraer alguna conclusión.
- **Individuos**: Cada uno de los elementos de la población (unidad estadística).
- **Muestra**: Subconjunto (representativo) de la población, que seleccionamos con el objetivo de extraer información.

Conceptos básicos

La **variables estadísticas** son las características que se pueden observar o estudiar en los individuos de la población.

- **Cualitativas nominales**: Miden características que no toman valores numéricos (color del pelo, raza, tipo de pétalo,...).
- **Cualitativas ordinal**: Presentan entre sus posibles valores una relación de orden (grado de contaminación, calificación,...).
- **Cuantitativa discreta**: Toman un número discreto de valores (nº de hermanos, nº de materias,...).
- **Cuantitativa continua**: Toman valores numérico dentro de un intervalo real (peso, altura, anchura del pétalo, longitud y latitud,...).

Conceptos básicos

- **Tablas de frecuencias**:
  Las tablas de frecuencias se utilizan para representar la información contenida en una muestra de tamaño n extraida de una población, $(x_1, \ldots, x_n)$.
- **Modalidades**:
  Cada uno de los valores que puede tomar una variable (cualitativa o cuantitativa discreta), $c_i, i = 1, \ldots, k$.

  El número de individuos de la muestra en cada modalidad c_i se denota por n_i.

Conceptos básicos

Frecuencia absoluta:

Para cada modalidad c_i, la frecuencia absoluta es $n_i,\ i = 1, \ldots, k$.

Frecuencia relativa:

Para cada modalidad c_i, la frecuencia relativa es $f_i = n_i/n, i = 1, \ldots, k$.

Frecuencia absoluta acumulada:

La frecuencia absoluta acumulada de una modalidad c_i es $N_i = \sum_{j=1}^{i} n_j = n_1 + \ldots + n_i,\ i = 1, \ldots, k$.

Frecuencia relativa acumulada:

La frecuencia relativa acumulada de una modalidad c_i es $F_i = \sum_{j=1}^{i} f_j = f_1 + \ldots + f_i = \frac{N_i}{n},\ i = 1, \ldots, k$.

Modalidad	Frecuencia absoluta	Frecuencia relativa	Fr. abs. acumulada	Fr. rel. acumulada
c_1	n_1	f_1	N_1	F_1
c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_i	n_i	f_i	N_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_k	n_k	f_k	$N_k = n$	$F_k = 1$
TOTAL	n	1		

Table: Ejemplo de tabla de frecuencias.

Tablas de frecuencias

El conjunto de datos Titanic que contiene 4 variables cualitativas nominales de los 2201 pasajeros y tripulantes que se corresponden a: la clase del pasajero (1a, 2a, 3a y tripulación), edad (niño/adulto), supervivencia (si/no) y el sexo (hombre/mujer).

El siguiente código muestra la tabla de frecuencias para el conjunto de datos Titanic

```
> data(Titanic)
> fabs <- apply(Titanic, 1, sum)  # Frecuencia absoluta
> frel <- fabs/sum(fabs)  # Frecuencia relativa
> facum <- double(length(fabs))  # Se crea un vector de tamaño el de fabs
> for (i in 1:length(fabs)) {
+     facum[i] <- sum(fabs[1:i])
+ }  # Frec.Abs.Acum.
> facumrel <- facum/sum(fabs)  # Frecuencia Relativa Acumulada
> data.frame(fabs, frel, facum, facumrel)

     fabs   frel facum facumrel
1st   325 0.1477   325   0.1477
2nd   285 0.1295   610   0.2771
3rd   706 0.3208  1316   0.5979
Crew  885 0.4021  2201   1.0000
```

Tablas de frecuencias

- **Intervalos de clase:** para variables cuantitativas continuas, se agrupan en intervalos los distintos valores obtenidos en la muestra . Cada intervalo representará una *modalidad* en el caso de variables cuantitativas continuas.

1. Denotamos por $e_0 < e_1 < \ldots < e_k$ los extremos de los k intervalos de clase. Ejemplo de intervalo: (e_{i-1}, e_i).
2. Amplitud del intervalo: $a_i = e_i - e_{i-1}$.
3. Marca de clase: $c_i = \dfrac{e_{i-1} + e_i}{2}$.
4. Algunas cuestiones sobre los intervalos de clase:
   ¿Cuántos intervalos podemos construir? Entre 5 y 20. Una regla muy utilizada es hacer $k = \sqrt{n}$.
   ¿Siempre de la misma amplitud? Es deseable (salvo que otra información aconseje que sea de distinta amplitud).

Tablas de frecuencias

El conjunto de datos airquality dispone de medidas de calidad del aire en Nueva York con las variables cuantitativas Ozone (ozono en ppb), Solar.R (radiación solar en langleys), Wind (viento en mph), Temp (temperatura en F).

```
> # Ejemplo de cálculo de frecuencias en variables continuas
> data(airquality)
> attach(airquality)
> f <- cut(Temp, 5)  # Asigna datos numéricos uno de los 5 grupos
> tapply(Temp, f, length)  # Calcula la frec. abs.

  (56,64.2] (64.2,72.4] (72.4,80.6] (80.6,88.8]   (88.8,97]
         16          23          46          49          19

> table(f)  # Calcula la frec. abs.

f
  (56,64.2] (64.2,72.4] (72.4,80.6] (80.6,88.8]   (88.8,97]
         16          23          46          49          19

> # (Repetir pasos anteriores para obtener otras frecuencias)
```

Métodos gráficos para variables cualitativas

- Diagrama de barras
- Diagrama de sectores

Métodos gráficos para variables cuantitativas discretas

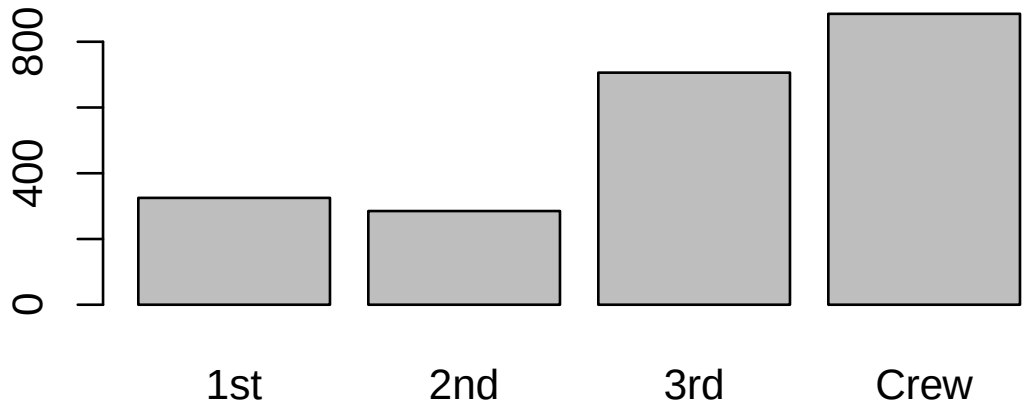
- Diagrama de barras
- Diagrama acumulativo de frecuencias

Métodos gráficos para variables cuantitativas continuas

- Histograma
- Diagrama de cajas (boxplot)

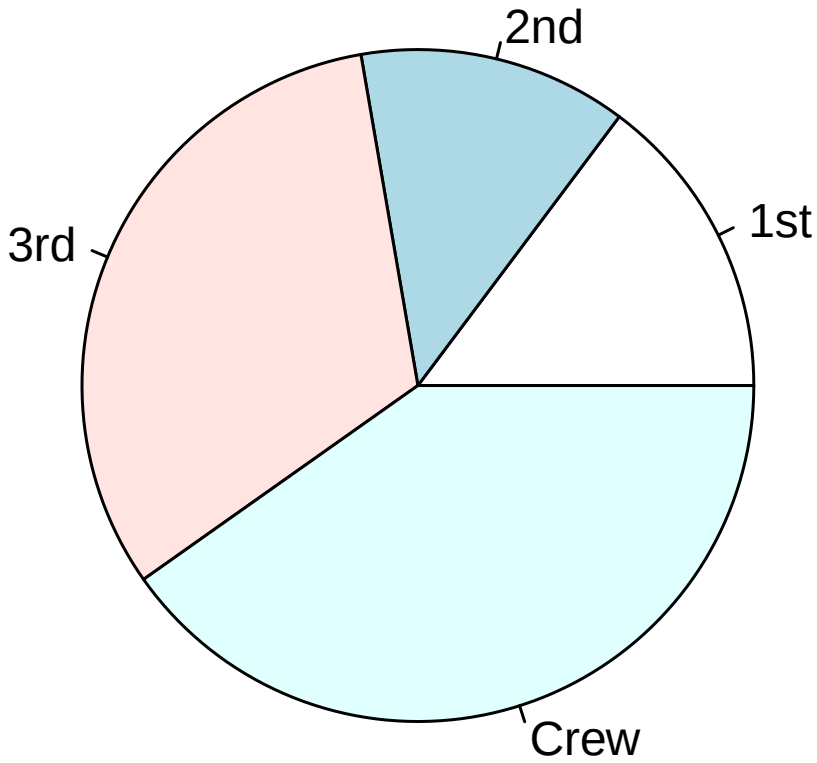
Variables cualitativas: Diagrama de barras

```
> barplot(apply(Titanic, 1, sum))
```



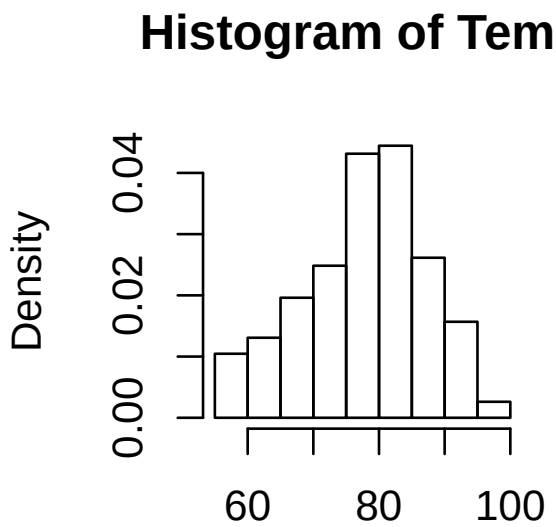
Variables cualitativas: Diagrama de sectores

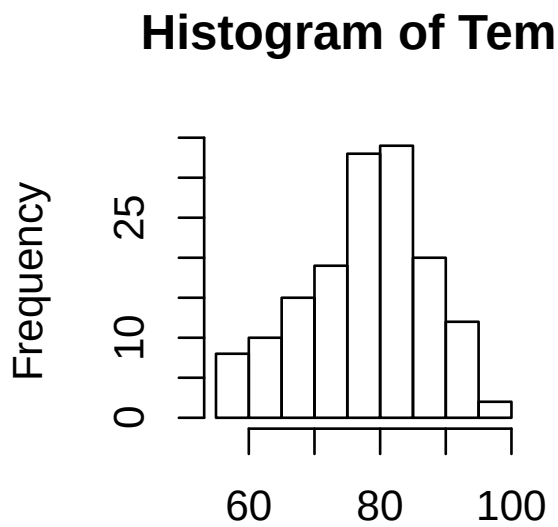
```
> pie(apply(Titanic, 1, sum))
```



Variables cuantitativas continuas: Histograma

```
> # rm(list=ls())
> attach(airquality)

The following object is masked from airquality (position 3):

    Day, Month, Ozone, Solar.R, Temp, Wind

> n <- length(Temp)
> par(mfrow = c(1, 2))
> hist(Temp, xlab = "(F)", freq = FALSE, breaks = floor(sqrt(n)))
> hist(Temp, xlab = "(F)", freq = TRUE, breaks = floor(sqrt(n)))
```





Medidas de centralización: Su objetivo es obtener un representante del conjunto de los datos.

Media aritmética: Se define la media aritmética (o simplemente media) como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La media aritmética tiene interesantes propiedades:

1. Entre el mínimo y el máximo:

$$\min\{x_1, \ldots, x_n\} \leq \bar{x} \leq \max\{x_1, \ldots, x_n\}$$

2. Tiene las mismas unidades que los datos originales.
3. Media de las desviaciones con respecto a la media es cero (centro de gravedad de los datos) $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$
4. Media de los cuadrados de las desviaciones con respecto a una constante es mínima para la media.
5. Lineal. $y_i = a + bx_i \Rightarrow \bar{y} = a + b\bar{x}$

Problema: no es una medida robusta (media recortada, media truncada).

Media truncada

Truncar la media consiste en calcular la media aritmética de un porcentaje central de los datos.

Una media truncada al 10% calcularía la media aritmética del 90% de los valores centrales despreciando el 5% de los valores más bajos y el 5% de los más altos.

```
> mean(Wind)

[1] 9.958

> mean(Wind[-2])

[1] 9.97

> mean(Wind, trim = 0.05)

[1] 9.896

> mean(Wind, trim = 0.5)

[1] 9.7
```

Mediana

Si los datos están ordenados de menor a mayor, la mediana es el valor hasta el cual se encuentran el 50% de los casos.

- Si n es impar, la mediana será el dato central.
- Si n es par, entonces se tomará como mediana la media de los dos datos centrales.

Si los datos se han agrupado, se determina primero el intervalo mediano (aquel intervalo donde la frecuencia relativa acumulada es menor o igual que 0,5 en su extremo inferior y mayor que 0,5 en su extremo superior).

La mediana sería la medida de posición central más robusta (más insensible a datos anómalos).

```
> median(Wind)

[1] 9.7

> mean(Wind, trim = 0.5)

[1] 9.7

> (table(f))/sum(table(f))  #cumsum(table(f))/sum(table(f))

f
  (56,64.2] (64.2,72.4] (72.4,80.6] (80.6,88.8]   (88.8,97]
     0.1046      0.1503      0.3007      0.3203      0.1242
```

Moda

Para variables discretas o cualitativas, la moda es el valor o valores que más se repiten.

- La moda no tiene porqué ser única.
- Si los datos se encuentran agrupados, se puede obtener el intervalo modal como aquel que tiene una mayor frecuencia.

```
> f <- cut(Temp, 5)
> table(f)

f
  (56,64.2] (64.2,72.4] (72.4,80.6] (80.6,88.8]   (88.8,97]
         16          23          46          49          19
```

Valores no observados en los datos (NA)

```
> mean(Ozone, na.rm = T)

[1] 42.13

> mean(Ozone[!is.na(Ozone)])

[1] 42.13

> Ozone2 <- na.omit(Ozone)
> mean(Ozone2)

[1] 42.13

> # Ozone2 na.action(Ozone2)
```

Otras medidas de posición

Cuartiles: los cuartiles Q_1, Q_2 y Q_3 dividen la muestra en cuatro partes iguales.

Deciles: $d_1, \ldots, d_9$ dividen la muestra en 10 partes iguales (intervalos del 10%).

Percentiles: $p_1, \ldots, p_{99}$ dividen la muestra en 100 partes iguales (intervalos del 1%)

Cuantiles en general, para cualquier $0 < p < 1$.

```
> quantile(Ozone2, probs = c(0, 0.25, 0.5, 0.75, 1))

    0%    25%    50%    75%   100%
  1.00  18.00  31.50  63.25 168.00

> summary(Ozone2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    18.0    31.5    42.1    63.2   168.0
```

Varianza y desviación típica

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Propiedades de la varianza:

1. Valores no negativos
2. No linealidad
3. Otra forma de calcular:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

```
> n <- length(Ozone2)
> (n - 1)/n * var(Ozone2)  # Varianza

[1] 1079

> sqrt((n - 1)/n) * sd(Ozone2)  # Desviación estandar

[1] 32.85
```

Otras medidas de dispersión

Desviación absoluta respecto a la media: $D_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$.

Desviación absoluta respecto a la mediana: $D_{Q_2} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Q_2|$.

Recorrido o rango: $R = max(x_i) - min(x_i)$.

Rango intercuartílico: $RI = Q_3(x) - Q_1(x)$.

Coeficiente de variación: El coeficiente de variación es una medida de dispersión relativa (no depende de las unidades de los datos):

$$CV = \frac{s}{\bar{x}}, \quad \bar{x} > 0$$

```
> mean(abs(Ozone2 - mean(Ozone2)))  # Desv. Abs.

[1] 26.35

> mean(abs(Ozone2 - median(Ozone2)))  # Desv. Absoluta Mediana

[1] 24.89

> range(Ozone2)  # Varianza

[1]   1 168

> sd(Ozone2)/mean(Ozone2)  # CV

[1] 0.783
```

Medidas de forma

Las medidas de forma tratan de medir el grado de simetría y apuntamiento en los datos.

1. **Coeficiente de asimetría de Fisher**: Toma valor 0 cuando la distribución de los datos es simétrica con respecto a la media.

$$\gamma_F = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

   - Valores positivos: asimetría positiva
   - Valores negativos: asimetría negativa

2. **Coeficiente de curtosis**: El coeficiente de curtosis mide el grado de apuntamiento de la distribución.

$$\gamma_C = \frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

   - Valores > 3: distribución leptocúrtica (apuntada)
   - Valores < 3: distribución platicúrtica (achatada)

```
> momento.centrado <- function(x, orden) {
+     x.new <- x[!is.na(x)]
+     mean((x.new - mean(x.new))^orden)
+ }
> momento.centrado(Ozone2, 3)/sd(Ozone2, na.rm = T)^3  # Asimetría de Fisher

[1] 1.21

> momento.centrado(Ozone2, 4)/sd(Ozone2, na.rm = T)^4  # Curtosis

[1] 4.112
```

Diagramas de caja (Boxplot)

Los diagramas de caja se construyen a partir de las siguientes medidas:

- El primer y el tercer cuartil, Q_1 y Q_3, que delimitan la caja central. La longitud de la caja viene dada por el RI, que es una medida de dispersión absoluta.
- Los límites inferior y superior se calculan como:

$$LI = \max\{\min\{x_i\}, Q_1 - 1.5(Q_3 - Q_1)\},$$

$$LS = \min\{\max\{x_i\}, Q_3 + 1.5(Q_3 - Q_1)\}.$$

  En el cálculo de los límites inferior y superior se utiliza el $RI = Q_3 - Q_1$.
- La mediana (Q_2) se representa con una línea horizontal en la caja central.

```
> data(airquality)
> boxplot(Temp)
```

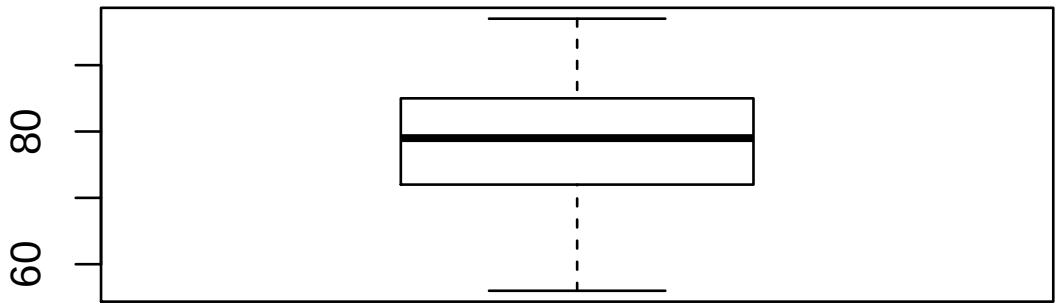


Diagrama de cajas por grupos

```
> boxplot(Temp ~ Month)
```

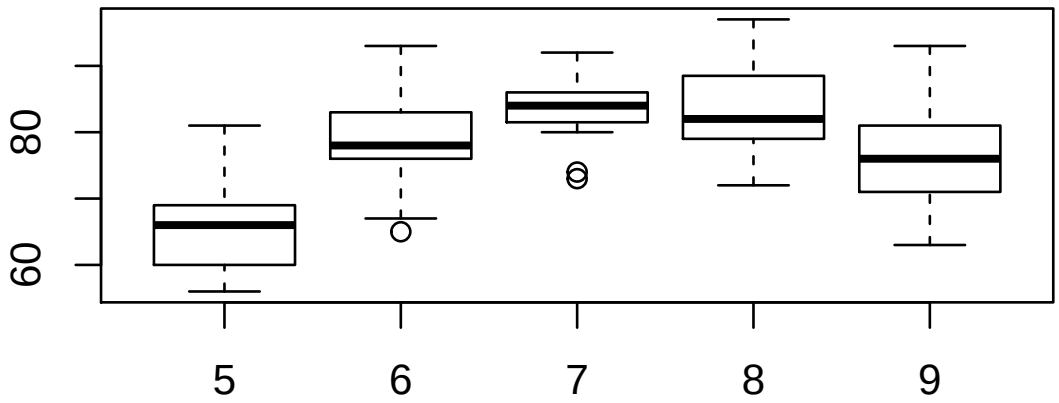


Diagrama de dispersión o nube de puntos

Se representan las parejas de datos (x_i, y_i) con $i = 1, \ldots, n$, de las dos variables (X, Y) (también llamada variable bidimensional).

```
> plot(Temp, Ozone, xlab = "Temp", ylab = "Ozone")
```

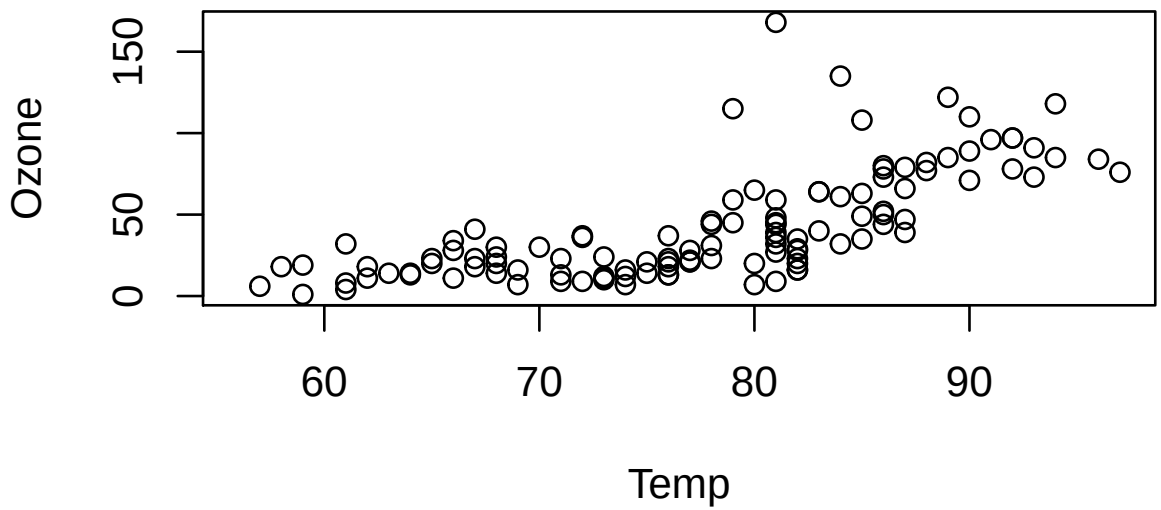
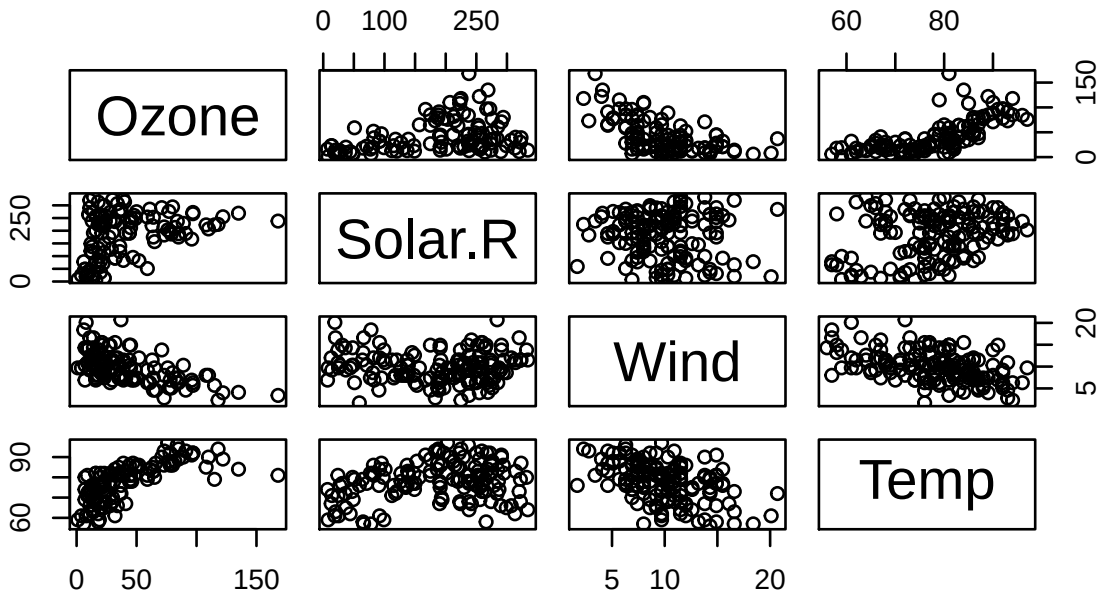


Diagrama de dispersión

```
> pairs(airquality[, 1:4])
```



Covarianza

La covarianza entre dos variables S_{xy} es una medida que indica la variabilidad conjunta de X e Y y se calcula como:

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y}$$

Relación entre variables y signo de la covarianza

- Si la relación entre las variables es directa, entonces $S_{xy} > 0$.
- Si la relación entre las variables es inversa, entonces $S_{xy} < 0$.
- Si no hay relación lineal entre las variables, entonces $S_{xy} = 0$.

Coeficiente de correlación lineal

A partir de una muestra de datos $\{(x_i, y_i)\}_{i=1}^n$, el coeficiente de correlación lineal se calcula como:

$$r = \frac{S_{xy}}{s_x s_y},$$

donde S_{xy} es la covarianza muestral y s_x, s_y son las respectivas desviaciones típicas muestrales.

- No tiene dimensiones
- Toma valores en $[-1, 1]$
- Si no existe relación lineal entre las variables, $r = 0$

Coeficiente de correlación lineal

```
> cor(Temp, Wind)

[1] -0.458

> ind.na <- numeric()
> for (i in 1:4) ind.na <- c(ind.na, na.action(na.omit(airquality[,
+     i])))
> cor(airquality[-unique(ind.na), 1:4])

          Ozone Solar.R    Wind    Temp
Ozone    1.0000  0.3483 -0.6125  0.6985
Solar.R  0.3483  1.0000 -0.1272  0.2941
Wind    -0.6125 -0.1272  1.0000 -0.4972
Temp     0.6985  0.2941 -0.4972  1.0000
```

Otras herramientas para el análisis descriptivo

- NA, outiliers, library Chron, datos repetidos
- CPU time, código C, Fortran
- image, persp, contour

script descriptivo.r

Tipificación de datos

Si tenemos una muestra $x_1, \ldots, x_n$ con media $\bar{x}$ y varianza s^2, los datos tipificados se construyen como:

$$z_i = \frac{x_i - \bar{x}}{s}$$

La muestra resultante $z_1, \ldots, z_n$ tendrá media 0 y varianza 1. La tipificación de datos permite comparar la posición relativa de las observaciones dentro de cada grupo.

Transformación Box-Cox

$$X^{(\lambda)} = \frac{(X+m)^{(\lambda)} - 1}{\lambda} \; si \; \lambda \neq 0,$$

$$X^{(\lambda)} = ln(X+m) \; si \; \lambda = 0,$$

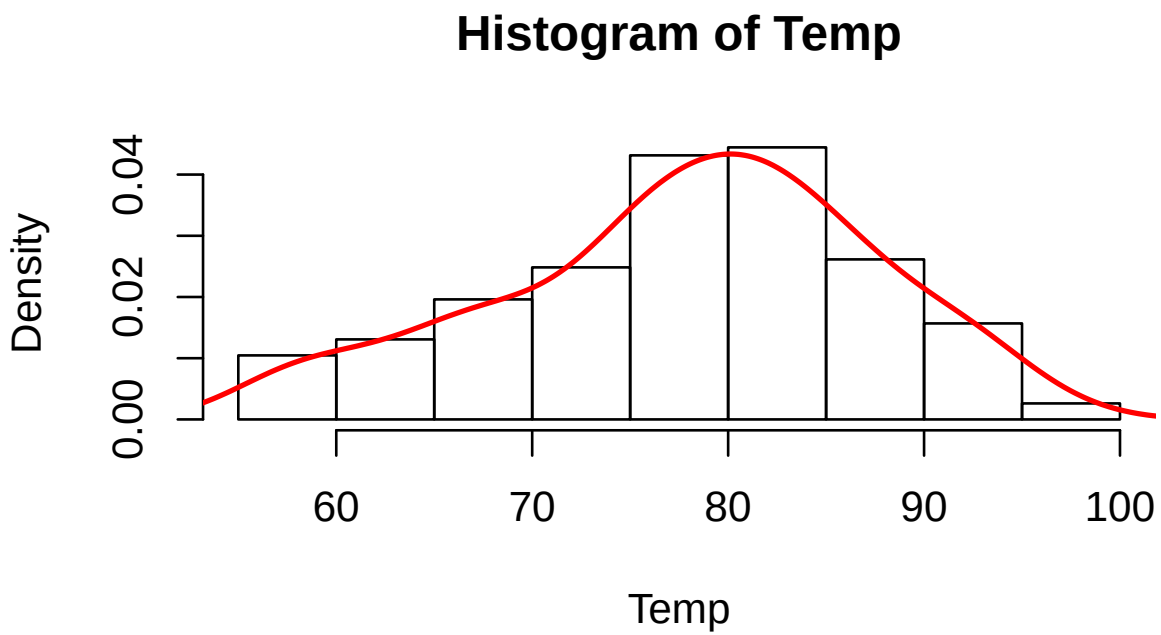
siendo $X + m > 0$.

Para valores de $\lambda > 1$ se corrigen asimetría a la izquierda

Para valores $\lambda < 1$ se corrigen asimetría a la derecha.

Alternativa al histograma: la función de densidad

```
> library(sm)
> hist(Temp, freq = F)
> lines(density(Temp), col = 2, lwd = 2)  # Kernel Density Estimation
```



Regresion lineal y regresion no paramétrica (suavizado por Kernel)

Recta de regresión

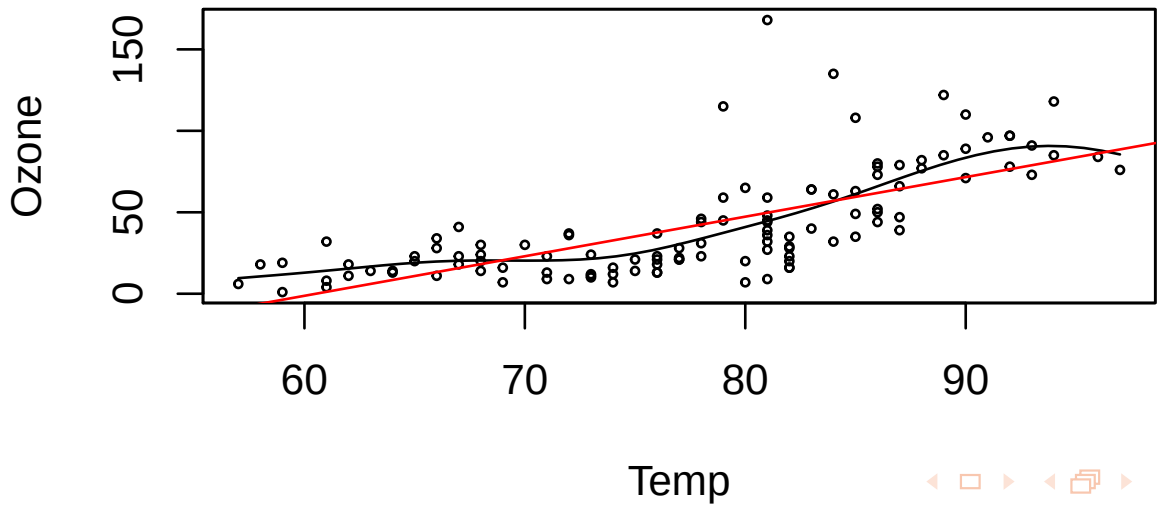
La recta de regresión de Y sobre X tendrá la siguiente expresión:

$$y = a + bx,$$

donde a representa la ordenada en el origen o intercepto y b es la pendiente. X se denomina variable explicativa o independiente e Y será la variable respuesta, o variable dependiente.

Regresion lineal y regresion no paramétrica (suavizado por Kernel)

```
> a <- lm(Ozone ~ Temp, data = airquality, na.action = na.exclude)
> # library(sm)
> b <- sm.regression(x = Temp, y = Ozone)

missing data are removed

> abline(a = a$coef[1], b = a$coef[2], col = 2)
> # ?sm.regression
```



CPU time

```
> system.time(apply(toeplitz(1:1000), 1, mean))

   user  system elapsed
   0.15    0.00    0.15

> system.time(colMeans(toeplitz(1:1000)))

   user  system elapsed
   0.01    0.01    0.03
```

Análisis de datos descriptivos multivariante

- **Análisis de Correspondencias** basado en la matriz de frecuencias.
- **Análisis de Componentes Principales** basado en la matriz de covarianzas.
- **Análisis de Discriminante** basado en la matriz de distancias.
- **Análisis Clúster** basado en la matriz de distancias.

script descriptivo.r

Análisis de Correspondencias (Matriz de frecuencias)

El Análisis de Correspondencias estudia la relación entre dos variables discretas a través de su distribución conjunta de frecuencias. A la tabla formada por dicha distribución conjunta se le suele denominar **tabla de contingencia**.

Si el Análisis de Correspondencias pretende estudiar la relación entre las dos variables, la situación extrema de no relación se producirá cuando las dos variables sean independientes. Esto es equivalente a que la distribución conjunta resulte del producto de las marginales, y se puede contrastar mediante el test ji–cuadrado.
Ejemplo de usos notables

- Epidemiología: Enfermos-Sanos, Fumador-No fumador.
- Estudios socio-económicos: Nivel de estudios-Ingresos.

```
> ct <- corresp(~Age + Eth, data = quine)
> # corresp(caith) biplot(corresp(caith, nf = 2))
```

Análisis de Componentes Principales (ACP)

El ACP es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos, se pasa de la **matrices de datos**, que contiene toda la información recogida sobre el fenómeno de estudio, a unas representaciones (visuales) de aquella información.
El ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados mediante la descomposición en autovalores de la matriz de covarianza.
Usos:

- Análisis exploratorio de datos
- Modelos predictivos: regresión y clasificación

- Inconveniente: pérdida de información (que intentaremos minimizar), los datos se tienen que distribuir de manera gaussiana.
- Ventaja: ganancia en significación, reducimos la dimensión de los datos a unas pocas CP que el usuario puede representar e interpretar.

Análisis de Componentes Principales (ACP)

- *svd*(): Singular Value Decomposition of a Matrix, $X = UDV'$
- *eigen*() Spectral Decomposition of a Matrix, $X = V\lambda V^{-1}$

```
> hilbert <- function(n) {
+     i <- 1:n
+     1/outer(i - 1, i, "+")
+ }
> X <- hilbert(6)[, 1:6]
> s <- svd(X)
> D <- diag(s$d)
> range(X - s$u %*% D %*% t(s$v))  #  X = U D V')

[1] -1.665e-16  2.220e-16

> e <- eigen(X)  # para matrices cuadradas
> range(e$vectors %*% (diag(e$values)) %*% solve((e$vectors)) -
+     X)

[1] -4.996e-16  1.388e-16

> # y sin el numero de variables mayor que el tamaño muestral
> # p>n?
```

Análisis discriminante

El análisis discriminante o clasificación supervisada es un tema fundamental en la Estadística

El objetivo es encontrar las localizaciones y el número de grupos alrededor de los cuales se agrupan los datos, "El arte de encontrar grupos".

Generalmente el objetivo último es siempre clasificar nuevas observaciones (teniendo en cuenta la información presente en el conjunto de datos de entrenamiento).

Métodos de análisis discriminante

- Vecinos más cercanos, kernel.
- Árboles de clasificación y regresión (CART).
- Modelización de las probabilidades a posteriori como función de regresión: redes neuronales, projection pursuit, discriminación logística.
- Análisis discriminante lineal (generalizaciones): lda, qda, fda.
- Support Vector Machines.

Análisis cluster o clasificacion no supervisada

El principal objetivo del análisis cluster es agrupar un conjunto de observaciones o de datos de forma que cada elemento contenido en el grupo sea lo más parecido posible al resto de elementos del mismo grupo. Además los grupos formados deberán ser lo más distintos posibles los unos de los otros.

En el caso de Análisis cluster aplicado a las variables podemos señalar que mientras que otras técnicas como el Análisis de Componentes Principales utilizan la matriz de correlación de los datos para reducir la dimensión, la mayor parte de técnicas cluster utilizan medidas de distancia para hacer la clasificación.
En el análisis cluster debemos atender a la siguiente distinción en cuanto a los métodos utilizados

- Algoritmos Jerárquicos: dendograma
- Algoritmos del tipo k-medias
- Algoritmos basados en métodos de máxima verosimilitud
- Métodos basados en distancias: distancias entre individuos y medidas de distancias entre grupos de individuos.

Análisis cluster: Métodos jerárquicos

El proceso para la construcción de grupos o clúster en los métodos jeráquicos se basa en la construcción del árbol Jerárquico o Dendograma.

- Elegir entre métodos aglomerativos o divisivos.
- Estrategias para mezclar clúster (o separarlos) (vecino más cercano, vecino más lejano, distancia media,.)
- Decidir por algún método apropiado el número de clústeres que se deben seleccionar.

Análisis cluster: Métodos de partición

- El método de aplicación más usual del algoritmo k-medias trata de buscar la partición de los n individuos en k grupos de manera que se minimize la suma de cuadrados dentro de los grupos sobre todo el conjunto de las variables.
- Normalmente se establece un umbral de cambio en ese término de suma de cuadrados o un número máximo de iteraciones para detener el proceso.
- Aunque parece un proceso sencillo, en realidad es un proceso muy costoso computacionalmente

Técnicas estadísticas notables para el Análisis de Datos Complejos

- Análisis de Datos Funcionales
- Análisis de Datos Direccionales
- Análisis de Datos Composicionales

script descriptivo.r

Análisis de Datos Funcionales (ADF)

El Análisis de Datos Funcionales se ocupa de la modelización estadística de variables aleatorias que toman valores en un espacio de funciones (variables funcionales).

Ejemplo en 1 dimensión, el nivel de Ozono medido en minutos durante un dia, (dato funcional=curva).

Ejemplo en 2 dimensiones, un dato funcional es una superficie.

Métodos de descripitivos

- Media, mediana y varianza funcional.
- Boxplot functional.
- Métricas y semimétricas.
- ACP funcionales.

```
> # library(fda) library(fda.usc)
```

Análisis de Datos Direccionales

Los datos circulares constituyen el caso más simple de esta categoría de datos llamada datos direccionales, donde la medida no es escalar, sino que es angular o direccional.

Para trabajar con datos de esta naturaleza es necesario construir nuevos estadísticos pues los estadísticos usuales empleados para datos lineales son inapropiados, ya que no tienen en cuenta la naturaleza periódica de esta clase de datos.

Por tanto, cada punto x en el círculo unidad puede ser representado por un ángulo θ, $X = (cos\theta, sin\theta)$.

Algunas funciones de distribución utilizadas son: Distribución Uniforme, Cardioide, Normal proyectada, Wrapped, von Mises.

- Medidas de localización
- Medidas de concentración y dispersión
- Estimación de la densidad circular
- Estimación de la regresión circular-lineal

```
> # library(circular) library(NPCirc)
```

Análisis de Datos Composicionales

Cualquier vector x, cuyas componentes representan partes de un todo, está sujeto a la restricción de que la suma de sus componentes sea la unidad, o en el caso general, una constante.

Un dato composicional $x = (x1, \ldots, x_D)$ con D partes, es un vector con componentes estrictamente positivas, tal que la suma de todas ellas es k.

```
> # library(compositions)
```

Contenidos: Métodos de estimación puntual. Métodos de estimación de máxima Verosimilitud.

▶ Índice del curso

Conceptos básicos

Las tareas vinculadas a la Estadística se clasifican en tres grandes disciplinas:

- **Estadística Descriptiva**: Se ocupa de recoger, clasificar y resumir la información contenida en la muestra.
- **Cálculo de Probabilidades**: Es una parte de la matemática teórica que estudia las leyes que rigen los mecanismos aleatorios.
- **Inferencia Estadística**: Pretende extraer conclusiones para la población a partir del resultado observado en la muestra.

Conceptos básicos

- **Población**: Colectivo de individuos sobre los que se quiere extraer alguna conclusión.
- **Individuos**: Cada uno de los elementos de la población (unidad estadística).
- **Muestra**: Subconjunto (representativo) de la población, que seleccionamos con el objetivo de extraer información.

Conceptos básicos

El muestreo aleatorio simple (m.a.s) es aquel en el que cada vez que seleccionamos un individuo de la muestra, este tiene la misma probabilidad de ser elegido para formar parte de la muestra que cualquier otro independientemente de que otros individuos hayan sido ya seleccionados. Por tanto, bajo muestreo aleatorio simple un individuo podría aparecer en la muestra más de una vez (con reemplazamiento). En este caso las variables aleatorias que conforman la muestra pueden suponerse independientes e idénticamente distribuidas (según la distribución poblacional).

Distibuciones continuas en R

- beta: **beta**
- Cauchy **cauchy**
- chi-square **chisq**
- exponential **exp**
- F **f**
- gamma **gamma**
- normal **norm**
- student's t **t**
- uniform **unif**
- Weibull **weibull**

script distribuciones.r

Distibuciones continuas en R

Ejemplo de prefijos para la distribución normal

- **dnorm** densidad
- **pnorm** probabilidad acumulada
- **qnorm** cuantil
- **rnorm** valor aleatorio

Distibuciones discretas en R

- bionimal: **binom**
- geometrica **geom**
- hipergeometrica **hyper**
- binomial negativa **nbinom**
- Poisson **pois**

Distribución muestral y función de verosimilitud

Estamos interesados en el estudio de una variable aleatoria X, cuya distribución, F, es en mayor o menor grado desconocida. Conociendo la distribución podemos extraer conclusiones acerca de la población en estudio. En la inferencia paramétrica suponemos que tenemos una familia de distribuciones cuya distribución de probabilidad se supone conocida salvo los valores que toman ciertos coeficientes (parámetros), es decir, $\mathcal{F} = \left\{F_\theta | \theta \in \Theta \subset R^k\right\}$.

Llamaremos muestra aleatoria simple (m.a.s.) de tamaño n de una variable aleatoria X con distribución teórica F a n variables aleatorias $\{x_1, \ldots, x_n\}$ independientes e igualmente distribuidas con distribución común F.

Llamaremos espacio muestral al conjunto de muestras posibles que pueden obtenerse al seleccionar una muestra de un tamaño determinado de una cierta población. Llamaremos estadístico a cualquier función T de la muestra. El estadístico $T(x_1, \ldots, x_n)$ como función de variables aleatorias, es también una variable aleatoria y tendrá por tanto una distribución de probabilidad que llamaremos distribución en el muestreo de T.

Ejemplos de estadísticos serían

- La media muestral: $T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^{n} X_k$.
- La varianza muestral: $T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{x})^2$.
- La cuasivarianza muestral: $T(x_1, \ldots, x_n) = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{x})^2$.

Notación $\hat{\theta} = T(x_1, \ldots, x_n)$

Propiedades deseables en un estadístico

- Insesgado (centrado): $E(\hat{\theta}_n) = \theta$, con $Sesgo = E(\hat{\theta}_n) - \theta$.
- Asintóticamente insesgado: $lim_{n\to\infty} E(\hat{\theta}_n) = \theta$.
- Consistente en media cuadrática: $lim_{n\to\infty} ECM(\hat{\theta}_n) = 0$ siendo $ECM(\hat{\theta}_n) = Sesgo_{\hat{\theta}_n}(\theta)^2 + Var(\hat{\theta}_n)$.
- Eficiencia: $Efic(\hat{\theta}_n) = \frac{1}{ECM(\hat{\theta}_n)}$.

Función de verosimilitud

Sea una v.a. continua X con función de densidad $f(\bullet|\Theta)$ donde se especifica Θ para indicar que depende de ciertos parámetros desconocidos.

Dada una m.a.s. $x_1, \ldots, x_n$ llamaremos función de verosimilitud a la función de densidad conjunta de la muestra, es decir, $L(\Theta, x_1, \ldots, x_n) = \prod f(x_i|\Theta)$.

El objetivo en la inferencia es obtener información sobre la población a partir de una muestra. En este planteamiento aquel valor Θ que maximice esta función, será el mejor estimador de los parámetros desconocidos de la población. El estimador así construido se llamará estimador máximo-verosímil.

Propiedades de los estimadores máximo-verosímiles para distribuciones cuyos rangos de valores no depende de parámetros:

- Asintóticamente centrados
- Asintóticamente normales
- La varianza de varianza mínima (eficientes)
- Si existe un estadístico suficiente, el estimador máximo–verosímil es suficiente.

Máxima verosimilitud para distibuciones discretas

Ejemplo: Se lanza una moneda al aire ocho veces y se observa la siguiente secuencia

¿Cuál es un buen estimador para p, probabilidad de cara, basado en la secuencia ?

Sea $X_i \sim Bern(p)$, donde $X_i = 1$ indica cara.

$$P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 1, X_8 = 0) =$$

$$P(X_1 = 1)P(X_2 = 1)P(X_3 = 1)P(X_4 = 0)P(X_5 = 1)P(X_6 = 0)P(X_7 = 1)P(X_8 = 0)$$
$$= p^5(1-p)^3$$

Es razonable buscar la p que maximice la función $L(p) = p^5(1-p)^3$: $L'(p) = 0$ se tiene $p = 5/8$.

La función $L(p) = p^5(1-p)^3$ se llama funcion de verosimilitud de parámetro p y $\hat{p} = 5/8$ es la estimación máximo verosimil de p.

Máxima verosimilitud para distibuciones discretas

Sea $f(x;\theta)$ la función de probabilidad para una distribución discreta de parámetro θ. Suponemos que $X_1, \ldots, X_n$ son v.a de esa distribución y $x_1, \ldots, x_n$ son los valores observados. Entonces la función de verosimilitud de θ es

$$\begin{aligned} L(\theta) = L(\theta|x_1, \ldots, x_n) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \mathrm{P}(X_1 = x_1), \ldots, P(X_n = x_n) = f(x_1;\theta), \ldots, f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta). \end{aligned}$$

Sea $X_1, \ldots, X_n$ n v.a. independientes Bernoulli, $Bern(p)$, $0 < p < 1$. El estimador máximo verosimil de p para la realización X número de unos $\mathrm{X}=\sum_{i=1}^{n} X_i es=\mathrm{X/n}$.

$$L(p) = p^x(1-p)^{n-x}$$

Máxima verosimilitud para distibuciones discretas

Ejemplo 2: Observamos $x_1 = 3, x_2 = 4, x_3 = 3, x_4 = 7$ de una distribución de Poisson de parámetro λ desconocido
Función de masa de probabilidad

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \ x = 0, 1, 2, \ldots.$$

Función de verosimilitud

$$L(\lambda) = f(x_1;\lambda)(x_2;\lambda)(x_3;\lambda)(x_4;\lambda) = \frac{\lambda^3 e^{-\lambda}}{3!}\frac{\lambda^4 e^{-\lambda}}{4!}\frac{\lambda^3 e^{-\lambda}}{3!}\frac{\lambda^7 e^{-\lambda}}{7!} = \frac{\lambda^{17} e^{-4\lambda}}{3!4!3!7!}$$

¿L'$(\lambda) = 0$?
Tomar logaritmos, derivar respecto λ e igualar a 0.

$$ln(L(\lambda)) = 17ln(\lambda) - 4\lambda - ln(3!4!3!7!)$$

$$\frac{L'(\lambda)}{L(\lambda)} = 17\frac{1}{\lambda} - 4 = 0$$

El máximo de $\hat{\lambda} = 17/4$ es el valor maximize la función $L(p)$ y corresponde a la media de las cuatro observaciones.

Máxima verosimilitud para la distribución de Poisson

Sea $X_1, \ldots, X_n$ n v.a. de una distribución de Poisson de parámetro λ desconocido. El estimador máximo verosimil de λ es $\hat{\lambda} = \bar{x}$, la media muestral.

Distribución de Poisson en R

Suponemos que un servidor web se cuelga 5 veces al año. Si asumimos que esa frecuencia sigue una distribución de Poisson, ¿Cúal es la probabilidad que la página no se cuelgue en un año?

```
> dpois(0, 5)

[1] 0.006738

> # 5^0*exp(-5)/(1)
```

¿y que se cuelgue tres o más veces?

```
> 1 - (dpois(0, 5) + dpois(1, 5) + dpois(2, 5))

[1] 0.8753

> 1 - ppois(2, 5)

[1] 0.8753
```

Máxima verosimilitud para distibuciones continuas

Sea $X_1, \ldots, X_n$ n v.a. de una distribución continua de parámetro θ desconocido. Función de verosimilitud de θ para una m.a.s X de esa distribución es

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Sea $X_1 = x_1, \ldots, X_n = x_n$ son una m.a.s de una distribución normal $N(\mu, \sigma^2)$ las estimaciones máximo verosimiles de μ y σ son

$$\hat{\mu} = \sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Ejemplo, fábrica de neumáticos

Una frábrica de neumáticos mide el tiempo de vida X (en kms) de 3 tipos de neumáticos de coche: intermedios X_i, blandos X_b y duros X_d.

$$X_i \sim exp(\lambda), X_b \sim exp(0.77/\lambda), X_d \sim exp(1.25/\lambda)$$

Se selecciona de forma aleatoria e independiente un neumático de cada tipo y después del test se observan: $x_i = 28, x_b = 25, x_d = 31$ en miles de kms.

Distribución exponencial: $f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0, E(x) = 1/\lambda, Var(X) = 1/\lambda^2$

Encontrar la estimación máximo verosimil de λ.

$$\begin{aligned} L(\lambda) &= f_i(x_i;\lambda) f_b(x_b;\lambda) f_d(x_d;\lambda) \\ &= \lambda e^{-\lambda x_i}(1.3\lambda e^{-1.3\lambda x_b})(0.8\lambda e^{-0.8\lambda x_d}) \\ &= 1.04\lambda^3 e^{-(X_i + 1.3X_b + 0.8X_d)\lambda} \end{aligned}$$

La log-verosimilitud es $ln(L(\lambda)) = ln(1.04) + 3ln(\lambda) - (X_i + 1.3X_b + 0.8X_d)\lambda$

El estimador máximo verosimil es $\hat{\lambda} = 0.0395$. Por tanto el tiempo de vida esperado (miles de kms) para los tres tipos de neumático es:

$$\hat{E}(X_i) = 25.3, \quad \hat{E}(X_b) = 19.5, \quad \hat{E}(X_d) = 31.6$$

Ejemplo, generación de energía eólica

En un parque eólico la generación de energía de cada turbina depende de la velocidad del viento.

Se observa la velocidad del viento durante 168 dias, ¿Cómo podemos modelar la velocidad del viento?

La distribución de Weibull ha sido muy utilizada para modelar la velocidad del viento.

$$f(x;k,\lambda) = \frac{kx^{k-1}}{\lambda^k} e^{-(x/\lambda)^k}, x \geq 0$$

Queremos encontrar la estimación máximo verosimil de λ y k.

$$L(\lambda, k) = \prod_{i=1}^{n} \frac{kx_i^{k-1}}{\lambda^k} e^{-(x_i/\lambda)^k}$$

Ejemplo, generación de energía eólica

Calcular las derivadas parciales de esta función respecto a λ y k e igualar cada ecuación a 0.

$$\frac{\partial(ln(L(\lambda,k)))}{\partial\lambda} = \frac{-kn}{\lambda} + \frac{k}{\lambda^{k+1}}\sum_{i=1}^{n} {x_i}^k = 0 \qquad (1)$$

se tiene $\lambda^k = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k$

$$\frac{\partial(ln(L(\lambda,k)))}{\partial k} = \frac{n}{k} - n\ ln(\lambda) - \sum_{i=1}^{n} ln(x_i) - \sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^k ln\left(\frac{x_i}{\lambda}\right) = 0 \qquad (2)$$

substituyendo λ^k en (2) nos queda

$$\frac{1}{k} + \sum_{i=1}^{n} ln(x_i) - \frac{1}{\alpha}\sum_{i=1}^{n} x_i^k ln(x_i) = 0$$

donde $\alpha = \sum_{i=1}^{n} x_i^k$.

En ocasiones, no existe una forma cerrada para encontrar la solución del EMV. En este ejemplo, los métodos númericos permiten encontrar una buena aproximación del valor de k y por tanto de λ.

script Turbina.r

Teorema Central del Límite (TCL)

Diremos que una sucesión de variables aleatorias $\{Xn\}_{n=1}^{\infty}$ satisface el Teorema Central del Límite si existe una variable aleatoria X, una sucesión de números reales $\{An\}_{n=1}^{\infty}$, y otra de números positivos $\{Bn\}_{n=1}^{\infty}$ tal que $lim_{n\to\infty} Bn = +\infty$ verificando que

$$\frac{\sum_{k=1}^{n} X_k - A_n}{Bn}$$

Teorema Central del Límite (TCL)

Si $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas con media μ y desviación típica σ, entonces $\sum_{k=1}^{n} X_k$ tiene asintóticamente una distribución normal de media $n\mu$ y desviación típica $\sqrt{n}\sigma$. Por las propiedades de la normal, tendremos que:

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \xrightarrow{n\to\infty} N(0,1)$$

.

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} N(0,1)$$

.

Estimación de la media de una población

Sea $X = \{x_1, \ldots, x_n\}$ una m.a. que sigue una $X \sim N(\mu, \sigma^2)$. Un estimador razonable del parámetro μ es la media muestral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_n$ que verifica las siguientes propiedades:

- Insesgado : $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$
- Consistente: $Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2\right) = \frac{\sigma^2}{n}$

Por tanto, la media muestral

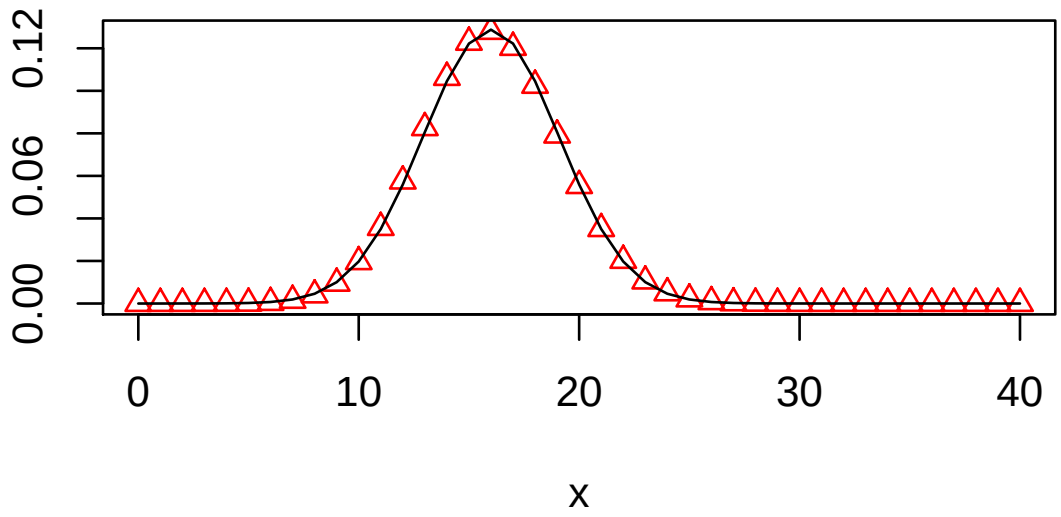
$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

En el caso de que la población de partida no sea normal, por el **teorema central de límite** si el tamaño de la muestra es grande (≥ 30) la distribución de $\bar{x}$ se aproximará a la normal.

Teorema Central del Límite (TCL)

Veamos el Teorema Central del Límite en el caso de la binomial. Como sabemos la binomial de parámetros n y p es la suma de n Bernoullis de parámetro p. Tomamos, $n = 40$ y $p = 0,4$. El valor esperado y la desviación típica de la binomial de parámetros $n = 40$ y $p = 0,4$ son 16 y $3,0983$,respectivamente.

```
> x = seq(0, 40, by = 1)  #Comparamos ambas distribuciones.
> plot(x, dbinom(x, 40, 0.4), pch = 2, col = 2, ylab = "")
> lines(x, dnorm(x, 16, 3.0983))
```



Estimación de la varianza de una población

En este caso, un estimador razonable de la varianza de la población puede ser la varianza muestral $S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{x})^2$ que verifica las siguientes propiedades:

- Es asintóticamente insesgado :

$$E(nS^2) = E\left(\sum_{i=1}^{n}(X_i - \bar{x})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \bar{x})^2\right) = (n-1)\sigma^2$$

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

- Consistente: Bajo normalidad se puede demostrar que $\frac{nS^2}{\sigma^2} \in \chi^2_{n-1}$.

Intervalo de confianza

- **Intervalos de Confianza**. Dado que la estimación puntual conlleva un cierto error, construímos un intervalo que con alta probabilidad contenga al parámetro. La amplitud del intervalo nos da idea del margen de error de nuestra estimación.
- Un intervalo de confianza es un intervalo construido en base a la muestra y, por tanto, aleatorio, que contiene al parámetro con una cierta probabilidad, conocida como **nivel de confianza**.

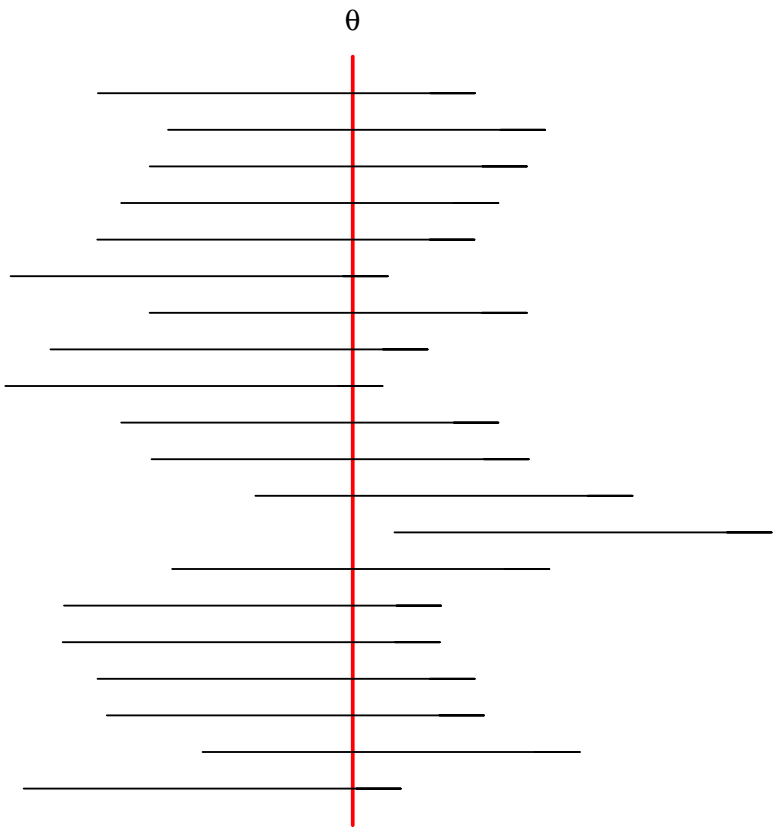
Intervalo de confianza

- Sea θ el parámetro desconocido y $\alpha \in [0, 1]$.
- Se dice que el intervalo $[L_1, L_2]$ tiene un nivel de confianza $1 - \alpha$ si

$$P(L_1 \leq \theta \leq L_2) \geq 1 - \alpha$$

- Los valores de L_1 y L_2 **dependerán de la muestra!!!!**.
- El nivel de confianza con frecuencia se expresa en porcentaje. Así, un intervalo de confianza del 95% es un intervalo de extremos aleatorios que contiene al parámetro con una probabilidad de 0.95.

Interpretación del nivel de confianza $1-\alpha$



- Dada una realización muestral, el intervalo construido puede contener o no al parámetro desconocido
- Esperamos que el $100(1-\alpha)\%$ de los intervalos contengan al parámetro desconocido

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$

- Recordamos que es este caso

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in N(0,1)$$

- Este estadístico (pivote) nos servirá para construir un intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ cuando la **varianza σ^2 es conocida**.

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

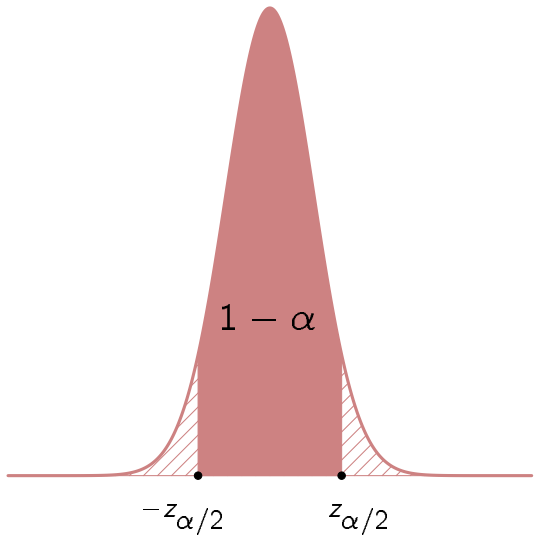
Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$. Supongamos que σ^2 es conocida.

- Sea $z_{\alpha/2}$ el valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$. Entonces:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

- Equivalentemente,

$$P\left(\bar{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



Intervalo de confianza de nivel $1 - \alpha$ para la media μ cuando σ^2 es conocida

$$\left(\bar{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)$$

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es conocida

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Ejemplo: Un investigador está interesado en determinar el nivel medio de determinada proteína en el cuerpo humano. Para ello toma una muestra de 10 individuos y obtiene el nivel de proteína de cada uno de ellos. Los resultados son los siguientes:

$$22, 20, 24, 18, 23, 25, 26, 20, 19, 23$$

- ¿Cómo estimarías el nivel medio de proteína a partir de esta muestra?
- Nuevas investigaciones determinan que la variable de interés es aproximadamente normal con varianza igual a 25. Construye un intervalo de confianza para el nivel medio de proteína en el cuerpo humano con nivel de confianza del 95%.
- ¿Cuál sería el intervalo de confianza para un nivel de confianza del 90%?

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Solución: La media es 22 y la desviación $s = 2.667$, $\alpha/2 = 0.05$. Así pues, el quatil del 95% es $q_{0.95} = 1.645$ tal que $P(Z < 1.645) = 0.95$.

El intervalo de convianza al 95% es

$$\left(22 - 1.645\sqrt{\frac{25}{n}}, 22 + 1.645\sqrt{\frac{25}{n}}\right)$$

Tenemos con un 90% de confianza que el nivel medio de la proteina esá en el intervalo $(19.399, 24.601)$

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Margen de error (ME)

$$ME = q(\sigma/\sqrt{n}).$$

En el ejemplo anterior, el valor μ es deconocido pero con varianza conocida $\sigma^2 = 25$, ¿ Cuantos individuos necesitamos, con una confianza del 95%, para que el margen de error sea como mucho de 1 unidad?

Solución:

$$1.96\sqrt{\frac{25}{n}} \leq 1$$

$$n \geq 96.04$$

$$n \geq 97$$

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.

- En la práctica no es habitual conocer la varianza de la variable de interés.
- Cuando la varianza σ^2 es desconocida, usaremos como estadístico (pivote) para construir un intervalo de confianza para la media μ

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

- Recuerda que:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- En este caso:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \in t_{n-1}$$

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

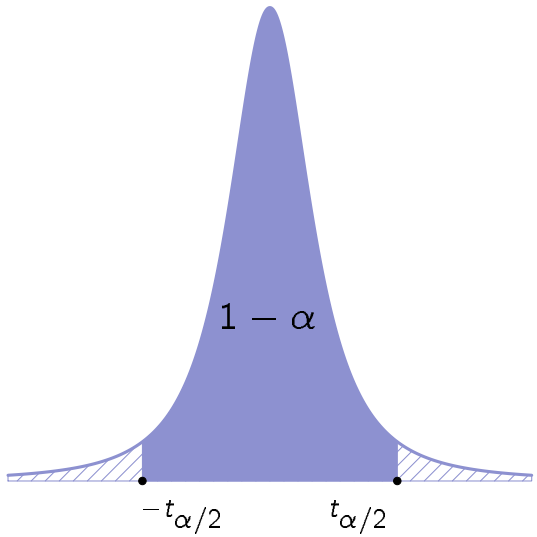
Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$. Supongamos que σ^2 es desconocida.

- Sea $t_{\alpha/2}$ el valor tal que $P(T > t_{\alpha/2}) = \alpha/2$, donde T es una variable t de Student con $n-1$ grados de libertad. Entonces:

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

- Equivalentemente,

$$P\left(\bar{X} - t_{\alpha/2}\,\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\,\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$



Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es desconocida

$$\left(\bar{X} - t_{\alpha/2}\,\frac{S}{\sqrt{n}}\ ,\ \bar{X} + t_{\alpha/2}\,\frac{S}{\sqrt{n}}\right)$$

t de Student con $n-1$ g.l.

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es desconocida

$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

t de Student con $n-1$ g.l.

Ejemplo: Considera las siguientes medidas, correspondientes al Volumen Espiratorio Forzado[1] (litros) de 10 sujetos de un estudio que examina la respuesta al ozono entre adolescentes que sufren asma.

$$3.50, 2.60, 2.75, 2.82, 4.05, 2.25, 2.68, 3.00, 4.02, 2.85$$

- ¿Cómo estimarías el Volumen Espiratorio Forzado medio?
- Construye un intervalo de confianza para el Volumen Espiratorio Forzado medio con nivel de confianza del 95%.
- ¿Cuál sería el intervalo de confianza para un nivel de confianza del 90%?

[1] El Volumen Espiratorio Forzado es la cantidad de aire expulsado durante el primer segundo de la espiración máxima, realizada tras una inspiración máxima

Intervalo de confianza para una proporción p

Intervalo de confianza de nivel $1-\alpha$ para la proporción p

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right)$$

Ejemplo: Una encuesta del proyecto "Pew Internet and American Life Project"[2] llevada a cabo en 2010 determina que el 16% de los usuarios de internet utilizan la red para consultar información sobre resultados de pruebas médicas. La encuesta, que forma parte de un estudio sobre el uso de internet en América, se basa en entrevistas telefónicas a un total de 3001 adultos. Asumimos que los encuestados fueron elegidos de manera aleatoria. Contruye un intervalo de confianza al 95% para la proporción de usuarios de internet que consultan información sobre resultados de pruebas médicas en América.

[2] http://www.pewinternet.org/

Contenidos: Terminología de contrastes de hipótesis

Índice del curso

Conceptos básicos

Las tareas vinculadas a la Estadística se clasifican en tres grandes disciplinas:

- **Estadística Descriptiva**: Se ocupa de recoger, clasificar y resumir la información contenida en la muestra.
- **Cálculo de Probabilidades**: Es una parte de la matemática teórica que estudia las leyes que rigen los mecanismos aleatorios.
- **Inferencia Estadística**: Pretende extraer conclusiones para la población a partir del resultado observado en la muestra.

Conceptos básicos

En general, no se pude conocer la totalidad del universo o población a estudiar (o su análisis resulta inviable en términos de tiempo o económicos). Por lo tanto, los parámetros de interés de la población son, en general, desconocidos.

La **Inferencia Estadística** tiene como objetivo extraer información sobre la población a partir de la obtención de muestras aleatorias

Conceptos básicos

Dependiendo de los objetivos, podremos clasificar las labores de inferencia en dos grandes categorías:

- en la que el interés se centra en **estimar o aproximar el valor de un parámetro**
- en la que el interés se centra en **decidir sobre la veracidad o no veracidad de cierta afirmación realizada sobre la población estudiada**

Hipótesis estadística y contraste de hipótesis

- Una **hipótesis estadística** es una conjetura o una afirmación sobre la población objeto de estudio (sobre la distribución de una o más variables aleatorias).
- Un **contraste de hipótesis** (o un test de hipótesis) se reduce a un problema de decisión sobre la veracidad de una hipótesis estadística.

Aplicaciones

Basándonos en una muestra aleatoria, podremos utilizar la estadística para verificar, por ejemplo:

- Si el tamaño medio de los ficheros de un sistema de archivos es igual a 29.3 KBytes.
- Si el número medio de usuarios concurrentes a una determinada aplicación supera al número medio de usuarios concurrentes a otra aplicación.
- Si la velocidad media de conexión es al menos de 4.500 Kbps, como garantiza un determinado operador de telecomunicaciones.
- Si el número de errores en software es independiente de la experiencia del programador.
- Si el tiempo de respuesta de una CPU sigue una distribución exponencial.
- ...

Contrastes paramétricos y no paramétricos

- **Constrastes paramétricos:** Las hipótesis que contrastamos hacen referencia a parámetros poblacionales. Conocida una variable aleatoria con su distribución, se establecen afirmaciones sobre los parámetros de dicha distribución.
- **Constrastes no paramétricos:** el desconocimiento de la población no se reduce al valor de un parámetro poblacional, sino que es más amplio.

Hipótesis nula y alternativa

- **Hipótesis nula:** es la hipótesis que se da por cierta antes de obtener la muestra. Se denota como H_0.
- **Hipótesis alternativa:** la denotamos por H_1 (o H_a) y es lo que sucede cuando no es cierta la hipótesis nula.

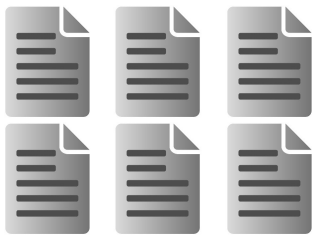
Contraste de hipótesis

- Cuando un investigador trata de entender o explicar algo, generalmente formula su problema de investigación por medio de una hipótesis
- Ejemplo: No sé si el tamaño medio de los ficheros de un sistema de archivos es igual a 29.3 KBytes.

Hipótesis nula

$$H_0 : \mu = 29.3$$

- Tomo una muestra de 6 ficheros del sistema de archivos



- $\bar{X} = 30.5$ KBytes.
- ¿Existe suficiente evidencia en los datos para rechazar H_0?
- ¿O la diferencia entre $\bar{X}$ y el valor hipotético de μ puede ser debido al azar?

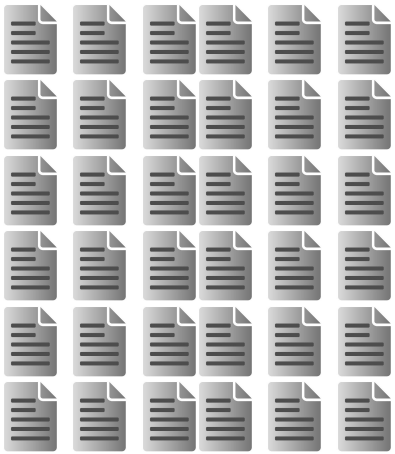
Contraste de hipótesis

- Cuando un investigador trata de entender o explicar algo, generalmente formula su problema de investigación por medio de una hipótesis
- Ejemplo: No sé si el tamaño medio de los ficheros de un sistema de archivos es igual a 29.3 KBytes.

Hipótesis nula

$$H_0 : \mu = 29.3$$

- Tomo una muestra de 36 ficheros del sistema de archivos



- $\bar{X} = 30.5$ KBytes.
- ¿Existe suficiente evidencia en los datos para rechazar H_0?
- ¿O la diferencia entre $\bar{X}$ y el valor hipotético de μ puede ser debido al azar?

Contraste de hipótesis

- Llamaremos **hipótesis nula**, y la denotamos por H_0, a la que se da por cierta antes de obtener la muestra. Goza de presunción de inocencia.
- Llamaremos **hipótesis alternativa**, y la denotamos por H_1 (o H_a) a lo que sucede cuando no es cierta la hipótesis nula.
- Por gozar la hipótesis nula de presunción de inocencia, sobre la hipótesis alternativa recae la carga de la prueba. Por tanto, cuando rechazamos H_0 en favor de H_1 es porque hemos encontrado pruebas significativas a partir de la muestra.

"Extraordinary claims require extraordinary evidence"
Carl Sagan

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera		
	Ho es falsa		

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	
	Ho es falsa		

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	
	Ho es falsa		Decisión correcta

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	Error de tipo I
	Ho es falsa		Decisión correcta

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	Error de tipo I
	Ho es falsa	Error de tipo II	Decisión correcta

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta	Error de tipo I
	Ho es falsa	Error de tipo II	Decisión correcta

Observamos que se puede tomar una decisión correcta o errónea.

- Error de tipo I: cuando rechazamos la hipótesis nula, siendo cierta.
- Error de tipo II: cuando no rechazamos la hipótesis nula, siendo falsa.

Contraste de hipótesis. Analogía con un juicio

Supongamos un juicio en el que se trata de decidir la culpabilidad o inocencia de un acusado.

- Hipótesis nula: el acusado es inocente (todo acusado es inocente hasta que se demuestre lo contrario).
- Hipótesis alternativa: el acusado es culpable.
- Juicio: es el procedimiento en el cual se trata de probar la culpabilidad del acusado y la evidencia debe ser muy fuerte para que se rechace la inocencia (H_0) en favor de la culpabilidad (H_a).
- Decisión: el veredicto.
- Error de tipo I: condenar a un inocente.
- Error de tipo II: absolver a un culpable.

Contraste de hipótesis

- La probabilidad del error de tipo I se denota por α y se denomina **nivel de significación**.

Nivel de significación

$$\alpha = P(\text{Rechazar } H_0 / H_0 \text{ es cierta})$$

- La probabilidad del error de tipo II se denota por β

$$\beta = P(\text{No rechazar } H_0 / H_0 \text{ es falsa})$$

- **Potencia:** Es la probabilidad de detectar que una hipótesis es falsa.

Potencia

$$\text{Potencia} = P(\text{Rechazar } H_0 / H_0 \text{ es falsa}) = 1 - \beta$$

Contraste de hipótesis

		Decisión	
		No se rechaza Ho	Se rechaza Ho
Realidad	Ho es verdadera	Decisión correcta Probabilidad 1-α	Error de tipo I Probabilidad α (significación)
	Ho es falsa	Error de tipo II Probabilidad β	Decisión correcta Probabilidad 1-β (potencia)

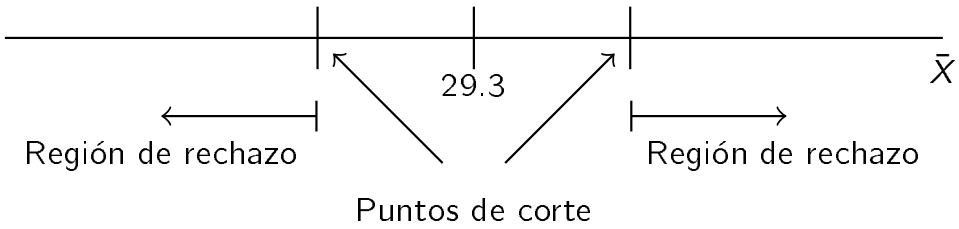
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?

Contraste bilateral

$$H_0 : \mu = 29.3$$
$$H_1 : \mu \neq 29.3$$

- Si estamos interesados en determinar si μ **difiere significativamente** de 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en ambas direcciones**.



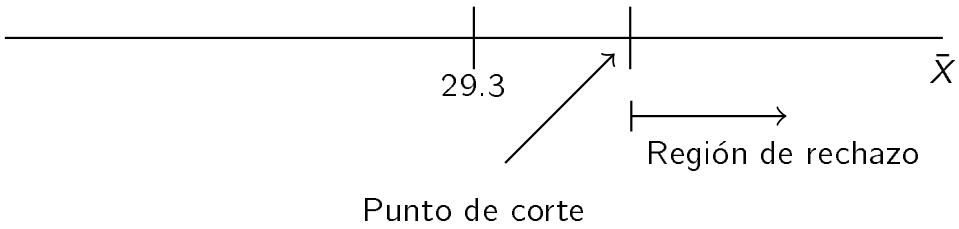
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Es el tamaño medio de los ficheros de un sistema de archivos mayor que 29.3 KBytes?

Contraste unilateral

$$H_0 : \mu \leq 29.3$$
$$H_1 : \mu > 29.3$$

- Si estamos interesados en determinar si μ **es significativamente mayor** que 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en una sola dirección**.



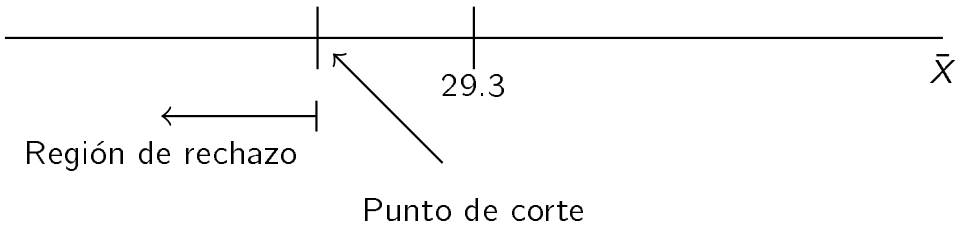
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Es el tamaño medio de los ficheros de un sistema de archivos menor que 29.3 KBytes?

Contraste unilateral

$$H_0 : \mu \geq 29.3$$
$$H_1 : \mu < 29.3$$

- Si estamos interesados en determinar si μ **es significativamente menor** que 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en una sola dirección**.



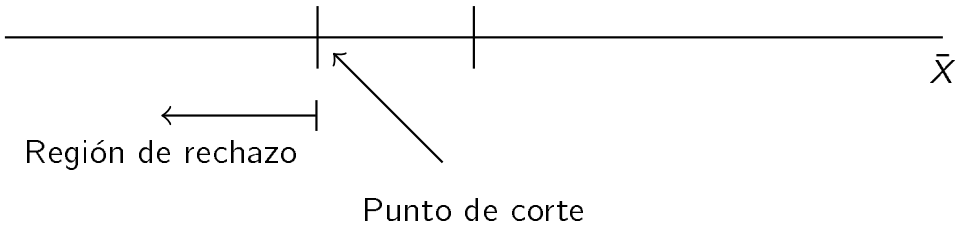
Elección de α y p-valor

- Volvamos a considerar un contraste del tipo:

Contraste unilateral

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- Deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 **en una sola dirección**.



- El punto de corte que determina la región de rechazo depende del nivel de significación α que elijamos.
- Se suele elegir α como 0.1, 0.05 o 0.01.
- El nivel elegido depende de lo importante que se considere rechazar equivocadamente H_0. Cuanto más desastrosas sean las consecuencias de rechazar H_0 siendo cierta, más pequeño se debe escoger α.

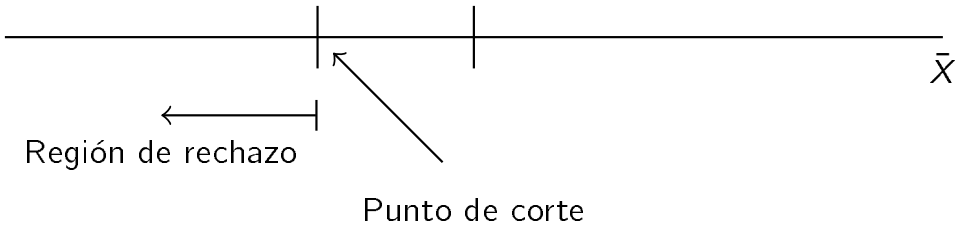
Elección de α y p-valor

- Volvamos a considerar un contraste del tipo:

Contraste unilateral

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- Deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 **en una sola dirección**.



- En lugar de fijar un nivel de significación α, podemos calcular el p-**valor**.
- El cálculo del p-valor se basa en suponer que H_0 es cierta y analizar si el valor observado en la muestra es un valor razonable para la correspondiente distribución.
- El p-valor de un contraste es la probabilidad de obtener un valor más "extremo" que el observado en el caso de que H_0 fuese cierta.

Contraste de hipótesis

Las etapas en la resolución de un contraste de hipótesis son:

- Especificar las hipótesis nula H_0 y alternativa H_1.
- Elegir un estadístico de contraste apropiado, para medir la discrepancia entre la hipótesis y la muestra.
- Fijar el nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es cierta.
- Al fijar un nivel de significación, α, se obtiene implícitamente una división en dos regiones del conjunto de posibles valores del estadístico de contraste:
  - La **región de rechazo** o región crítica que tiene probabilidad α (bajo H_0).
  - La **región de aceptación** que tiene probabilidad $1-\alpha$ (bajo H_0).
- Si el valor del estadístico cae en la región de rechazo, los datos no son compatibles con H_0 y la rechazamos. Entonces se dice que el contraste es **estadísticamente significativo**, es decir existe evidencia estadísticamente significativa a favor de H_1.
- Si el valor del estadístico cae en la región de aceptación, no existen razones suficientes para rechazar la hipótesis nula con un nivel de significación α, y el contraste se dice **estadísticamente no significativo**, es decir no existe evidencia a favor de H_1.

Contraste de hipótesis

La legislación establece que el nivel medio de ruido al que está expuesto un trabajador no debe sobrepasar los 80dB. Sin embargo, se sospecha que debido al continuo ruido que se produce en las salas de servidores, los trabajadores de un determinado centro de investigación están sometidos a niveles superiores de ruido.

- ¿Cómo plantearías un contraste de hipótesis en esta situación para poder estableces alguna conclusión sobre este problema?
- Describe los errores de tipo I y de tipo II en los que estarías incurriendo en este estudio y las consecuencias prácticas de los mismos.
- Decides hacer varias mediciones del nivel del ruido y observas que las mediciones son muy variables ¿Cómo crees que afectará este hecho a la decisión del contrastre?
- ¿Qué podrías hacer en ese caso para incrementar la potencia del test?

Contenidos: Métodos de construcción de contrastes de hipótesis

Índice del curso

Introducción

- Supongamos que se conoce que cierta característica X de una población sigue una distribución dada por:
  - una función de masa $P_\theta(x)$ en el caso discreto, o
  - una función de densidad $f_\theta(x)$ en el caso continuo,

  donde $\theta \in \Theta$ es desconocido.
- Sea $(X_1, \ldots, X_n)$ una muestra aleatoria X.
- Supongamos una partición del espacio paramétrico

$$\Theta = \Theta_0 \cup \Theta_1$$

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Un test para contrastar la hipótesis nula $H_0 : \theta \in \Theta_0$ frente a la hipótesis alternativa $H_1 : \theta \in \Theta_1$ consiste en decidir, para cada posible muestra, si rechazamos o no rechazamos H_0.

Introducción

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Se trata de dividir el espacio muestral (conjunto de todas las posibles muestras) en dos regiones: una región crítica R o de rechazo de H_0 y una región de no rechazo (complementario de R). El contraste queda así definido por la región de rechazo.
- El problema consiste entonces en definir un procedimiento para tomar una decisión sobre el contraste de la manera menos errónea posible.
- Recordamos que en un contraste se podían cometer básicamente dos tipos de errores:
  - Error de tipo I : rechazar H_0 cuando H_0 es cierta.
  - Error de tipo II: no rechazar H_0 cuando no H_0 es falsa.

Función de potencia

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- La **función de potencia** de un test con región crítica R, para contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$, es la función que a cada $\theta \in \Theta$ le hace corresponder el valor

$$P_\theta(R) = P_\theta(\text{rechazar } H_0)$$

- Nos interesará que la función de potencia tome valores:
  - próximos a 0 siempre que $\theta \in \Theta_0$.
  - próximos a 1 siempre que $\theta \in \Theta_1$.
- En la práctica, exigimos que la función de potencia no supere cierto valor pequeño α cuando $\theta \in \Theta_0$ y procuramos, después, que sea lo mayor posible cuando $\theta \in \Theta_1$.

Test uniformemente más potente

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Dado el conjunto de tests para contrastar el mismo sistema de hipótesis que tienen el mismo nivel de significación α, si existe uno que para todo $\theta \in \Theta_1$ es igual o más potente que los demás, diremos que es el test Uniformemente Más Potente (UMP)

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Abordamos ahora la cuestión de como construir tests de hipótesis de un modo sistemático y lo más objetivo posible.
- El método habitualmente utilizado es el método de **razón de verosimilitudes**.

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Supongamos que X es una característica de la población con función de masa P_θ (caso discreto).
- Para cada posible muestra, $(x_1, \ldots, x_n)$ se considera el siguiente cociente:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)}$$

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Supongamos que X es una característica de la población con función de masa P_θ (caso discreto).
- Para cada posible muestra, $(x_1, \ldots, x_n)$ se considera el siguiente cociente:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)}$$

- El cociente es siempre positivo.
- Puesto que $\Theta_0 \subset \Theta$, entonces se tiene necesariamente $0 \leq \Lambda \leq 1$.

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Supongamos que X es una característica de la población con función de masa P_θ (caso discreto).
- Para cada posible muestra, $(x_1, \ldots, x_n)$ se considera el siguiente cociente:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)}$$

- Si este cociente es pequeño, se considera poco plausible que el verdadero valor de θ esté en Θ_0. Lo más razonable será rechazar H_0.
- Si el cociente es grande, lo más razonable sería no rechazar H_0.
- La elección del punto crítico que separe la región de rechazo de H_0 de la de no rechazo se realiza fijando un determinado nivel de significación α.

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- El test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \quad \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)} \leq c \right\}$$

donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Supongamos que X es una característica de la población con función de densidad f_θ (caso continuo).
- Para cada posible muestra, $(X_1, \ldots, X_n)$ se considera el siguiente cociente:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} f_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} f_\theta(x_1, \ldots, x_n)}$$

- El test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \quad \frac{\sup_{\theta \in \Theta_0} f_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} f_\theta(x_1, \ldots, x_n)} \leq c \right\}$$

donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

Teorema de Neyman Pearson

El test UMP para contrastar

Contraste de hipótesis

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

es el definido por la región crítica

$$R = \left\{ (x_1, \ldots, x_n) : \quad \frac{\sup_{\theta \in \Theta_0} f_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} f_\theta(x_1, \ldots, x_n)} \leq c \right\}$$

donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

Test de razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- En el caso de poblaciones normales el método de razón de verosimilitudes da lugar contrastes sencillos y muy relacionados con los intervalos de confianza.

Contenidos: Contrastes en poblaciones normales

▶ Índice del curso

Contraste sobre la media de una población normal con varianza conocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ difiere de un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es conocida
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral $\bar{X}$ sea muy distinta de μ_0.

Contraste sobre la media de una población normal con varianza conocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal con una desviación típica $\sigma = 2$ KBytes.
- Tomamos una muestra de 36 ficheros del sistema de archivos. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = 29.3$$
$$H_1 : \mu \neq 29.3$$

- Observamos que $\bar{X} = 30.5$ KBytes. En base a la muestra, ¿podríamos concluir que el tamaño medio de los ficheros difiere de 29.3 KBytes?

Contraste sobre la media de una población normal con varianza conocida

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- Tenemos que $\Theta_0 = \{\mu = \mu_0\}$ y $\Theta_1 = \{\mu \neq \mu_0\}$.
- Hallamos el máximo de la función de verosimilitud en Θ_0 y en $\Theta = \Theta_0 \cup \Theta_1$.
- Derivando e igualando a cero, en Θ, el máximo se obtiene en $\hat{\mu} = \bar{x}$, y en Θ_0, el máximo se obtiene en el único valor posible $\hat{\mu} = \mu_0$.

Contraste sobre la media de una población normal con varianza conocida

- En este caso:

$$\begin{aligned}
\Lambda &= \frac{\sup_{\theta\in\Theta_0} f_\theta(x_1,\ldots,x_n)}{\sup_{\theta\in\Theta} f_\theta(x_1,\ldots,x_n)} = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2\right\}} \\
&= \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^n x_i^2 + n\mu_0^2 - 2\mu_0\sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2 - n\bar{x}^2 + 2\bar{x}\sum_{i=1}^n x_i\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(n\mu_0^2 - 2n\bar{x}\mu_0 + n\bar{x}^2\right)\right\} \\
&= \exp\left\{-\frac{n}{2}\left(\mu_0^2 - 2\bar{x}\mu_0 + \bar{x}^2\right)\right\} \\
&= \exp\left\{-\frac{n}{2}\left(\mu_0 - \bar{x}\right)^2\right\}
\end{aligned}$$

Contraste sobre la media de una población normal con varianza conocida

- El test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$, al nivel de significación α, es el que tiene como región crítica:

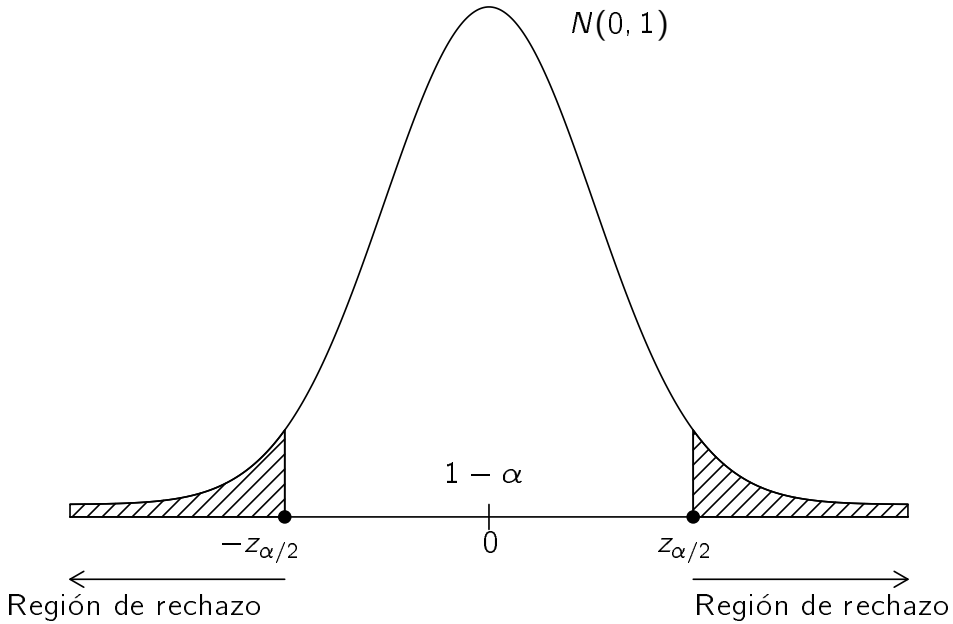
$$\begin{aligned} R &= \left\{(x_1, \ldots, x_n) : \quad \exp\left\{-\frac{n}{2}(\mu_0 - \bar{x})^2\right\} \leq c\right\} \\ &= \left\{(x_1, \ldots, x_n) : \quad n(\bar{x} - \mu_0)^2 \geq c_1\right\} \\ &= \left\{(x_1, \ldots, x_n) : \quad \left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)^2 \geq c_2\right\} \\ &= \left\{(x_1, \ldots, x_n) : \quad \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \geq c_3\right\} \end{aligned}$$

donde c_3 se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

Contraste sobre la media de una población normal con varianza conocida

- Ahora si H_0 es cierta, $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \in N(0,1)$. Por lo tanto, $c_3 = z_{\alpha/2}$.



Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Contraste sobre la media de una población normal con varianza conocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal con una desviación típica $\sigma = 2$ KBytes.
- Tomamos una muestra de 36 ficheros. Observamos que $\bar{X} = 30.5$ KBytes.

Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> sigma <- 2  # Desviación típica conocida
> mu0 <- 29.3  # Hipótesis nula
> n <- 36  # Tamaño muestral
> xbar <- 30.5  # Media muestral
> (xbar - mu0)/(sigma/sqrt(n))  # Estadístico de contraste

[1] 3.6

> alpha <- 0.05  # Nivel de significación
> qnorm(1 - alpha/2)  # z_alpha/2

[1] 1.96
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la media de una población normal con varianza conocida

- ¿Qué pasa con el contraste si no se cumple la hipótesis nula?
- Supongamos que en realidad $\mu = \mu_0 + \delta$. ¿Cuál es la potencia del test?
- ¿Podremos calcular el tamaño que debería tener la muestra para que el contraste realizado a nivel α sea estadísticamente significativo cuando el valor real de μ sea $\mu_0 + \delta$, con una potencia fija de antemano $\pi = 1 - \beta$?

$$n = \left(\frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta} \right)^2$$

Contraste sobre la media de una población normal con varianza conocida

- Ejemplo: En el contraste $H_0 : \mu = 29.3$ frente a $H_0 : \mu \neq 29.3$ ($\sigma = 2$ conocida), ¿cuál debería ser el tamaño muestral para detectar una desviación $\delta = 1$ con uan potencia $\pi = 0.9$?

```
> alpha <- 0.05  # Significación
> potencia <- 0.9  # Potencia
> beta <- 1 - potencia
> sigma <- 2  # Desviación típica conocida
> delta <- 1
> n <- ((qnorm(1 - alpha/2) + qnorm(1 - beta)) * sigma/delta)^2
> n

[1] 42.03
```

- Por lo tanto, deberíamos tomar una muestra de tamaño $n = 43$.

Contraste sobre la media de una población normal con varianza conocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ es mayor que un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es conocida
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es menor o igual que μ_0 cuando la media muestral $\bar{X}$ sea significativamente mayor que μ_0.

Contraste sobre la media de una población normal con varianza conocida

- Ejemplo: ¿Es el tamaño medio de los ficheros de un sistema de archivos mayor que 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal con una desviación típica $\sigma = 2$ KBytes.
- Tomamos una muestra de 36 ficheros del sistema de archivos. Queremos contrastar:

Contraste unilateral

$$H_0 : \mu \leq 29.3$$
$$H_1 : \mu > 29.3$$

- Observamos que $\bar{X} = 30.5$ KBytes. En base a la muestra, ¿podríamos concluir que el tamaño medio de los ficheros es mayor que 29.3 KBytes?

Contraste sobre la media de una población normal con varianza conocida

Contraste unilateral

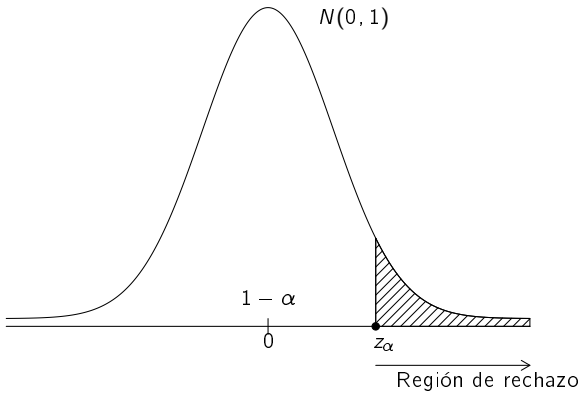
$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

- Tenemos que $\Theta_0 = \{\mu \leq \mu_0\}$ y $\Theta_1 = \{\mu > \mu_0\}$.
- Hallamos el máximo de la función de verosimilitud en Θ_0 y en $\Theta = \Theta_0 \cup \Theta_1$.

Contraste sobre la media de una población normal con varianza conocida

- Construyendo la razón de verosimilitudes, e imponiendo que la significación del contraste sea α, se llega a que el test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \right\}$$



Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0, 1)$

Contraste sobre la media de una población normal con varianza conocida

- Ejemplo: ¿Es el tamaño medio de los ficheros de un sistema de archivos mayor que 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal con una desviación típica $\sigma = 2$ KBytes.
- Tomamos una muestra de 36 ficheros. Observamos que $\bar{X} = 30.5$ KBytes.

Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0, 1)$

```
> sigma <- 2  # Desviación típica conocida
> mu0 <- 29.3  # Hipótesis nula
> n <- 36  # Tamaño muestral
> xbar <- 30.5  # Media muestral
> (xbar - mu0)/(sigma/sqrt(n))  # Estadístico de contraste

[1] 3.6

> alpha <- 0.05  # Nivel de significación
> qnorm(1 - alpha)  # z_alpha

[1] 1.645
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la media de una población normal con varianza conocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ es menor que un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es conocida
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral

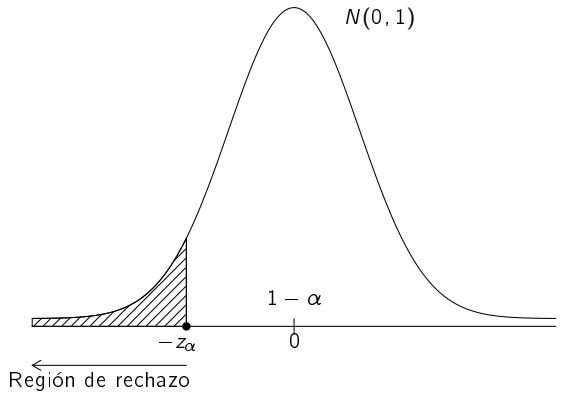
$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es mayor o igual que μ_0 cuando la media muestral $\bar{X}$ sea significativamente menor que μ_0.

Contraste sobre la media de una población normal con varianza conocida

- Construyendo la razón de verosimilitudes, e imponiendo que la significación del contraste sea α, se llega a que el test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha \right\}$$



Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0, 1)$

Contraste sobre la media de una población normal con varianza desconocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ difiere de un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es **desconocida**
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral $\bar{X}$ sea muy distinta de μ_0.

Contraste sobre la media de una población normal con varianza desconocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal pero se desconoce la desviación típica.
- Tomamos una muestra de 36 ficheros del sistema de archivos. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = 29.3$$
$$H_1 : \mu \neq 29.3$$

- Observamos que $\bar{X} = 30.5$ KBytes. En base a la muestra, ¿podríamos concluir que el tamaño medio de los ficheros difiere de 29.3 KBytes?

Contraste sobre la media de una población normal con varianza desconocida

Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- Tenemos ahora que:

$$\begin{aligned} \Theta_0 &= \{(\mu, \sigma) : \ \mu = \mu_0, \ \sigma^2 > 0\}, \\ \Theta_1 &= \{(\mu, \sigma) : \ \mu \neq \mu_0, \ \sigma^2 > 0\} \end{aligned}$$

- Hallamos el máximo de la función de verosimilitud en Θ_0 y en $\Theta = \Theta_0 \cup \Theta_1$.
- Derivando e igualando a cero, en Θ, el máximo se obtiene en

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- En Θ_0, el máximo se obtiene en

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2$$

Contraste sobre la media de una población normal con varianza desconocida

- En este caso, el test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \left| \frac{\bar{x} - \mu_0}{S_c/\sqrt{n}} \right| \geq c \right\}$$

donde

  - S_c^2 es la cuasivarianza muestral (S_c es la cuasidesviación muestral),

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  - c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

  - Ahora si H_0 es cierta, $\frac{\bar{x}-\mu_0}{S_c/\sqrt{n}} \in t_{n-1}$.

La distribución t de Student

- La t **de Student con** k **grados de libertad** es un modelo de variable aleatoria continua.

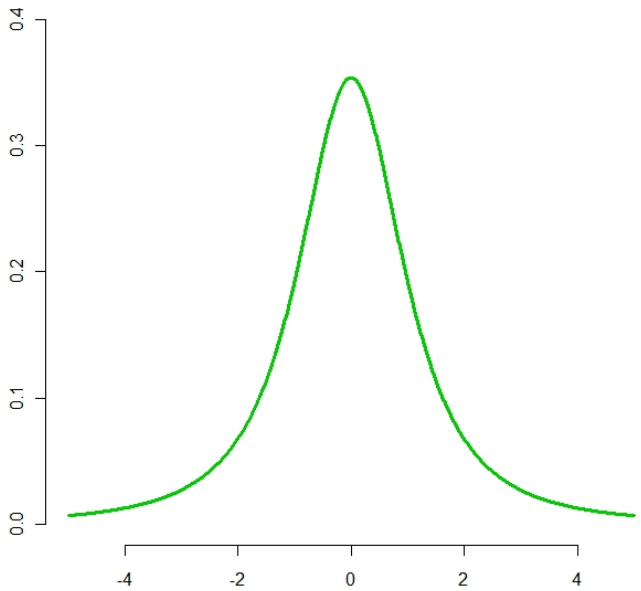


Figure: En verde densidad de una t de Student con 2 grados de libertad

La distribución t de Student

- La t **de Student con** k **grados de libertad** es un modelo de variable aleatoria continua.

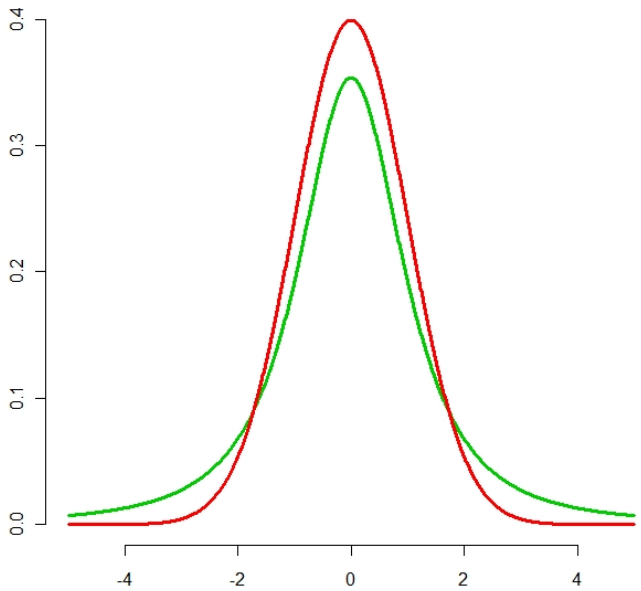


Figure: En verde densidad de una t de Student con 2 grados de libertad y en rojo densidad de una N(0,1)

La distribución t de Student

- La t **de Student con** k **grados de libertad** es un modelo de variable aleatoria continua.

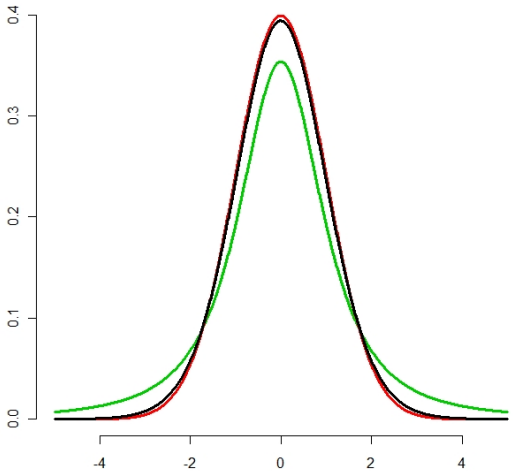
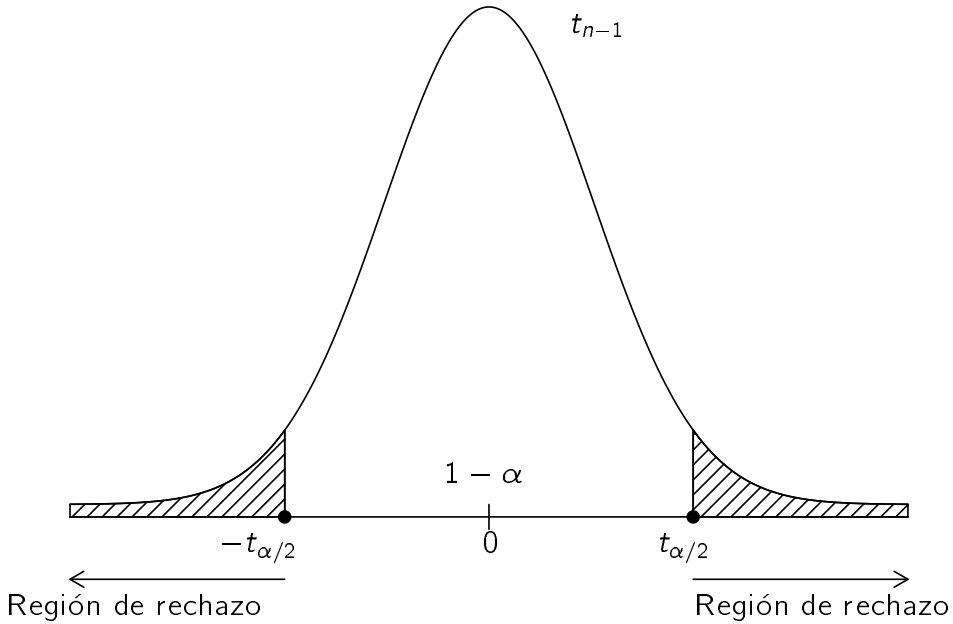


Figure: En verde densidad de una t de Student con 2 grados de libertad, en rojo N(0,1) y en negro densidad de una t de Student con 20 grados de libertad

- La variable t de Student toma valores en toda la recta real.
- La distribución t de Student es simétrica en torno al origen.
- $t_k \xrightarrow{d} N(0,1)$ cuando $k \to \infty$.

Contraste sobre la media de una población normal con varianza desconocida

- Ahora si H_0 es cierta, $\frac{\bar{x}-\mu_0}{S_c/\sqrt{n}} \in t_{n-1}$. Por lo tanto, $c = t_{\alpha/2}$.



Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X}-\mu_0}{S_c/\sqrt{n}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}-\mu_0}{S_c/\sqrt{n}} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

Contraste sobre la media de una población normal con varianza desconocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal (varianza desconocida).
- Tomamos una muestra de 36 ficheros. Observamos que $\bar{X} = 30.5$ KBytes y $S_c = 2.8$ KBytes.

Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

```
> mu0 <- 29.3  # Hipótesis nula
> n <- 36  # Tamaño muestral
> xbar <- 30.5  # Media muestral
> sc <- 2.8  # Cuasidesviación tipica
> (xbar - mu0)/(sc/sqrt(n))  # Estadístico de contraste

[1] 2.571

> alpha <- 0.05  # Nivel de significación
> qt(1 - alpha/2, n - 1)  # t_alpha/2

[1] 2.03
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la media de una población normal con varianza desconocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal (varianza desconocida).
- Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`
- Podemos realizar el contraste como hemos hecho anteriormente:

```
> x <- scan("ficheros.txt")
> mu0 <- 29.3  # Hipótesis nula
> n <- length(x)  # Tamaño muestral
> xbar <- mean(x)  # Media muestral
> sc <- sd(x)  # Cuasidesviación tipica
> (xbar - mu0)/(sc/sqrt(n))  # Estadístico de contraste

[1] 4.214

> alpha <- 0.05  # Nivel de significación
> qt(1 - alpha/2, n - 1)  # t_alpha/2

[1] 2.03
```

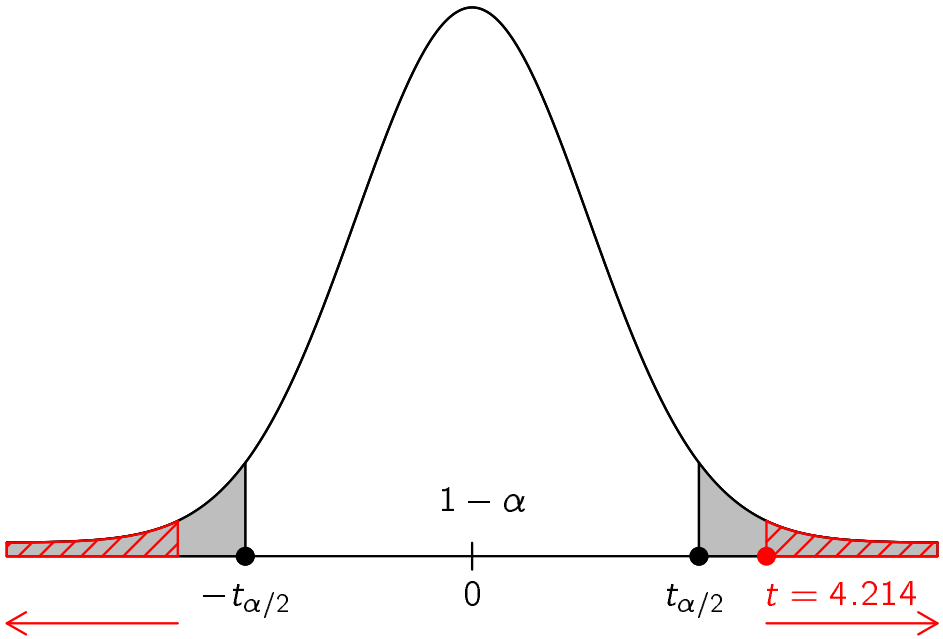
Contraste sobre la media de una población normal con varianza desconocida

- Ejemplo: ¿Difiere el tamaño medio de los ficheros de un sistema de archivos de 29.3 KBytes?
- Se sabe que el tamaño de los ficheros de un sistema de archivos sigue una distribución normal (varianza desconocida).
- Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`
- De forma equivalente, podemos usar la función `t.test`

```
> t.test(x, mu = mu0)


One Sample t-test

data:  x
t = 4.214, df = 35, p-value = 0.0001673
alternative hypothesis: true mean is not equal to 29.3
95 percent confidence interval:
 29.95 31.17
sample estimates:
mean of x
    30.56
```

El p-valor de un contraste

- A medida que el nivel de significación α disminuye, es más difícil rechazar la hipótesis nula (manteniendo los mismos datos).
- Dado un estadístico de contraste, hay un valor de α a partir del cual ya no podemos rechazar H_0.
- A dicho valor se le se le llama el p-valor del contraste y se denota por p.
- Si $\alpha < p$ no podemos rechazar H_0 a nivel α.
- Si $\alpha > p$ podemos rechazar H_0 a nivel α.



Contraste sobre la media de una población normal con varianza desconocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ es mayor que un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es **desconocida**.
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral

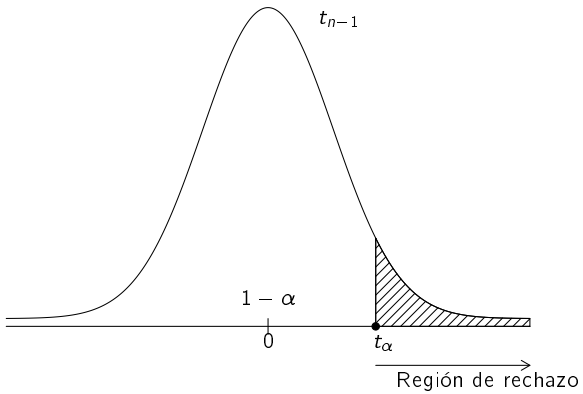
$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es menor o igual que μ_0 cuando la media muestral $\bar{X}$ sea significativamente mayor que μ_0.

Contraste sobre la media de una población normal con varianza desconocida

- Construyendo la razón de verosimilitudes, e imponiendo que la significación del contraste sea α, se llega a que el test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{S_c/\sqrt{n}} \geq t_\alpha \right\}$$



Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \geq t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n-1$ g.l.

Contraste sobre la media de una población normal con varianza desconocida

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la media μ es menor que un valor determinado μ_0?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Supongamos que la varianza σ^2 es **desconocida**.
  - Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral

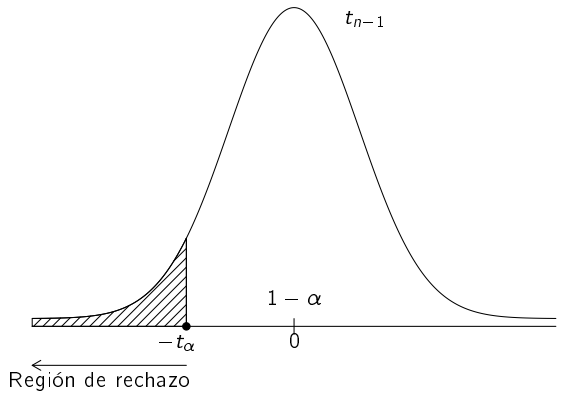
$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es mayor o igual que μ_0 cuando la media muestral $\bar{X}$ sea significativamente menor que μ_0.

Contraste sobre la media de una población normal con varianza desconocida

- Construyendo la razón de verosimilitudes, e imponiendo que la significación del contraste sea α, se llega a que el test de razón de verosimilitudes para contrastar la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \frac{\bar{x} - \mu_0}{S_c/\sqrt{n}} \leq -t_\alpha \right\}$$



Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \leq -t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n-1$ g.l.

Contraste sobre la varianza de una población normal

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la varianza σ^2 difiere de un valor determinado σ_0^2?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Se desea contrastar una hipótesis relativa a la varianza, σ^2.

Contraste bilateral (hipótesis nula simple)

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la varianza poblacional es σ_0^2 cuando la varianza muestral sea muy distinta de σ_0^2.

Contraste sobre la varianza de una población normal

- Ejemplo: Una empresa que fabrica baterías para portátiles asegura que un determinado modelo de batería tiene una vida media de 7 horas con una desviación típica de $\sqrt{5}$ horas. Un cliente que usa habitualmente dicho modelo de batería cree que el valor establecido por la compañía para la varianza no se ajusta a la realidad. ¿Difiere la varianza del tiempo de vida de la batería de lo establecido por la compañía?
- Se sabe que el tiempo de vida de la batería sigue una distribución normal.
- Tomamos una muestra los tiempos de vida de 10 baterías y obtenemos: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8.

Contraste bilateral (hipótesis nula simple)

$$H_0 : \sigma^2 = 5$$
$$H_1 : \sigma^2 \neq 5$$

Contraste sobre la varianza de una población normal

- El test para contrastar la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ frente a $H_1 : \sigma^2 \neq \sigma_0^2$, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \left| \frac{(n-1)S_c^2}{\sigma_0^2} \right| \geq c \right\}$$

  donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

- Ahora si H_0 es cierta, $\frac{(n-1)S_c^2}{\sigma_0^2} \in \chi^2_{n-1}$.

La distribución χ^2

- La χ^2_n **con** n **grados de libertad** es un modelo de variable aleatoria continua

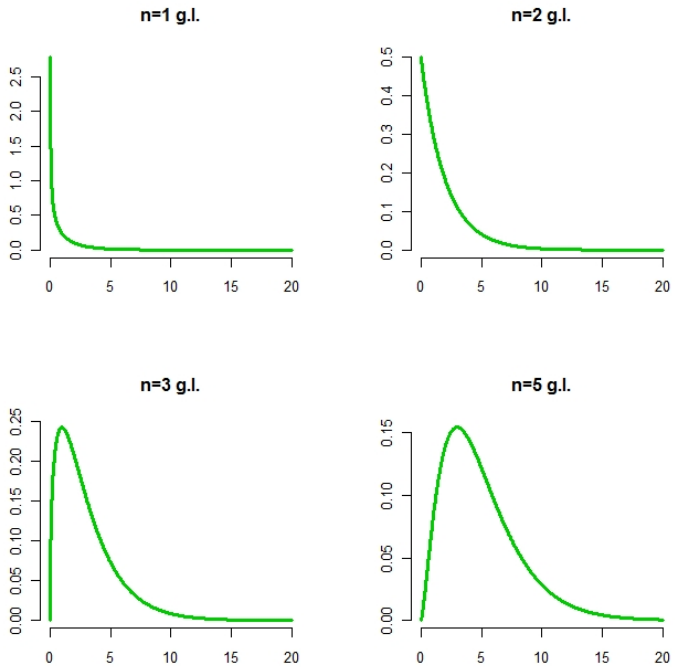
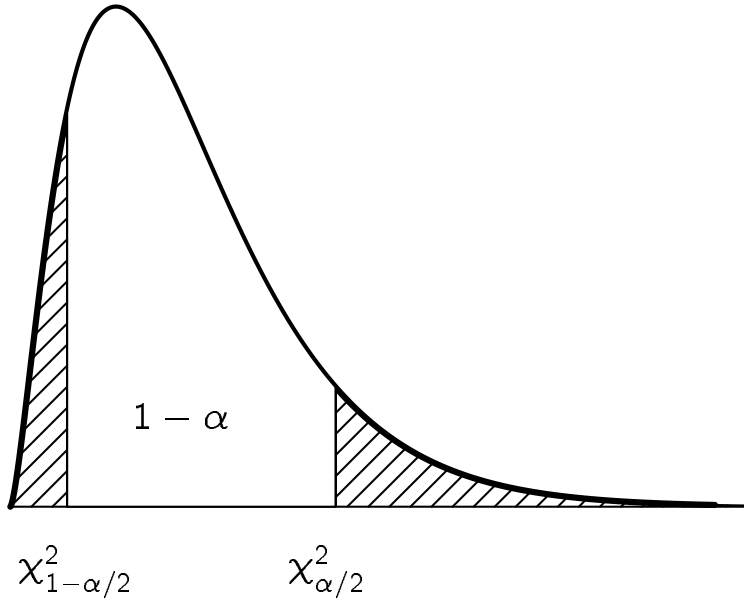


Figure: En verde densidades de variables χ^2_n para distintos valores de n.

- La variable Chi-cuadrado toma valores en $[0, +\infty)$.
- La distribución Chi-cuadrado es asimétrica.

Contraste sobre la varianza de una población normal

- Si H_0 es cierta, $\frac{(n-1)S_c^2}{\sigma_0^2} \in \chi^2_{n-1}$.



Rechazamos la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ frente a $H_1 : \sigma^2 \neq \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2} \quad \text{ó} \quad \frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi^2_{\alpha/2}$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

Contraste sobre la varianza de una población normal

- Ejemplo: ¿Difiere la varianza del tiempo de vida de la batería de $\sigma^2 = 5$ horas^2?
- Se sabe que el tiempo de vida de la batería sigue una distribución normal.
- Tomamos los tiempos de vida de 10 baterías: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8.

Rechazamos la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ frente a $H_1 : \sigma^2 \neq \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2} \quad \text{ó} \quad \frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi^2_{\alpha/2}$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

```
> x <- c(5, 6, 4, 3, 11, 12, 9, 13, 6, 8)  # Muestra
> sigma02 <- 5  # Hipótesis nula
> n <- length(x)  # Tamaño muestral
> xbar <- mean(x)  # Media muestral
> sc2 <- var(x)  # Cuasivarianza muestral
> (n - 1) * sc2/sigma02  # Estadístico de contraste

[1] 21.62

> alpha <- 0.05  # Nivel de significación
> qchisq(1 - alpha/2, n - 1)  # chi_alpha/2

[1] 19.02

> qchisq(alpha/2, n - 1)  # chi_{1-alpha/2}

[1] 2.7
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la varianza de una población normal

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la varianza σ^2 es mayor que un valor determinado σ_0^2?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Se desea contrastar una hipótesis relativa a la varianza, σ^2.

Contraste unilateral

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la varianza poblacional es menor o igual que σ_0^2 cuando la varianza muestral sea significativamente mayor que σ_0^2.

Contraste sobre la varianza de una población normal

- Ejemplo: Una empresa que fabrica baterías para portátiles asegura que un determinado modelo de batería tiene una vida media de 7 horas con una desviación típica de $\sqrt{5}$ horas. Un cliente que usa habitualmente dicho modelo de batería cree que el valor establecido por la compañía para la varianza no se ajusta a la realidad y que la variabilidad es mayor. ¿Es mayor la varianza del tiempo de vida de la batería de lo establecido por la compañía?
- Se sabe que el tiempo de vida de la batería sigue una distribución normal.
- Tomamos una muestra los tiempos de vida de 10 baterías y obtenemos: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8.

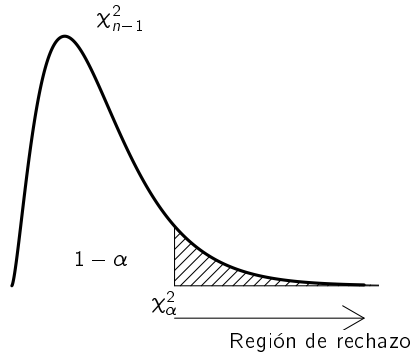
Contraste unilateral

$$H_0 : \sigma^2 \leq 5$$
$$H_1 : \sigma^2 > 5$$

Contraste sobre la varianza de una población normal

- Imponiendo que la significación del contraste sea α, se llega a que el test para contrastar la hipótesis nula $H_0 : \sigma^2 \leq \sigma_0^2$ frente a $H_1 : \sigma^2 > \sigma_0^2$ es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi_\alpha^2 \right\}$$



Rechazamos la hipótesis nula $H_0 : \sigma^2 \leq \sigma_0^2$ frente a $H_1 : \sigma^2 > \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi_\alpha^2$$

χ_α^2 denota el punto tal que $P(J > \chi_\alpha^2) = \alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

Contraste sobre la varianza de una población normal

- Ejemplo: ¿Es mayor la varianza del tiempo de vida de la batería que $\sigma^2 = 5$ horas^2?
- Se sabe que el tiempo de vida de la batería sigue una distribución normal.
- Tomamos los tiempos de vida de 10 baterías: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8.

Rechazamos la hipótesis nula $H_0 : \sigma^2 \leq \sigma_0^2$ frente a $H_1 : \sigma^2 > \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi_\alpha^2$$

χ_α^2 denota el punto tal que $P(J > \chi_\alpha^2) = \alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

```
> x <- c(5, 6, 4, 3, 11, 12, 9, 13, 6, 8)  # Muestra
> sigma02 <- 5  # Hipótesis nula
> n <- length(x)  # Tamaño muestral
> xbar <- mean(x)  # Media muestral
> sc2 <- var(x)  # Cuasivarianza muestral
> (n - 1) * sc2/sigma02  # Estadístico de contraste

[1] 21.62

> alpha <- 0.05  # Nivel de significación
> qchisq(1 - alpha, n - 1)  # chi_alpha

[1] 16.92
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la varianza de una población normal

- Ejemplo: ¿Es mayor la varianza del tiempo de vida de la batería que $\sigma^2 = 5$ horas^2?
- Se sabe que el tiempo de vida de la batería sigue una distribución normal.
- Tomamos los tiempos de vida de 10 baterías: 5, 6, 4, 3, 11, 12, 9, 13, 6, 8.

Rechazamos la hipótesis nula $H_0 : \sigma^2 \leq \sigma_0^2$ frente a $H_1 : \sigma^2 > \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

```
> x <- c(5, 6, 4, 3, 11, 12, 9, 13, 6, 8)  # Muestra
> sigma02 <- 5  # Hipótesis nula
> n <- length(x)  # Tamaño muestral
> xbar <- mean(x)  # Media muestral
> sc2 <- var(x)  # Cuasivarianza muestral
> est <- (n - 1) * sc2/sigma02  # Estadístico de contraste
> 1 - pchisq(est, n - 1)  # p-valor

[1] 0.01016
```

- Por lo tanto, para cualquier significación superior a 0.01016, rechazamos H_0.

Contraste sobre la varianza de una población normal

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la varianza σ^2 es menor que un valor determinado σ_0^2?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma^2)$.
  - Se desea contrastar una hipótesis relativa a la varianza, σ^2.

Contraste unilateral

$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la varianza poblacional es menor o igual que σ_0^2 cuando la varianza muestral sea significativamente menor que σ_0^2.

Contraste sobre la varianza de una población normal

- Imponiendo que la significación del contraste sea α, se llega a que el test para contrastar la hipótesis nula $H_0 : \sigma^2 \geq \sigma_0^2$ frente a $H_1 : \sigma^2 < \sigma_0^2$ es el que tiene como región crítica:

$$R = \left\{(x_1, \ldots, x_n) : \frac{(n-1)S_c^2}{\sigma_0^2} \leq \chi^2_{1-\alpha}\right\}$$

Rechazamos la hipótesis nula $H_0 : \sigma^2 \geq \sigma_0^2$ frente a $H_1 : \sigma^2 < \sigma_0^2$ si

$$\frac{(n-1)S_c^2}{\sigma_0^2} \leq \chi^2_{1-\alpha}$$

$\chi^2_{1-\alpha}$ denota el punto tal que $P(J > \chi^2_{1-\alpha}) = 1-\alpha$ siendo J una variable Ji-cuadrado con $n-1$ g.l.

Ejercicio

- En 20 días lectivos y a la misma hora se ha observado el número de terminales de una universidad conectados a Internet. Los resultados son los siguientes

$$1027, 1023, 1369, 950, 1436, 957, 634, 821, 882, 942,$$
$$904, 984, 1067, 570, 1063, 1307, 1212, 1045, 1047, 1178.$$

  Suponiendo normalidad, ¿se puede afirmar que el número de terminales de la universidad conectados a Internet supera los 1000 terminales?

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones con medias distintas?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).
- Suponemos que las **varianzas σ_1^2 y σ_2^2 son conocidas**.

Contraste bilateral (hipótesis nula simple)

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

- El sentido común nos aconseja rechazar la hipótesis nula cuando las medias muestrales correspondientes a ambas muestras sean muy distintas entre si.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el número de usuarios concurrentes a A sigue una distribución normal con desviación típica $\sigma = 10$.
- Se sabe que el número de usuarios concurrentes a B sigue una distribución normal con desviación típica $\sigma = 8$.
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

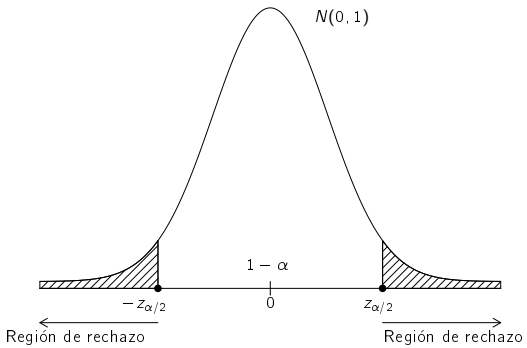
Contraste bilateral (hipótesis nula simple)

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

siendo μ_1 el número medio de usuarios concurrentes a A y μ_2 el número medio de usuarios concurrentes a B.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Si H_0 es cierta, la distribución de $\dfrac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ es $N(0,1)$.



Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A sigue una distribución normal con $\sigma = 10$.
- Se sabe que el nº de usuarios concurrentes a B sigue una distribución normal con $\sigma = 8$.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Observamos que $\bar{x}_1 = 140$ y $\bar{x}_2 = 134$

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> var1 <- 10^2; n1 <- 15; xbar1 <- 140  # Varianza, tamaño muestral y media muestral de A
> var2 <- 8^2; n2 <- 20; xbar2 <- 134  # Varianza, tamaño muestral y media muestral de B
> (xbar1 - xbar2)/sqrt(var1/n1 + var2/n2) # Estadístico de contraste

[1] 1.91

> alpha <- 0.05 # Nivel de significación
> qnorm(1 - alpha/2) # z_alpha/2

[1] 1.96
```

- Por lo tanto, para una significación del 5%, no rechazamos H_0.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A sigue una distribución normal con $\sigma = 10$.
- Se sabe que el nº de usuarios concurrentes a B sigue una distribución normal con $\sigma = 8$.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Observamos que $\bar{x}_1 = 140$ y $\bar{x}_2 = 134$

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- También podemos calcular el p-valor del contraste:

```
> var1 <- 10^2; n1 <- 15; xbar1 <- 140  # Varianza, tamaño muestral y media muestral de A
> var2 <- 8^2; n2 <- 20; xbar2 <- 134  # Varianza, tamaño muestral y media muestral de B
> est<-(xbar1 - xbar2)/sqrt(var1/n1 + var2/n2) # Estadístico de contraste
> 2*(1-pnorm(est)) # p-valor

[1] 0.05611
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones donde la media de la primera es superior a la media de la segunda?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).
- Suponemos que las **varianzas σ_1^2 y σ_2^2 son conocidas**.

Contraste unilateral

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando la media muestral de la primera muestra sea significativamente mayor que la media muestral de la segunda muestra.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A es significativamente mayor que el número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el número de usuarios concurrentes a A sigue una distribución normal con desviación típica $\sigma = 10$.
- Se sabe que el número de usuarios concurrentes a B sigue una distribución normal con desviación típica $\sigma = 8$.
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

Contraste unilateral

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

siendo μ_1 el número medio de usuarios concurrentes a A y μ_2 el número medio de usuarios concurrentes a B.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Si H_0 es cierta, la distribución de $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ es $N(0,1)$.

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A sigue una distribución normal con $\sigma = 10$.
- Se sabe que el nº de usuarios concurrentes a B sigue una distribución normal con $\sigma = 8$.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Observamos que $\bar{x}_1 = 140$ y $\bar{x}_2 = 134$.

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

```
> var1 <- 10^2; n1 <- 15; xbar1 <- 140  # Varianza, tamaño muestral y media muestral de A
> var2 <- 8^2; n2 <- 20; xbar2 <- 134  # Varianza, tamaño muestral y media muestral de B
> (xbar1 - xbar2)/sqrt(var1/n1 + var2/n2) # Estadístico de contraste

[1] 1.91

> alpha <- 0.05 # Nivel de significación
> qnorm(1 - alpha) # z_alpha/2

[1] 1.645
```

- Por lo tanto, para una significación del 5%, rechazamos H_0.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A sigue una distribución normal con $\sigma = 10$.
- Se sabe que el nº de usuarios concurrentes a B sigue una distribución normal con $\sigma = 8$.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Observamos que $\bar{x}_1 = 140$ y $\bar{x}_2 = 134$.

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

- También podemos calcular el p-valor del contraste:

```
> var1 <- 10^2; n1 <- 15; xbar1 <- 140  # Varianza, tamaño muestral y media muestral de A
> var2 <- 8^2; n2 <- 20; xbar2 <- 134  # Varianza, tamaño muestral y media muestral de B
> est<-(xbar1 - xbar2)/sqrt(var1/n1 + var2/n2) # Estadístico de contraste
> 1-pnorm(est) # p-valor

[1] 0.02806
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones donde la media de la primera es inferior a la media de la segunda?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).
- Suponemos que las **varianzas σ_1^2 y σ_2^2 son conocidas**.

Contraste unilateral

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando la media muestral de la primera muestra sea significativamente menor que la media muestral de la segunda muestra.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas conocidas

- Si H_0 es cierta, la distribución de $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ es $N(0,1)$.

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones con medias distintas?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).
- Suponemos que las **varianzas σ_1^2 y σ_2^2 son desconocidas pero iguales** $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$.

Contraste bilateral (hipótesis nula simple)

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

- El sentido común nos aconseja rechazar la hipótesis nula cuando las medias muestrales correspondientes a ambas muestras sean muy distintas entre si.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el número de usuarios concurrentes a A sigue una distribución normal.
- Se sabe que el número de usuarios concurrentes a B sigue una distribución normal.
- Las varianzas del número de usuarios concurrentes a cada una de las aplicaciones son desconocidas pero iguales.
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

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

siendo μ_1 el número medio de usuarios concurrentes a A y μ_2 el número medio de usuarios concurrentes a B.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

- Suponiendo que las **varianzas σ_1^2 y σ_2^2 son desconocidas pero iguales**, calculamos:

$$S_p^2 = \frac{(n_1-1)S_{c1}^2 + (n_2-1)S_{c2}^2}{n_1+n_2-2}.$$

donde S_{c1}^2 y S_{c2}^2 denotan las cuasivarianzas muestrales:

$$S_{c1}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2 \qquad S_{c2}^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2$$

- Si H_0 es cierta, la distribución de $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}}$ es $t_{n_1+n_2-2}$.

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_p^2}{n_1}+\frac{Sp^2}{n_2}}} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con n_1+n_2-2 g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A sigue una distribución normal.
- Se sabe que el nº de usuarios concurrentes a B sigue una distribución normal.
- Las varianzas del nº de usuarios concurrentes a cada una de las aplicaciones son desconocidas pero iguales.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Observamos que $\bar{x}_1 = 140$, $S^2_{c1} = 185$, $\bar{x}_2 = 134$ y , $S^2_{c2} = 40$

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{Sp^2}{n_2}}} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n_1 + n_2 - 2$ g.l.

```
> n1 <- 15; xbar1 <- 140; sc2_1 <- 185  # Tamaño, media y cuasivarianza muestral de A
> n2 <- 20; xbar2 <- 134; sc2_2 <- 40  # Tamaño, media y cuasivarianza muestral de B
> sp2 <- ((n1 - 1) * sc2_1 + (n2 - 1) * sc2_2)/(n1 + n2 - 2) # Varianza muestral ponderada
> est <- (xbar1 - xbar2)/sqrt(sp2/n1 + sp2/n2) # Estadístico de contraste
> 2 * (1 - pt(est, n1 + n2 - 2)) # p-valor

[1] 0.09056
```

- Por lo tanto, para una significación superior al p-valor, rechazamos H_0.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \geq t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n_1 + n_2 - 2$ g.l.

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \leq -t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n_1 + n_2 - 2$ g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

- Ejemplo: ¿Es el número medio de usuarios concurrentes a una aplicación A superior al número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Las varianzas del nº de usuarios concurrentes a cada una de las aplicaciones son desconocidas pero iguales.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \geq t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n_1 + n_2 - 2$ g.l.

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> n1 <- 15; xbar1 <- mean(A); sc2_1 <- var(A)  # Tamaño, media y cuasivarianza muestral de A
> n2 <- 20; xbar2 <- mean(B); sc2_2 <- var(B)  # Tamaño, media y cuasivarianza muestral de B
> sp2 <- ((n1 - 1) * sc2_1 + (n2 - 1) * sc2_2)/(n1 + n2 - 2) # Varianza muestral ponderada
> est <- (xbar1 - xbar2)/sqrt(sp2/n1 + sp2/n2) # Estadístico de contraste
> est

[1] 0.9864

> 1 - pt(est, n1 + n2 - 2) # p-valor

[1] 0.1656
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas pero iguales

- Ejemplo: ¿Es el número medio de usuarios concurrentes a una aplicación A superior al número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Las varianzas del nº de usuarios concurrentes a cada una de las aplicaciones son desconocidas pero iguales.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

- Utilizando la función `t.test`:

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> t.test(A, B, alternative = "greater", var.equal = TRUE)


Two Sample t-test

data:  A and B
t = 0.9864, df = 33, p-value = 0.1656
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -2.445    Inf
sample estimates:
mean of x mean of y
    139.3     135.8
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones con medias distintas?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).
- Suponemos que las **varianzas σ_1^2 y σ_2^2 son desconocidas**.

Contraste bilateral (hipótesis nula simple)

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

- El sentido común nos aconseja rechazar la hipótesis nula cuando las medias muestrales correspondientes a ambas muestras sean muy distintas entre si.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el número de usuarios concurrentes a A sigue una distribución normal.
- Se sabe que el número de usuarios concurrentes a B sigue una distribución normal.
- Las varianzas del número de usuarios concurrentes a cada una de las aplicaciones son desconocidas.
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

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

siendo μ_1 el número medio de usuarios concurrentes a A y μ_2 el número medio de usuarios concurrentes a B.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

- Si H_0 es cierta, la distribución de $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_{c1}^2}{n_1}+\frac{S_{c2}^2}{n_2}}}$ converge lentamente a una $N(0,1)$.
- Para muestras pequeñas se suele utilizar la aproximación de Welch, según la cual el estadístico sigue una distribución t de Student con ν grados de libertad, siendo ν los grados de libertad aproximados:

$$\nu = \frac{\left(\frac{S_{c1}^2}{n_1}+\frac{S_{c2}^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_{c1}^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_{c2}^2}{n_2}\right)^2}-2$$

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_p^2}{n_1}+\frac{Sp^2}{n_2}}} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con ν g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Las varianzas del nº de usuarios concurrentes a cada una de las aplicaciones son desconocidas.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> n1 <- 15; xbar1 <- mean(A); sc2_1 <- var(A)  # Tamaño, media y cuasivarianza muestral de A
> n2 <- 20; xbar2 <- mean(B); sc2_2 <- var(B)  # Tamaño, media y cuasivarianza muestral de B
> est <-(xbar1 - xbar2)/sqrt(sc2_1/n1 + sc2_2/n2)
> est

[1] 0.8957

> (sc2_1/n1 + sc2_2/n2)^2/(1/(n1 - 1) * (sc2_1/n1)^2 + 1/(n2 - 1) * (sc2_2/n2)^2) # g.l.

[1] 18.43

> 2 * (1 - pt(est, 19))

[1] 0.3816
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

- Ejemplo: ¿El número medio de usuarios concurrentes a una aplicación A difiere del número medio de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Las varianzas del nº de usuarios concurrentes a cada una de las aplicaciones son desconocidas.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

- Utilizando la función `t.test`:

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> t.test(A, B)


Welch Two Sample t-test

data:  A and B
t = 0.8957, df = 18.43, p-value = 0.382
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.584 11.418
sample estimates:
mean of x mean of y
    139.3     135.8
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras independientes, varianzas desconocidas y desiguales

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} \geq t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con ν g.l.

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} \leq -t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con ν g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

Hipótesis: ¿Se puede concluir que dos muestras dependientes proceden de dos poblaciones con medias distintas?

- En ocasiones nos interesará comparar dos métodos o tratamientos.
- En ese caso es natural que los individuos donde se aplican los tratamientos sean los mismos.
- Se supone $X_1 \in N\left(\mu_1, \sigma_1^2\right)$ y $X_2 \in N\left(\mu_2, \sigma_2^2\right)$ pero X_1 y X_2 **no son independientes**.
- Consideraremos en es caso la variable $D = X_1 - X_2$

Contraste bilateral (hipótesis nula simple)

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

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo medio de ejecución de nuestro código original difiere del tiempo medio de ejecución tras modificar la función?
- Se sabe que el tiempo de ejecución antes y después del cambio sigue una distribución normal.
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

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

siendo μ_1 el tiempo medio antes de la modificación y μ_2 el tiempo medio después de la modificación.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

- Si H_0 es cierta, $\frac{\bar{D}}{S_D/\sqrt{n}}$ es t_{n-1}.

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{D}}{S_D/\sqrt{n}} \geq t_{\alpha/2}$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n-1$ g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

- Ejemplo: ¿El tiempo medio de ejecución de nuestro código original difiere del tiempo medio de ejecución tras modificar la función?
- Se sabe que el tiempo de ejecución antes y después del cambio sigue una distribución normal.
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado. Los datos se muestran a continuación.

```
> torig <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> tmodif <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> n <- length(torig)
> Dif <- torig - tmodif # D = X1 - X2 [antes - después]
> Difbar <- mean(Dif); sDif <- sd(Dif)
> est <- Difbar/(sDif/sqrt(n))
> est

[1] 3.867

> 2 * (1 - pt(est, n-1))

[1] 0.003806
```

- Por lo tanto, para una significación superior al p-valor, rechazamos H_0.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

- Ejemplo: ¿El tiempo medio de ejecución de nuestro código original difiere del tiempo medio de ejecución tras modificar la función?
- Se sabe que el tiempo de ejecución antes y después del cambio sigue una distribución normal.
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado. Los datos se muestran a continuación.

- Utilizando la función `t.test`:

```
> torig <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> tmodif <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> t.test(torig, tmodif, paired = TRUE)


Paired t-test

data:  torig and tmodif
t = 3.867, df = 9, p-value = 0.003806
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6466 2.4694
sample estimates:
mean of the differences
                  1.558
```

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \geq t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n-1$ g.l.

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \leq -t_\alpha$$

t_α denota el punto tal que $P(T > t_\alpha) = \alpha$ siendo T una variable t de Student con $n-1$ g.l.

Contraste sobre la diferencia de medias de dos poblaciones normales: muestras apareadas

- Ejemplo: ¿El tiempo medio de ejecución de nuestro código original es significativamente superior al tiempo medio de ejecución tras modificar la función?
- Se sabe que el tiempo de ejecución antes y después del cambio sigue una distribución normal.
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado. Los datos se muestran a continuación.
- Utilizando la función `t.test`:

```
> torig <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> tmodif <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> t.test(torig, tmodif, paired = TRUE, alternative = "greater")


Paired t-test

data:  torig and tmodif
t = 3.867, df = 9, p-value = 0.001903
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.8195    Inf
sample estimates:
mean of the differences
                  1.558
```

- Observamos que el tiempo de ejecución ha disminuido significativamente al modificar el código.

Contraste sobre la razón de varianzas de dos poblaciones normales

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones con varianzas distintas?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).

Contraste bilateral (hipótesis nula simple)

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando las varianzas muestrales correspondientes a ambas muestras sean muy distintas entre si.

Contraste sobre la razón de varianzas de dos poblaciones normales

- Ejemplo: En un ejemplo anterior nos preguntábamos si el número medio de usuarios concurrentes a una aplicación A difería del número medio de usuarios concurrentes a otra aplicación B. En uno de los casos supusimos que las varianzas eran desconocidas pero iguales. Tiene sentido, por tanto, preguntarse si dicha suposición era aceptable.
- Se sabe que el número de usuarios concurrentes a A y a B sigue una distribución normal.
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Contraste sobre la razón de varianzas de dos poblaciones normales

- Consideramos el estadístico:

$$\frac{S_{c1}^2}{S_{c2}^2}$$

donde S_{c1}^2 y S_{c2}^2 denotan las cuasivarianzas muestrales:

$$S_{c1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 \qquad S_{c2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2$$

- Si H_0 es cierta, $\frac{S_{c1}^2}{S_{c2}^2}$ es F_{n_1-1,n_2-1}.

La distribución F de Snedecor

- La F **con** d_1, d_2 **grados de libertad** es un modelo de variable aleatoria continua

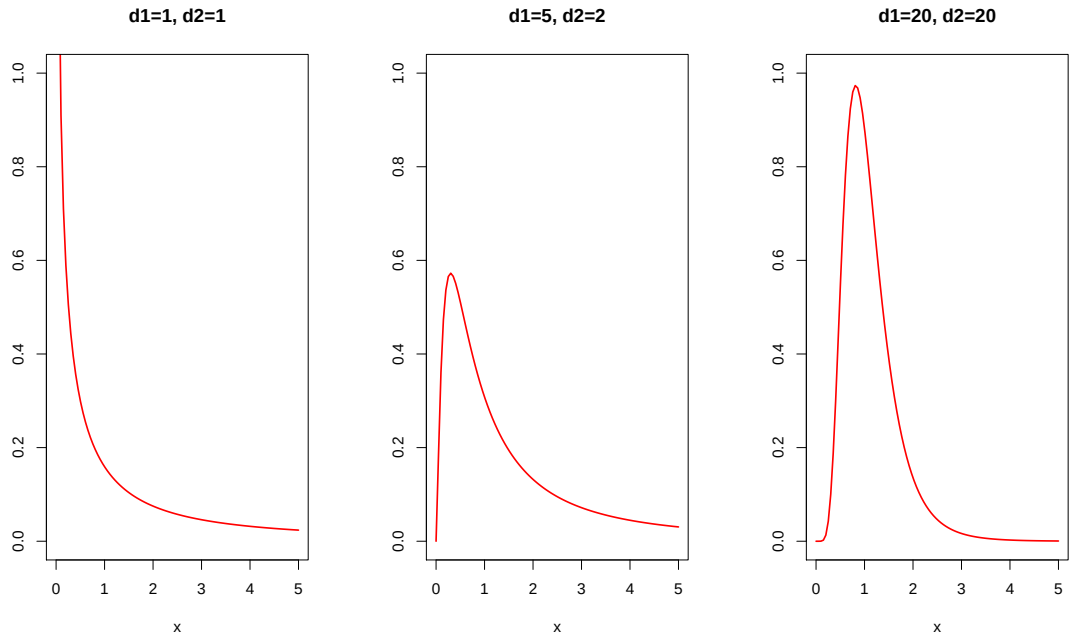


Figure: Densidades de variables F para distintos valores de d_1, d_2.

- La variable F toma valores en $[0, +\infty)$.
- La distribución F es asimétrica.

Contraste sobre la razón de varianzas de dos poblaciones normales

- Si H_0 es cierta, $\frac{S_{c1}^2}{S_{c2}^2}$ es F_{n_1-1,n_2-1}.

Rechazamos la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$ si

$$\frac{S_{c1}^2}{S_{c2}^2} \leq f_{1-\alpha/2} \quad \text{ó} \quad \frac{S_{c1}^2}{S_{c2}^2} \geq f_{\alpha/2}$$

f_α denota el punto tal que $P(F > f_\alpha) = \alpha$ siendo F una variable F de Snedecor con $n_1 - 1, n_2 - 1$ g.l.

Contraste sobre la razón de varianzas de dos poblaciones normales

- Ejemplo: ¿Difiere la varianza del número de usuarios concurrentes a una aplicación A de la varianza del número de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> n1 <- 15; xbar1 <- mean(A); sc2_1 <- var(A)  # Tamaño, media y cuasivarianza muestral de A
> n2 <- 20; xbar2 <- mean(B); sc2_2 <- var(B)  # Tamaño, media y cuasivarianza muestral de B
> sc2_1/sc2_2 # Estadístico de contraste

[1] 4.746

> alpha <- 0.05
> qf(1 - alpha/2, n1 -1 , n2 - 1)

[1] 2.647

> qf(alpha/2, n1 -1 , n2 - 1)

[1] 0.3496
```

Contraste sobre la razón de varianzas de dos poblaciones normales

- Ejemplo: ¿Difiere la varianza del número de usuarios concurrentes a una aplicación A de la varianza del número de usuarios concurrentes a otra aplicación B?
- Se sabe que el nº de usuarios concurrentes a A y a B sigue una distribución normal.
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, en 20 ocasiones el nº de usuarios concurrentes a la aplicación B. Los datos se encuentran en los ficheros `usuariosA.txt` y `usuariosB.txt`.

- Utilizando la función `var.test`:

```
> A <- scan("usuariosA.txt"); B <- scan("usuariosB.txt")
> var.test(A,B)


F test to compare two variances

data:  A and B
F = 4.746, num df = 14, denom df = 19, p-value =
0.00209
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.793 13.578
sample estimates:
ratio of variances
             4.746
```

Contraste sobre la razón de varianzas de dos poblaciones normales

Rechazamos la hipótesis nula $H_0 : \sigma_1^2 \leq \sigma_2^2$ frente a $H_1 : \sigma_1^2 > \sigma_2^2$ si

$$\frac{S_{c1}^2}{S_{c2}^2} \geq f_\alpha$$

f_α denota el punto tal que $P(F > f_\alpha) = \alpha$ siendo F una variable F de Snedecor con $n_1 - 1, n_2 - 1$ g.l.

Rechazamos la hipótesis nula $H_0 : \sigma_1^2 \geq \sigma_2^2$ frente a $H_1 : \sigma_1^2 < \sigma_2^2$ si

$$\frac{S_{c1}^2}{S_{c2}^2} \leq f_{1-\alpha}$$

$f_{1-\alpha}$ denota el punto tal que $P(F > f_{1-\alpha}) = 1 - \alpha$ siendo F una variable F de Snedecor con $n_1 - 1, n_2 - 1$ g.l.

Contenidos: Contrastes sobre la proporción

Índice del curso

Contrastes sobre una proporción

Hipótesis: ¿Se puede concluir a partir de una muestra que la proporción p de individuos de la población que cumplen una determinada característica difiere de un valor predeterminado p_0?

Contraste bilateral (hipótesis nula simple)

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando la proporción de individuos en la muestra que cumplen la característica de interés difiera mucho de p_0.

Contrastes sobre una proporción

- Ejemplo: Una encuesta del proyecto "Pew Internet and American Life Project"[3] llevada a cabo en 2010 determina que el 16% de los usuarios de internet utilizan la red para consultar información sobre resultados de pruebas médicas. ¿Podemos concluir que el porcentaje real difiere de los establecido por este estudio?
- Realizamos un estudio alternativo en el que encuestamos a 3000 usuarios de internet. Del total, 520 afirman consultar información sobre resultados de pruebas médicas. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : p = 0.16$$
$$H_1 : p \neq 0.16$$

siendo p la probabilidad de que un usuario de internet consulte información sobre resultados de pruebas médicas.

[3] http://www.pewinternet.org/

Contrastes sobre una proporción

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple donde

$$X_i = \begin{cases} 1 & \text{, con probabilidad } p \\ 0 & \text{, con probabilidad } 1-p \end{cases}$$

Estimación puntual de una proporción p

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

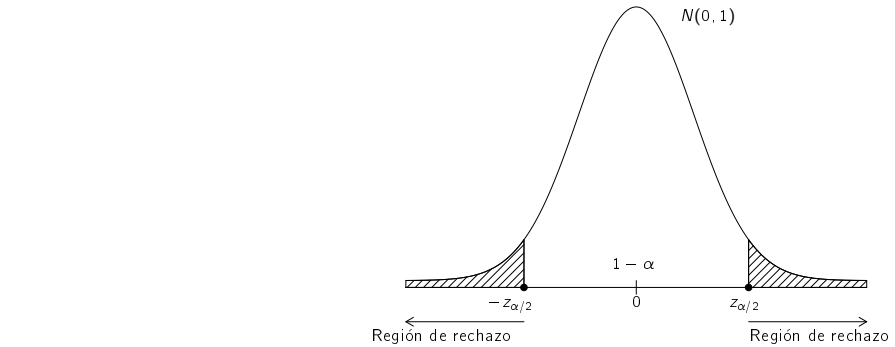
Para n grande, por el Teorema Central de Límite:

Distribución de $\hat{p}$

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Contrastes sobre una proporción

- Por lo tanto, si H_0 es cierta, la distribución de $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ se aproxima para n grande a una $N(0,1)$.



Rechazamos la hipótesis nula $H_0 : p = p_0$ frente a $H_1 : p \neq p_0$ si

$$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Contrastes sobre una proporción

- Ejemplo: Una encuesta del proyecto "Pew Internet and American Life Project"[4] llevada a cabo en 2010 determina que el 16% de los usuarios de internet utilizan la red para consultar información sobre resultados de pruebas médicas. ¿Podemos concluir que el porcentaje real difiere de lo establecido por este estudio?
- Realizamos un estudio alternativo en el que encuestamos a 3000 usuarios de internet. Del total, 520 afirman consultar información sobre resultados de pruebas médicas. Queremos contrastar:

Rechazamos la hipótesis nula $H_0 : p = p_0$ frente a $H_1 : p \neq p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> p0 <- 0.16  # Hipótesis nula
> n <- 3000  # Tamaño muestral
> exitos <- 520  # Número de éxitos
> phat <- exitos/n  # Proporción muestral
> est <- (phat - p0)/sqrt(p0 * (1 - p0)/n)  # Estadístico del contraste
> 2 * (1 - pnorm(est))  # p-valor

[1] 0.04637
```

[4] http://www.pewinternet.org/

Contrastes sobre una proporción

- Ejemplo: Una encuesta del proyecto "Pew Internet and American Life Project"[5] llevada a cabo en 2010 determina que el 16% de los usuarios de internet utilizan la red para consultar información sobre resultados de pruebas médicas. ¿Podemos concluir que el porcentaje real difiere de lo establecido por este estudio?
- Realizamos un estudio alternativo en el que encuestamos a 3000 usuarios de internet. Del total, 520 afirman consultar información sobre resultados de pruebas médicas. Queremos contrastar:
- Podemos realizar el test exacto con la función `binom.test`:

```
> binom.test(520, 3000, 0.16)  # Argumentos: éxitos, tamaño muestral, p0


Exact binomial test

data:  520 and 3000
number of successes = 520, number of trials = 3000,
p-value = 0.04912
alternative hypothesis: true probability of success is not equal to 0.16
95 percent confidence interval:
 0.1599 0.1874
sample estimates:
probability of success
                0.1733
```

[5] http://www.pewinternet.org/

Contrastes sobre una proporción

Rechazamos la hipótesis nula $H_0 : p \leq p_0$ frente a $H_1 : p > p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Rechazamos la hipótesis nula $H_0 : p \geq p_0$ frente a $H_1 : p < p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Contrastes sobre dos proporciones

Hipótesis: ¿Se puede concluir a partir de dos muestra independientes que las proporciones de individuos que cumplen una determinada característica en dos poblaciones son distintas?

- En algunas ocasiones estamos interesados en hacer contrastes sobre dos proporciones p_1 y p_2 de dos poblaciones.
- Tenemos dos muestras:
  - Una muestra formada por n_1 variables independientes de la población 1.
  - Una muestra formada por n_2 variables independientes de la población 2.
- Suponemos que las muestras son **independientes**.

Contraste bilateral (hipótesis nula simple)

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando las proporciones de individuos que cumplen la característica de interés en cada muestra sean muy diferentes entre si.

Contrastes sobre dos proporciones

- Ejemplo: ¿Existen diferencias significativas entre los porcentajes de usuarios de internet que utilizan la red para consultar información sobre resultados de pruebas médicas en EEUU y en España?
- Realizamos un estudio en España en el que encuestamos a 3000 usuarios de internet. Del total, 520 afirman consultar información sobre resultados de pruebas médicas. En EEUU se ha realizado un estudio análogo sobre 4500 individuos, de los cuales 720 afirmaron consultar información sobre resultados de pruebas médicas.

Contraste bilateral (hipótesis nula simple)

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

siendo p_1 y p_2 las probabilidades de que un usuario de internet consulte información sobre resultados de pruebas médicas en España y EEUU, respectivamente.

Contrastes sobre dos proporciones

- De los n_1 individuos de la población 1, m_1 presentan la característica de interés.

$$\hat{p}_1 = \frac{m_1}{n_1}$$

- De los n_2 individuos de la población 2, m_2 presentan la característica de interés.

$$\hat{p}_2 = \frac{m_2}{n_2}$$

- Bajo H_0, las proporciones poblacionales son iguales, y por tanto estimaremos la proporción común mediante:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}$$

Rechazamos la hipótesis nula $H_0 : p_1 = p_2$ frente a $H_1 : p_1 \neq p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Contrastes sobre dos proporciones

- Ejemplo: ¿Existen diferencias significativas entre los porcentajes de usuarios de internet que utilizan la red para consultar información sobre resultados de pruebas médicas en EEUU y en España?
- Realizamos un estudio en España en el que encuestamos a 3000 usuarios de internet. Del total, 520 afirman consultar información sobre resultados de pruebas médicas. En EEUU se ha realizado un estudio análogo sobre 4500 individuos, de los cuales 720 afirmaron consultar información sobre resultados de pruebas médicas.

Rechazamos la hipótesis nula $H_0 : p_1 = p_2$ frente a $H_1 : p_1 \neq p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> n1 <- 3000  # Tamaño muestral 1
> exitos1 <- 520  # Número de éxitos 1
> phat1 <- exitos1/n1  # Proporción muestral 1
> n2 <- 4500  # Tamaño muestral 2
> exitos2 <- 720  # Número de éxitos 2
> phat2 <- exitos2/n2  # Proporción muestral 2
> phat <- (exitos1 + exitos2)/(n1 + n2)  # Proporción estimada global
> est <- (phat1 - phat2)/sqrt(phat * (1 - phat)/n1 + phat * (1 - phat)/n2)
> est

[1] 1.523

> 2 * (1 - pnorm(est))  # p-valor

[1] 0.1278
```

Contrastes sobre dos proporciones

Rechazamos la hipótesis nula $H_0 : p_1 \leq p_2$ frente a $H_1 : p_1 > p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Rechazamos la hipótesis nula $H_0 : p_1 \geq p_2$ frente a $H_1 : p_1 < p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \leq -z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Ejercicio

Deseamos contrastar si la tasa de respuesta de una población a un determinado tratamiento difiere de 50%. Fijado un nivel de significación $\alpha = 0.05$, ¿cuál debería ser el tamaño muestral para tener una potencia del 80% para detectar que la hipótesis no se cumple si en realidad la tasa de respuesta es del 60%?

Contenidos: Tests Chi-cuadrado

▶ Índice del curso

Tests Chi-cuadrado

- Hasta este momento hemos supuesto que la muestra aleatoria $X_1, \ldots, X_n$ con la que contamos procedía de una población con cierto modelo de probabilidad (por ejemplo, normal) y que desconocíamos el valor de los parámetros de la distribución.
- Sin embargo, en ocasiones desconocemos el tipo de modelo que sigue la variable objeto de estudio.
- Cuando las hipótesis se realizan sobre el modelo, y no sobre algun parámetro, nos encontramos con los contrastes no paramétricos.
- En esta sección analizaremos algunos de los contrastes de este tipo, los llamados tests Chi-cuadrado.
- Al construir la región crítica asociada a dichos contrastes aparecerán estadísticos que seguirán, aproximadamente, una distribución χ^2 de Pearson.

Tests de bondad de ajuste

- Uno de los problemas más importantes con el que nos encontramos en la práctica es el de contrastar la validez de un modelo.
- Por ejemplo, nos gustaría saber si un conjunto de medidas $X_1, \ldots, X_n$ proceden de una variable X con distribución normal.
- Los **tests de bondad de ajuste** tienen por objetivo determinar si los datos se ajustan a una determinada distribución. Dicha distribución:
  - puede estar completamente especificada (hipótesis simple) o
  - puede pertenecer a una clase paramétrica (hipótesis compuesta)

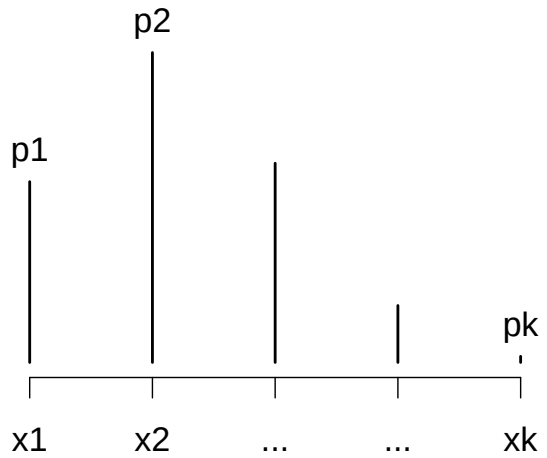
Test de bondad de ajuste Chi-cuadrado

- Una variable aleatoria X viene determinada su función de distribución F.

$$F(x) = P(X \leq x), \ x \in \mathbb{R}$$

- Para presentar el test de bondad de ajuste Chi-cuadrado, partimos en primer lugar de una variable X discreta, que puede tomar únicamente k valores, $x_1, x_2, \ldots, x_k$.
- Si X es discreta, podremos calcular su función de masa.

X	Función de masa
x_1	$p_1 = P(X = x_1)$
x_2	$p_2 = P(X = x_2)$
$\vdots$	$\vdots$
x_k	$p_k = P(X = x_k)$



Test de bondad de ajuste Chi-cuadrado

Hipótesis: ¿Es razonable admitir a la vista de una muestra que la distribución F de la variable de la cual procede es una distribución F_0 determinada?

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

- Suponemos que X es una variable discreta que toma valores $x_1, x_2, \ldots, x_k$.
- Tomamos una muestra aleatoria de tamaño n.
- Sean $n_1, n_2 \ldots, n_k$ las frecuencias observadas de las modalidades $x_1, x_2, \ldots, x_k$, respectivamente $(n_1 + n_2 + \ldots, +n_k = n)$.

Contraste de hipótesis

$$H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$$
$$H_1 : (p_1, p_2, \ldots, p_k) \neq (p_1^0, p_2^0, \ldots, p_k^0)$$

- El sentido común nos aconseja rechazar la hipótesis nula cuando las frecuencias observadas en la muestran disten mucho de las frecuencias que esperaríamos encontrar si la hipótesis nula fuese cierta.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: El administrador de la página web del CITIUS quiere determinar si el número de accesos a la web varía según el día de la semana.
- Para ello, registra el número de visitas a la página durante una semana obteniendo las siguientes frecuencias:

Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
20	14	18	17	22	29	25

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : (p_1, p_2, \ldots, p_7) = (1/7, 1/7, \ldots, 1/7)$$
$$H_1 : (p_1, p_2, \ldots, p_7) \neq (1/7, 1/7, \ldots, 1/7)$$

siendo p_i la probabilidad de que un usuario acceda a la página web del CITIUS el día i de la semana $(i = 1, \ldots, 7)$.

Test de bondad de ajuste Chi-cuadrado: razón de verosimilitudes

Contraste de hipótesis

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

- Supongamos que X es una característica de la población con función de masa P_θ (caso discreto).
- Para cada posible muestra, $(x_1, \ldots, x_n)$ se considera el siguiente cociente:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)}$$

- El test de razón de verosimilitudes para este contraste, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : \quad \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)} \leq c \right\}$$

donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

La distribución multinomial

- Sea $(N_1, N_2, \ldots, N_k)$ el vector aleatorio que cuenta el número de observaciones de la muestra en cada una de las modalidades $x_1, x_2, \ldots, x_k$.
- Se tiene que

$$P(N_1 = n1, N_2 = n_2, \ldots, N_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

- Se dice que $(N_1, N_2, \ldots, N_k)$ sigue una distribución multinomial de parámetros n y $p = (p_1, \ldots, p_k)$.
- Observa que $p_1 + \ldots + p_k = 1$. Por tanto $p_k = 1 - (p_1 + p_2 + \ldots + p_{k-1})$ y el espacio parámetrico tiene dimensión $k - 1$.

Test de bondad de ajuste Chi-cuadrado: razón de verosimilitudes

Contraste de hipótesis

$$H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$$
$$H_1 : (p_1, p_2, \ldots, p_k) \neq (p_1^0, p_2^0, \ldots, p_k^0)$$

- Sean $n_1, n_2 \ldots, n_k$ las frecuencias observadas de las modalidades $x_1, x_2, \ldots, x_k$, respectivamente $(n_1 + n_2 + \ldots, +n_k = n)$.
- En este caso:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P_\theta(x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} P_\theta(x_1, \ldots, x_n)} = \frac{(p_1^0)^{n_1}(p_2^0)^{n_2} \cdots (p_k^0)^{n_k}}{(\hat{p}_1)^{n_1}(\hat{p}_2)^{n_2} \cdots (\hat{p}_k)^{n_k}}$$

siendo

$$\hat{p}_i = \frac{n_i}{n}, \quad i = 1, \ldots, k$$

- El test de razón de verosimilitudes para este contraste, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : -2\log(\Lambda) = 2\sum_{i=1}^{k} n_i \left(log(\hat{p}_i) - \log(p_i^0) \right) \geq c \right\}$$

donde c se obtiene de la condición:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R).$$

Test de bondad de ajuste Chi-cuadrado: razón de verosimilitudes

Contraste de hipótesis

$$H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$$
$$H_1 : (p_1, p_2, \ldots, p_k) \neq (p_1^0, p_2^0, \ldots, p_k^0)$$

- Si H_0 es cierta, $-2\log(\Lambda)$ sigue asintóticamente una distribución χ^2_{k-1}.
- Por lo tanto, el test de razón de verosimilitudes para este contraste, al nivel de significación α, es el que tiene como región crítica:

$$R = \left\{ (x_1, \ldots, x_n) : 2 \sum_{i=1}^{k} n_i \left(\log(\hat{p}_i) - \log(p_i^0) \right) \geq \chi^2_\alpha \right\}$$

con χ^2_α tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k-1$ g.l.

Test de bondad de ajuste Chi-cuadrado: test clásico

Contraste de hipótesis

$$H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$$
$$H_1 : (p_1, p_2, \ldots, p_k) \neq (p_1^0, p_2^0, \ldots, p_k^0)$$

- Sin embargo, la tradición estadística, iniciada por Pearson en 1900 antes del desarrollo de los test de razón de verosimilitudes, hace que normalmente no se emplee el estadístico $-2\log(\Lambda)$
- Lo usual es sustituirlo por una medida de la discrepancia entre las $\hat{p}_i$ y las p_i^0

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}$$

- Si H_0 es cierta, el estadístico también sigue asintóticamente una distribución χ^2_{k-1}.
- La aproximación es buena cuando $np_i^0 \geq 5$ para cada $i = 1, \ldots, k$.

Rechazamos $H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$ si

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k-1$ g.l.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: El administrador de la página web del CITIUS quiere determinar si el número de accesos a la web varía según el día de la semana.
- Registra el número de visitas a la página durante una semana obteniendo las frecuencias:

Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
20	14	18	17	22	29	25

Rechazamos $H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$ **si**

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k-1$ g.l.

```
> p0 <- rep(1/7, 7)  # Hipótesis nula
> ni <- c(20, 14, 18, 17, 22, 29, 25)  # Frecuencias observadas
> n <- sum(ni)  # Tamaño muestral
> ei <- n * p0  # Frecuencias esperadas: n * p0 = 20.71
> sum((ni - ei)^2/ei)  # Estadístico Chi-cuadrado

[1] 7.503

> alpha <- 0.05
> qchisq(1 - alpha, 6)  # k - 1 = 6

[1] 12.59
```

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: El administrador de la página web del CITIUS quiere determinar si el número de accesos a la web varía según el día de la semana.
- Para ello, registra el número de visitas a la página durante una semana obteniendo las siguientes frecuencias:

Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
20	14	18	17	22	29	25

Rechazamos $H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$ **si**

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k-1$ g.l.

- También podemos calcular el p-valor del contraste:

```
> p0 <- rep(1/7, 7)  # Hipótesis nula
> ni <- c(20, 14, 18, 17, 22, 29, 25)  # Frecuencias observadas
> n <- sum(ni)  # Tamaño muestral
> ei <- n * p0  # Frecuencias esperadas
> est <- sum((ni - ei)^2/ei)
> 1 - pchisq(est, 6)  # p-valor

[1] 0.2768
```

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: El administrador de la página web del CITIUS quiere determinar si el número de accesos a la web varía según el día de la semana.
- Para ello, registra el número de visitas a la página durante una semana obteniendo las siguientes frecuencias:

Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
20	14	18	17	22	29	25

Rechazamos $H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$ **si**

$$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k-1$ g.l.

- Podemos utilizar la función `chisq.test`:

```
> ni <- c(20, 14, 18, 17, 22, 29, 25)  # Frecuencias observadas
> chisq.test(ni)


Chi-squared test for given probabilities

data:  ni
X-squared = 7.503, df = 6, p-value = 0.2768
```

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro $\lambda = 2$. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

nº de caídas del servidor	0	1	2	3	4	5	6	7
nº de días (n_i)	22	53	58	39	20	5	2	1

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

siendo F_0 la distribución de una Poisson de parámetro $\lambda = 2$.

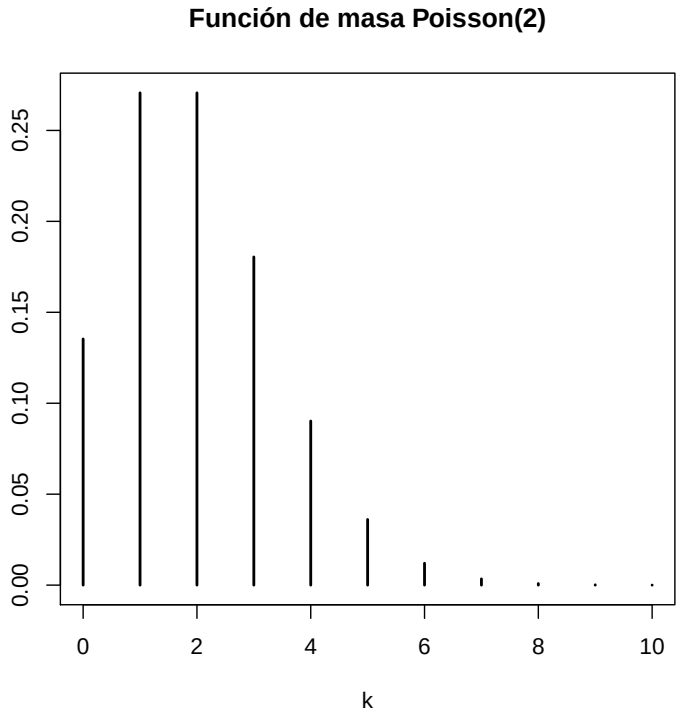
Test de bondad de ajuste Chi-cuadrado

- Una variable X con distribución de Poisson de parámetro λ tiene función de masa

$$P(X = k) = \exp^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

- En Particular, si X sigue una distribución de Poisson de parámetro $\lambda = 2$:

X	$P(X = k)$
0	$P(X = 0) = 0.1353$
1	$P(X = 1) = 0.2706$
2	$P(X = 2) = 0.2706$
⋮	⋮
6	$P(X = 6) = 0.0120$
7	$P(X = 7) = 0.0034$
8	$P(X = 8) = 0.0008$
⋮	⋮



Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro $\lambda = 2$. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

número de caídas del servidor	0	1	2	3	4	5	6	7
número de días	22	53	58	39	20	5	2	1

- Por tanto, podemos calcular las frecuencias esperadas bajo H_0 (el número de caídas diarias sigue una distribución de Poisson de parámetro $\lambda = 2$).

```
> ni <- c(22, 53, 58, 39, 20, 5, 2, 1)
> n <- sum(ni)
> p0 <- dpois(0:7, 2)
> ei <- n * p0
> ei

[1] 27.07 54.13 54.13 36.09 18.04  7.22  2.41  0.69
```

nº de caídas	0	1	2	3	4	5	6	7
nº de días (n_i)	22	53	58	39	20	5	2	1
$n_i p_i^0$	27.07	54.13	54.13	36.09	18.04	7.22	2.41	0.69

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro $\lambda = 2$. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

número de caídas del servidor	0	1	2	3	4	5	6	7
número de días	22	53	58	39	20	5	2	1

- Si queremos que $n_i p_i^0 \geq 5$ debemos agrupar los valores más extremos. Por ejemplo, consideramos como última categoría es "5 accidentes o más".

```
> ni <- c(22, 53, 58, 39, 20, 8)
> n <- sum(ni)
> p0 <- c(dpois(0:4, 2), 1 - ppois(4, 2))
> ei <- n * p0
> ei <- round(n * p0, 2)
> ei

[1] 27.07 54.13 54.13 36.09 18.04 10.53
```

nº de caídas	0	1	2	3	4	5 o más
nº de días (n_i)	22	53	58	39	20	8
$n_i p_i^0$	27.07	54.13	54.13	36.09	18.04	10.53

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro $\lambda = 2$. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

nº de caídas	0	1	2	3	4	5 o más
nº de días (n_i)	22	53	58	39	20	8
$n_i p_i^0$	27.07	54.13	54.13	36.09	18.04	10.53

```
> ni <- c(22, 53, 58, 39, 20, 8)
> n <- sum(ni)
> p0 <- c(dpois(0:4, 2), 1 - ppois(4, 2))
> ei <- n * p0
> est <- sum((ni - ei)^2/ei)  # Estadístico Chi-cuadrado
> est

[1] 2.303

> 1 - pchisq(est, 5)  # p-valor

[1] 0.8058
```

- Por lo tanto, no existe evidencia significativa para rechazar la hipótesis nula.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro $\lambda = 2$. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

nº de caídas	0	1	2	3	4	5 o más
nº de días (n_i)	22	53	58	39	20	8
$n_i p_i^0$	27.07	54.13	54.13	36.09	18.04	10.53

- Utilizando la función `chisq.test`

```
> ni <- c(22, 53, 58, 39, 20, 8)
> p0 <- c(dpois(0:4, 2), 1 - ppois(4, 2))
> chisq.test(ni, p = p0)


Chi-squared test for given probabilities

data:  ni
X-squared = 2.303, df = 5, p-value = 0.8058
```

Test de bondad de ajuste Chi-cuadrado

- Aunque el test de bondad de ajuste Chi-cuadrado es especialmente adecuado para distribuciones discretas, también se puede aplicar para contrastar si a la vista de una muestra de una variable continua la distribución F de la cual procede es una distribución F_0 determinada.
- En el caso de que la distribución sea continua, se divide la recta real en k intervalos $A_1, A_2, \ldots, A_k$. De este modo, se pueden calcular las frecuencias de observaciones que pertenecen a cada A_i, y llevar a cabo el test Chi-cuadrado igual que en el caso discreto.
- Para que el test sea capaz de distinguir si la distribución verdadera es F_0 o no, conviene considerar muchos conjuntos A_i.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones obteniendo los siguientes errores de medición (en décimas de grado):

  1.2, 2.3, -1.4, -0.4, -0.6, 3.2, 3.9, -2.5, 0.8,
  -0.1, 1.3, 0.2, 3.8, 4.1, -2.6,2.4, -4.1, -2.6, 0.6,
  -0.3, 1.5, 1.9, -2.7, -2.4, -3.7, 0.7, -0.2, 0.5, -1.2, 1.7

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

siendo F_0 la distribución de una normal de media $\mu = 0$ y desviación típica $\sigma = 3$.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones obteniendo los siguientes errores de medición (en décimas de grado). Los datos se encuentran en el fichero `termometro.txt`.

```
> x <- scan("termometro.txt")
> # Dividimos la recta real en intervalos
> ni <- table(cut(x, c(-Inf, -2, 0, 2, Inf)))  # [-inf,-2],[-2,0],[0,2],[2,inf]
> ni


(-Inf,-2]    (-2,0]     (0,2]  (2, Inf]
        7         7        10         6

> n <- sum(ni)
> # Calculamos las probabilidades de cada intervalo bajo H0
> p0 <- numeric()
> p0[1] <- pnorm(-2, 0, 3)  # P[X < -2], X normal de media 0 y desviación 3
> p0[2] <- pnorm(0, 0, 3) - pnorm(-2, 0, 3)  # P[-2 < X < 0]
> p0[3] <- pnorm(2, 0, 3) - pnorm(0, 0, 3)  # P[ 0 < X < 2]
> p0[4] <- 1 - pnorm(2, 0, 3)  # P[X > 2]
> p0

[1] 0.2525 0.2475 0.2475 0.2525
```

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones obteniendo los siguientes errores de medición (en décimas de grado). Los datos se encuentran en el fichero `termometro.txt`.

```
> ei <- n * p0  # Frecuencias esperadas de cada intervalo
> ei

[1] 7.575 7.425 7.425 7.575

> est <- sum((ni - ei)^2/ei)  # Estadístico Chi-cuadrado
> est

[1] 1.288

> 1 - pchisq(est, 3)  # p-valor

[1] 0.7319
```

- Por tanto, no rechazamos la hipótesis nula (Las discrepancias respecto de H_0 no son significativas a ningún nivel de significación razonable).

Test de bondad de ajuste Chi-cuadrado

Hipótesis: ¿Es razonable admitir a la vista de una muestra que la distribución F de la variable de la cual procede pertenece a una familia de distribuciones F_θ determinada?

Contraste de hipótesis

$$H_0 : F \in \{F_\theta / \theta \in \Theta\}$$
$$H_1 : F \notin \{F_\theta / \theta \in \Theta\}$$

- El test Chi-cuadrado también se puede emplear para contrastar una hipótesis nula de este tipo, en la que nos planteamos si la distribución pertenece al modelo paramétrico representado por F_θ, donde θ representa el/los parámetro/s del modelo.
- Gran parte de las ideas coinciden con el caso anterior, referido a una hipótesis nula simple. Así, suponemos disponible una muestra de observaciones independientes.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson de parámetro λ. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

nº de caídas del servidor	0	1	2	3	4	5	6	7
nº de días (n_i)	22	53	58	39	20	5	2	1

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : F \in \{F_\lambda / \lambda \in \mathbb{R}\}$$
$$H_1 : F \notin \{F_\lambda / \lambda \in \mathbb{R}\}$$

donde F_λ denota la distribución de una Poisson de parámetro λ.

Test de bondad de ajuste Chi-cuadrado

Contraste de hipótesis

$$H_0 : F \in \{F_\theta / \theta \in \Theta\}$$
$$H_1 : F \notin \{F_\theta / \theta \in \Theta\}$$

- Ahora para calcular las frecuencias esperadas bajo la hipótesis nula, necesitamos estimar el valor del parámetro θ. Consideremos entonces un estimador adecuado $\hat{\theta}$ y a continuación calcularemos las probabilidades $p_1(\hat{\theta}), p_2(\hat{\theta}), \ldots, p_k(\hat{\theta})$ bajo la distribución $F_{\hat{\theta}}$.
- Las frecuencias esperadas serán $np_1(\hat{\theta}), np_2(\hat{\theta}), \ldots, np_k(\hat{\theta})$.
- El estadístico Chi-cuadrado se construye de igual modo,

$$\sum_{i=1}^{k} \frac{(n_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

- De nuevo rechazaremos cuando el estadístico tome un valor grande, pero ahora su distribución bajo la hipótesis nula es diferente, como consecuencia de haber tenido que estimar θ. En este caso se aproxima también por una distribución Chi-cuadrado pero el número de grados de libertad será $k - q - 1$ donde q representa el número de parámetros que hemos tenido que estimar.

Test de bondad de ajuste Chi-cuadrado

Contraste de hipótesis

$$H_0 : F \in \{F_\theta / \theta \in \Theta\}$$
$$H_1 : F \notin \{F_\theta / \theta \in \Theta\}$$

- Consideremos entonces un estimador adecuado $\hat{\theta}$ y a continuación calcularemos las probabilidades $p_1(\hat{\theta}), p_2(\hat{\theta}), \ldots, p_k(\hat{\theta})$ bajo la distribución $F_{\hat{\theta}}$.

Rechazamos $H_0 : F \in \{F_\theta / \theta \in \Theta\}$ si

$$\sum_{i=1}^{k} \frac{(n_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $k - q - 1$ g.l.

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.

número de caídas del servidor	0	1	2	3	4	5	6	7
número de días	22	53	58	39	20	5	2	1

- En primer lugar estimamos el parámetro λ

```
> Ai <- 0:7
> ni <- c(22, 53, 58, 39, 20, 5, 2, 1)
> n <- sum(ni)
> lambdahat <- sum(Ai * ni)/n  # Número medio de caídas diarias
> lambdahat

[1] 2.05
```

- Por tanto, podemos calcular las frecuencias esperadas suponiendo que el número de caídas diarias sigue una distribución de Poisson de parámetro $\hat{\lambda}$.

```
> p0 <- dpois(0:7, lambdahat)
> ei <- n * p0
> ei

[1] 25.75 52.78 54.10 36.97 18.95  7.77  2.65  0.78
```

Test de bondad de ajuste Chi-cuadrado

- Ejemplo: Análisis previos sugieren que el número de caídas diarias de un servidor de PSN sigue una distribución de Poisson. Se decide comprobar esta hipótesis y para ello se registran las caídas del servidor ocurridas a lo largo de 200 días.
- Si queremos que $n_i p_i(\hat{\lambda}) \geq 5$ debemos agrupar los valores más extremos. Por ejemplo, consideramos como última categoría es "5 accidentes o más".

```
> ni <- c(22, 53, 58, 39, 20, 8)
> n <- sum(ni)
> p0 <- c(dpois(0:4, lambdahat), 1 - ppois(4, lambdahat))
> ei <- n * p0
> ei <- round(n * p0, 2)
> ei

[1] 25.75 52.78 54.10 36.97 18.95 11.46

> est <- sum((ni - ei)^2/ei)
> est

[1] 2.042

> 1 - pchisq(est, 4)  # k - q - 1 = 6 - 1 - 1 = 4 g.l

[1] 0.7279
```

Tests de independencia

Hipótesis: ¿Es razonable admitir en base a la observación de 2 características en n individuos que dichas características son independendientes?

- Supongamos que de n elementos de una población se han observado dos características X e Y, obteniéndose una muestra aleatoria simple bidimensional $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
- Sobre la base de estas observaciones se desea contrastar si las características poblacionales X e Y son independientes o no.

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

Test de independencia

- Ejemplo: Se hizo un estudio consistente en experimentar la efectividad de dos configuraciones de procesador sobre el rendimiento de 165 equipos. Se registró el tipo de configuración de cada equipo y su rendimiento (muy bajo, bajo, moderado o alto).
  - Se registraron 83 equipos con configuración A:
    - 12 presentan rendimiento muy bajo,
    - 24 presentan rendimiento bajo,
    - 31 presentan rendimiento moderado,
    - 16 presentan rendimiento alto.
  - Se registraron 82 equipos con configuración B,
    - 20 presentan rendimiento muy bajo,
    - 18 presentan rendimiento bajo,
    - 30 presentan rendimiento moderado,
    - 14 presentan rendimiento alto.
- Queremos determinar si la configuración está relacionada con el tipo de rendimiento. Es decir, planteamos un contraste:

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

siendo X el tipo de configuración e Y el rendimiento alcanzado.

Tests de independencia

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

- Siguiendo la idea de los tests tipo Chi-cuadrado, los valores posibles de X se agrupan en k conjuntos: $A_1, A_2, \ldots, A_k$; mientras que los valores de Y lo hacen en los conjuntos: $B_1, B_2, \ldots, B_s$.
- Denotamos n_{ij} al número de observaciones cuyo valor de X se encuentra en el conjunto A_i, y cuyo valor de Y se encuentra en el conjunto B_j. Por tanto, los n_{ij} representan frecuencias observadas, y conforman una **tabla de contingencia** de la siguiente naturaleza:

$X \backslash Y$	B_1	B_2	$\cdots$	B_r	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kr}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	n

donde n es el tamaño muestral.

Tests de independencia

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

$X \backslash Y$	B_1	B_2	$\cdots$	B_r	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kr}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	n

- Sea $p_{ij} = P(X \in A_i, Y \in B_j)$.
- La hipótesis de independencia entre X e Y implica que se cumpla

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \qquad \text{para todo} \quad i \in \{1, 2, \ldots, k\} \quad j \in \{1, 2, \ldots, r\}$$

- Es decir las probabilidades conjuntas p_{ij} han de ser el producto de las marginales $p_{i\cdot}$ y $p_{\cdot j}$.
- Las frecuencias esperadas bajo H_0 serían entonces:

$$\frac{n_{i\cdot} n_{\cdot j}}{n}$$

Tests de independencia

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

$X \backslash Y$	B_1	B_2	$\cdots$	B_r	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kr}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	n

- De este modo, el estadístico tipo Chi-cuadrado para contrastar la hipótesis de independencia adoptaría la forma:

$$\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij}-n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

- Su distribución se aproxima por una Chi-cuadrado con $(k-1)(r-1)$ grados de libertad.

Tests de independencia

Contraste de hipótesis

H_0 : X e Y son independientes
H_1 : X e Y no son independientes

$X \setminus Y$	B_1	B_2	$\cdots$	B_r	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kr}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	n

Rechazamos H_0 : X e Y son independientes si

$$\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij}-n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $(k-1)(r-1)$ g.l.

Test de independencia

- Ejemplo: Se hizo un estudio consistente en experimentar la efectividad de dos configuraciones de procesador sobre el rendimiento de 165 equipos. Se registró el tipo de configuración de cada equipo y su rendimiento (muy bajo, bajo, moderado o alto).

	Rendimiento				
Configuración	Muy bajo	Bajo	Moderado	Alto	Total
A	12	24	31	16	83
B	20	18	30	14	82
Total	32	42	61	30	165

- Queremos determinar si la configuración está relacionada con el tipo de rendimiento. Calculamos los valores esperados de cada celda suponiendo independencia.

	Rendimiento				
Configuración	Muy bajo	Bajo	Moderado	Alto	Total
A	12(16.09)	24(21.12)	31(30.68)	16(15.09)	83
B	20(15.90)	18(20.87)	30(30.31)	14(14.90)	82
Total	32	42	61	30	165

Test de independencia

- Ejemplo: Se hizo un estudio consistente en experimentar la efectividad de dos configuraciones de procesador sobre el rendimiento de 165 equipos. Se registró el tipo de configuración de cada equipo y su rendimiento (muy bajo, bajo, moderado o alto).

	Rendimiento				
Configuración	Muy bajo	Bajo	Moderado	Alto	Total
A	12(16.09)	24(21.12)	31(30.68)	16(15.09)	83
B	20(15.90)	18(20.87)	30(30.31)	14(14.90)	82
Total	32	42	61	30	165

```
> A <- c(12, 24, 31, 16)
> B <- c(20, 18, 30, 14)
> nij <- rbind(A, B)
> chisq.test(nij)


Pearson's Chi-squared test

data:  nij
X-squared = 3.001, df = 3, p-value = 0.3915
```

Tests de independencia. Caso particular de tablas 2×2

- Consideremos ahora el caso particular en que queremos llevar a cabo un contraste de independencia y las características X e Y sólo presentan dos posibles valores o categorías.
- Una tabla de contingencia 2×2 está formada por dos filas y dos columnas.

	Variable 1		
Variable 2	Valor 1	Valor 2	Total
Valor 1	a	b	a + b
Valor 2	c	d	c + d
Total	a + c	b + d	a + b + c + d

- Por lo visto anteriormente:

Rechazamos H_0 : X e Y son independientes si

$$\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij}-n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con 1 g.l.

- En tablas 2×2 la aproximación a la Chi-cuadrado puede no ser buena y, por eso, se suele aplicar la llamada **corrección por continuidad de Yates**.

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|n_{ij}-n_{i.}n_{.j}/n|-0.5)^2}{n_{i.}n_{.j}/n}.$$

Tests de homogeneidad

Hipótesis: ¿Es razonable admitir a partir de m muestras aleatorias simples de otras tantas poblaciones que la distribución poblacional es la misma en todos los casos, o por el contrario se trata de poblaciones heterogéneas con diferentes distribuciones?

- Se dispone de m muestras aleatorias simples, cuyos tamaños son respectivamente $n_1, n_2, \ldots, n_m$.
- Sobre la base de estas observaciones se desea contrastar si la distribución poblacional es la misma en todos los casos, o por el contrario si se trata de poblaciones heterogéneas con diferentes distribuciones.

Contraste de hipótesis

H_0 : Las m poblaciones son homogéneas
H_1 : Las m poblaciones no son homogéneas

Test de homogeneidad

- Ejemplo: Estamos interesados en estudiar la fiabilidad de cierto componente informático con relación al distribuidor que nos lo suministra. Para realizar esto, tomamos una muestra de 100 componentes de cada uno de los 3 distribuidores que nos sirven el producto comprobando el número de defectuosos en cada lote. Obtenemos los siguientes resultados:

	Componentes defectuosos	Componentes correctos	
Distribuidor 1	16	84	100
Distribuidor 2	24	76	100
Distribuidor 3	9	91	100
	49	251	300

- Queremos determinar si entre los distribuidores existen diferencias de fiabilidad referente al mismo componente.

Tests de homogeneidad

Contraste de hipótesis

H_0 : m poblaciones son homogéneas
H_1 : m poblaciones no son homogéneas

- De nuevo, resolveremos mediante un test de tipo Chi-cuadrado.
- Así, dividiremos el conjunto de valores posibles en k conjuntos $A_1, A_2, \ldots, A_k$ y clasificaremos en ellos las observaciones de cada muestra.
- Representamos los valores observados mediante la siguiente tabla de contingencia:

Muestra	A_1	A_2	$\cdots$	A_k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	n_m
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

Tests de homogeneidad

Contraste de hipótesis

H_0 : m poblaciones son homogéneas
H_1 : m poblaciones no son homogéneas

Muestra	A_1	A_2	$\cdots$	A_k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	n_m
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

- La hipótesis H_0 se traduce en que cada conjunto A_j debe tener una probabilidad teórica p_j, desconocida, pero que no varía de una población a otra.
- Así, si p_{ij} es la probabilidad que presenta A_j en la población i-ésima, el contraste de homogeneidad se puede plantear como el contraste de la hipótesis nula

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{mj} \quad (= p_j) \qquad \text{para todo} \quad j \in \{1, \ldots, k\}$$

Tests de homogeneidad

Contraste de hipótesis

H_0 : m poblaciones son homogéneas
H_1 : m poblaciones no son homogéneas

Muestra	A_1	A_2	$\cdots$	A_k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	n_m
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

- Para calcular las frecuencias esperadas, pensemos que bajo la hipótesis nula hay un única distribución de probabilidad $p_1, p_2, \ldots, p_k$ y

$$\hat{p}_j = \frac{n_{\cdot j}}{n}$$

- Entonces las frecuencias esperadas se calculan como:

$$n_i \hat{p}_j = \frac{n_i n_{\cdot j}}{n}$$

Tests de homogeneidad

Contraste de hipótesis

H_0 : m poblaciones son homogéneas
H_1 : m poblaciones no son homogéneas

Muestra	A_1	A_2	$\cdots$	A_k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	n_m
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

- Entonces, el estadístico Chi-cuadrado adopta la forma

$$\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij}-n_i n_{\cdot j}/n)^2}{n_i n_{\cdot j}/n}$$

- La distribución del estadístico bajo H_0 se aproxima por una Chi-cuadrado, en este caso con $(m-1)(k-1)$ grados de libertad.

Tests de homogeneidad

Contraste de hipótesis

H_0 : m poblaciones son homogéneas
H_1 : m poblaciones no son homogéneas

Muestra	A_1	A_2	$\cdots$	A_k	Total
1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	n_m
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

Rechazamos H_0 : m poblaciones son homogéneas **si**

$$\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij}-n_i n_{\cdot j}/n)^2}{n_i n_{\cdot j}/n} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Chi-cuadrado con $(m-1)(k-1)$ g.l.

Test de homogeneidad

- Ejemplo: Estamos interesados en estudiar la fiabilidad de cierto componente informático con relación al distribuidor que nos lo suministra. Para realizar esto, tomamos una muestra de 100 componentes de cada uno de los 3 distribuidores que nos sirven el producto comprobando el número de defectuosos en cada lote. Obtenemos los siguientes resultados:

	Componentes defectuosos	Componentes correctos	
Distribuidor 1	16	84	100
Distribuidor 2	24	76	100
Distribuidor 3	9	91	100
	49	251	300

- Queremos determinar si entre los distribuidores existen diferencias de fiabilidad referente al mismo componente.

```
> D1 <- c(16, 94)
> D2 <- c(24, 76)
> D3 <- c(9, 81)
> nij <- rbind(D1, D2, D3)
> chisq.test(nij)


Pearson's Chi-squared test

data:  nij
X-squared = 7.2, df = 2, p-value = 0.02732
```

- Rechazamos H_0 ($\alpha = 0.05$). Existe evidencia significativa de que existen diferencias de fiabilidad entre los distribuidores.

Test de homogeneidad

- Como hemos visto, existe una gran similitud entre el test Chi-cuadrado de homogeneidad y el test Chi-cuadrado de independencia.
- De hecho, el método de resolución es idéntico en ambos casos.
- La diferencia estriba en la interpretación del problema.
- Desde un punto de vista técnico, los totales por fila y columna en el contraste de independencia son aleatorios, mientras que están fijados por el experimentador cuando se plantea un contraste de homogeneidad.

Contenidos: Test de Kolmogorov–Smirnov y otros tests de bondad de ajuste

▶ Índice del curso

Test de Kolmogorov–Smirnov

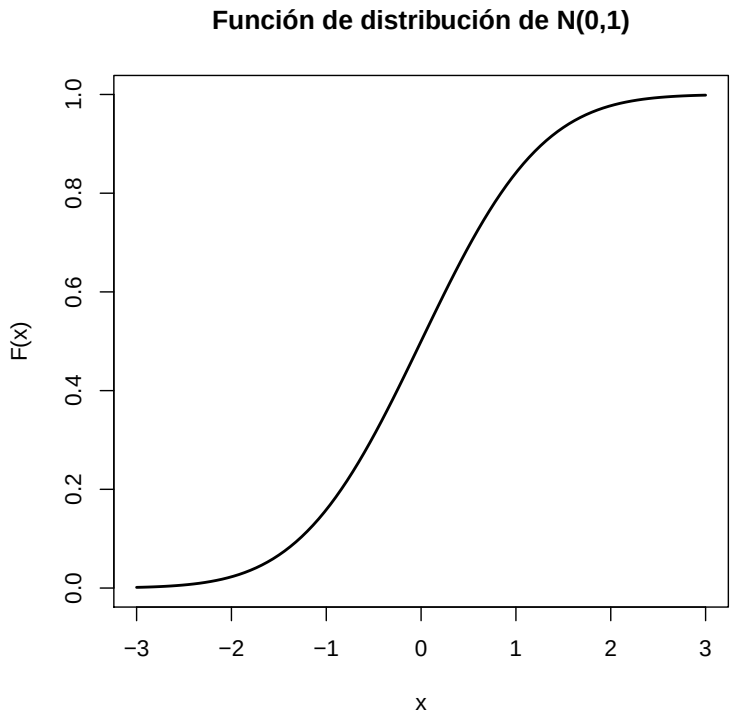
- El test de Kolmogorov-Smirnov es otro método que podemos emplear para efectuar el contraste de bondad de ajuste de la distribución.
- El test de Kolmogorov-Smirnov se aplica sólo al caso continuo, y en lugar de tomar una partición de intervalos disjuntos (como hace el test Chi-cuadrado para el caso de distribuciones continuas), consiste en tomar intervalos crecientes de izquierda a derecha.
- Después se observa si las frecuencias observadas en los intervalos crecientes (frecuencias acumuladas), discrepan mucho o poco de las esperadas.

Test de Kolmogorov–Smirnov

- Una variable aleatoria X viene determinada su función de distribución F.

$$F(x) = P(X \leq x), \ x \in \mathbb{R}$$

- Representamos como ejemplo la función de distribución de una variable normal estándar:



Test de Kolmogorov–Smirnov

Hipótesis: ¿Es razonable admitir a la vista de una muestra que la distribución F de la variable de la cual procede es una distribución F_0 determinada?

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

- Suponemos que X es una variable continua.
- Tomamos una muestra aleatoria de tamaño n.

Test de Kolmogorov–Smirnov

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones obteniendo los siguientes errores de medición (en décimas de grado):

  1.2, 2.3, -1.4, -0.4, -0.6, 3.2, 3.9, -2.5, 0.8,
  -0.1, 1.3, 0.2, 3.8, 4.1, -2.6,2.4, -4.1, -2.6, 0.6,
  -0.3, 1.5, 1.9, -2.7, -2.4, -3.7, 0.7, -0.2, 0.5, -1.2, 1.7

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

siendo F_0 la distribución de una normal de media $\mu = 0$ y desviación típica $\sigma = 3$.

La función de distribución empírica

- Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple de una variable X con función de distribución F.
- La función de distribución empírica se define como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\} \qquad x \in \mathbb{R}$$

  esto es, para cada valor de x, es la frecuencia relativa de valores muestrales menores o iguales que x.
- La función de distribución empírica es:
  - Una función de distribución: no decreciente, continua por la derecha, con límites por la izquierda, que parte de cero y llega a uno.
  - Una distribución discreta que presenta saltos de amplitud $1/n$ en las observaciones muestrales.
  - Otorga el mismo peso $1/n$ a cada observación muestral.

Test de Kolmogorov–Smirnov

Contraste de hipótesis

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

- Ya que la distribución empírica, F_n, es un estimador adecuado de F, parece razonable rechazar la hipótesis nula $H_0 : F = F_0$ cuando F_n sea muy distinta de F_0 y aceptarla cuando la discrepancia entre ambas no sea muy grande.
- Una forma de medir la discrepancia entre las funciones F_n y F_0 se obtiene mediante el **estadístico de Kolmogorov–Smirnov**, definido así:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

Test de Kolmogorov–Smirnov

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

- Se define la muestra ordenada como el estadístico $(X_{1:n}, \ldots, X_{n:n})$ resultante de disponer la muestra $(X_1, \ldots, X_n)$ en orden creciente.
- Para el cálculo de D_n en la práctica basta con efectuar la diferencia en los datos muestrales ordenados $X_{j:n}$.

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| = \max(D_n^+, D_n^-)$$

siendo:

  - $D_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x))$,
  - $D_n^- = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x))$.

- Además, D_n^+ y D_n^- se pueden calcular así:

$$D_n^+ = \max_{j \in \{1,\ldots,n\}} \left(\frac{j}{n} - F_0\left(X_{j:n}\right) \right)$$

$$D_n^- = \max_{j \in \{1,\ldots,n\}} \left(F_0\left(X_{j:n}\right) - \frac{j-1}{n} \right)$$

Test de Kolmogorov–Smirnov

- Rechazaremos la hipótesis nula cuando el estadístico de Kolmogorov–Smirnov tome un valor grande.
- Para saber el punto concreto a partir del cual debemos rechazar la hipótesis nula, es preciso conocer la distribución del estadístico.
- Se ha demostrado que para cualquier distribución F_0 continua, el estadístico de Kolmogorov–Smirnov tiene la misma distribución (que está tabulada).

Rechazamos $H_0 : F = F_0$ si

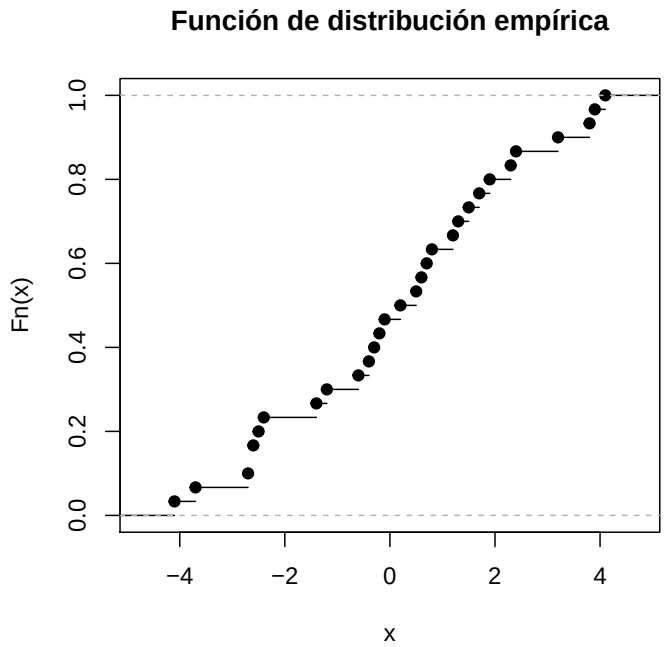
$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \geq d_{n,\alpha}$$

$d_{n,\alpha}$ denota el punto tal que $P(D_n > d_{n,\alpha}) = \alpha$

Test de Kolmogorov–Smirnov

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones. Los datos se encuentran en el fichero `termometro.txt`.

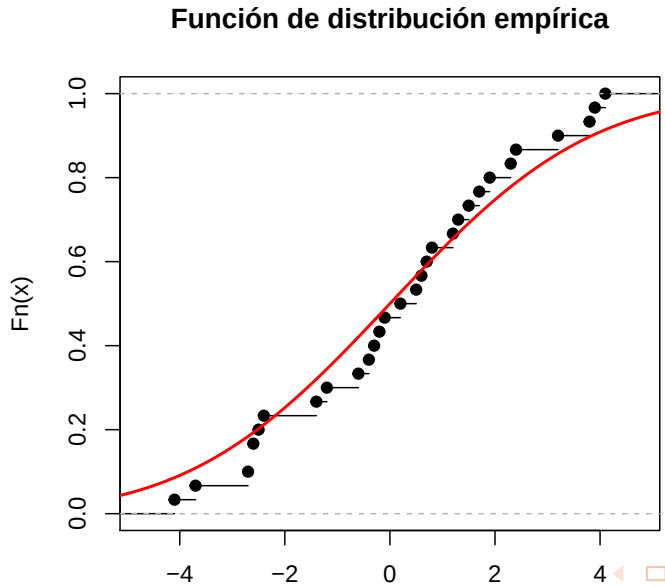
```
> x <- scan("termometro.txt")
> plot(ecdf(x), main = "Función de distribución empírica")
```



Test de Kolmogorov–Smirnov

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones. Los datos se encuentran en el fichero `termometro.txt`.

```
> x <- scan("termometro.txt")
> plot(ecdf(x), main = "Función de distribución empírica")
> eje <- seq(-6, 6, length = 1000)
> # Añado la distribución de la normal con mu = 0 y sigma = 3
> lines(eje, pnorm(eje, 0, 3), col = 2, lwd = 2)
```



Test de Kolmogorov–Smirnov

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal de media cero y desviación típica 3 décimas de grado. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones. Los datos se encuentran en el fichero `termometro.txt`.
- Llevamos a cabo el contraste con la función `ks.test`:

```
> x <- scan("termometro.txt")
> ks.test(x, "pnorm", 0, 3)

Warning:  ties should not be present for the Kolmogorov-Smirnov test


One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1207, p-value = 0.7743
alternative hypothesis: two-sided
```

Test de Kolmogorov–Smirnov

Hipótesis: ¿Es razonable admitir a la vista de una muestra que la distribución F de la variable de la cual procede pertenece a una familia de distribuciones F_θ determinada?

Contraste de hipótesis

$$H_0 : F \in \{F_\theta / \theta \in \Theta\}$$
$$H_1 : F \notin \{F_\theta / \theta \in \Theta\}$$

- El test de Kolmogorov–Smirnov también se puede adaptar para contrastar una hipótesis nula de este tipo, en la que nos planteamos si la distribución pertenece al modelo paramétrico representado por F_θ, donde θ representa el/los parámetro/s del modelo.
- Gran parte de las ideas coinciden con el caso anterior, referido a una hipótesis nula simple. Así, suponemos disponible una muestra de observaciones independientes.

Test de Kolmogorov–Smirnov

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones obteniendo los siguientes errores de medición (en décimas de grado):

1.2, 2.3, -1.4, -0.4, -0.6, 3.2, 3.9, -2.5, 0.8,
-0.1, 1.3, 0.2, 3.8, 4.1, -2.6,2.4, -4.1, -2.6, 0.6,
-0.3, 1.5, 1.9, -2.7, -2.4, -3.7, 0.7, -0.2, 0.5, -1.2, 1.7

- Queremos contrastar:

Contraste de hipótesis

$$H_0 : F \in \left\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$
$$H_1 : F \notin \left\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$

Test de Kolmogorov–Smirnov

Contraste de hipótesis

$$H_0 : F \in \{F_\theta / \theta \in \Theta\}$$
$$H_1 : F \notin \{F_\theta / \theta \in \Theta\}$$

- En este caso podemos considerar la distancia entre la distribución empírica (estimador no paramétrico de F) y el estimador paramétrico $F_{\hat{\theta}}$:

$$D_n = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_{\hat{\theta}}(x) \right|$$

  donde $\hat{\theta}$ es un estimador adecuado de los parámetros θ.
- El cálculo del estadístico se puede efectuar igual que antes, a través de D_n^+ y D_n^-.
- Sin embargo, la distribución del estadístico ya no es la misma que la de D_n para una distribución completamente especificada.
- En su lugar, se han elaborado tablas con las distribuciones del estadístico para cada modelo de distribución: normal, exponencial, u otros.

Rechazamos $F \in \{F_\theta / \theta \in \Theta\}$ si

$$D_n = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_{\hat{\theta}}(x) \right| \geq d_{n,\alpha}$$

$d_{n,\alpha}$ denota el punto tal que $P(D_n > d_{n,\alpha}) = \alpha$

Test de Lilliefors

Contraste de hipótesis

$$H_0 : F \in \left\{ N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$$
$$H_1 : F \notin \left\{ N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$$

- El test de Lilliefors es la adaptación del test de Kolmogorov-Smirnov al caso en que se contrasta normalidad

Test de Lilliefors

- Ejemplo: En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones. Los datos se encuentran en el fichero `termometro.txt`.
- Utilizaremos la función `lillie.test` incuida en el paquete `nortest`.

```
> library(nortest)
> x <- scan("termometro.txt")
> lillie.test(x)


        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.1064, p-value = 0.5215
```

Test de Kolmogorov–Smirnov para dos muestras

Hipótesis: ¿Es razonable admitir a la vista de dos muestras que las distribuciones F_1 y F_2 de las poblaciones de la cuales proceden son iguales?

Contraste de hipótesis

$$H_0 : F_1 = F_2$$
$$H_1 : F_1 \neq F_2$$

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución.
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).

Test de Kolmogorov–Smirnov para dos muestras

Contraste de hipótesis

$$H_0 : F_1 = F_2$$
$$H_1 : F_1 \neq F_2$$

- Ya que la distribución empírica, F_{n_1}, es un estimador adecuado de F_1, y la la distribución empírica, F_{n_2}, es un estimador adecuado de F_2, parece razonable rechazar la hipótesis nula $H_0 : F_1 = F_2$ cuando las distribuciones empíricas sean muy distintas.
- Una forma de medir la discrepancia entre las funciones F_{n_1} y F_{n_2} se obtiene mediante el **estadístico de Kolmogorov–Smirnov para dos muestras**, definido así:

$$D_n = \sup_{x \in \mathbb{R}} \left| F_{n_1}(x) - F_{n_2}(x) \right|$$

Test de Kolmogorov–Smirnov para dos muestras

```
> x1 <- rnorm(50)  # Normal estándar
> x2 <- rnorm(50, -1)  # Normal de media -1
> ks.test(x1, x2)


Two-sample Kolmogorov-Smirnov test

data:  x1 and x2
D = 0.5, p-value = 4.808e-06
alternative hypothesis: two-sided
```

Tests de bondad de ajuste

- Tanto el test Chi-cuadrado como el test de Kolmogorov–Smirnov resuelven el mismo problema de contraste de bondad de ajuste de la distribución.
- El test Chi-cuadrado se puede aplicar tanto a variables discretas como a variables continuas, aunque en este último caso requiere agrupamiento de los datos.
- El test de Kolmogorov–Smirnov sólo se puede aplicar a variables continuas. En el caso de una variable continua, es más aconsejable emplear el test de Kolmogorov–Smirnov, ya que no requiere agrupamiento de los datos y por lo general resulta más eficaz para detectar los posibles incumplimientos de la hipótesis nula.
- El test de Kolmogorov–Smirnov es capaz de evaluar contrastes unilaterales y permite el cálculo bandas de confianza para la función de distribución.
- En la ejecución práctica del contraste, debemos tener en cuenta que el test Chi-cuadrado se puede usar para el contraste de cualquier modelo paramétrico, sin más que modificar el número de grados de libertad en función del número de parámetros que haya que estimar, mientras que el test de Kolmogorov-Smirnov requiere tablas nuevas para cada modelo paramétrico.

Tests de bondad de ajuste

Hipótesis: ¿Es razonable admitir a la vista de una muestra que la distribución de la variable de la cual procede es normal?

- Aunque tanto el test Chi-cuadrado de bondad de ajuste como el test de Kolmogorov-Smirnov son los que más se utilizan, se han desarrollado otros tests de bondad de ajuste.
- Puesto que uno de los principales supuestos sobre el que se asienta muchos modelos estadísticos es que las variables observadas siguen una distribución normal, se han desarrollado procedimientos específicos para contrastar la normalidad.
- Por ejemplo, existen contrastes para determinar si la forma de la distribución de las observaciones muestrales se aleja significativamente de la de un modelo normal en lo que a su simetría y curtosis se refiere.
- Otro contraste de normalidad muy extendido es el de Shapiro-Wilk, que es considerado como uno de los test más potentes para muestras pequeñas.

Test basado en la asimetría

- La distribución normal es simétrica en torno a su media. Como además, la forma más común de desviarse respecto de la normalidad es por falta de simetría, por ejemplo en variables positivas, parece lógico construir un método de contraste en base a cierta medida de asimetría.
- Si $x_1, \ldots, x_n$ es una muestra aleatoria simple, el coeficiente de asimetría muestral se define como

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

  siendo $\overline{x}$ la media muestral y s^2 la varianza muestral.
- El coeficiente de asimetría es la media de las potencias cúbicas (también llamado momento de orden tres) de los valores estandarizados. Al estar estandarizados, habrá observaciones a ambos lados de la media, manteniendo un equilibrio (en orden uno).
- La potencia tres que se emplea en el coeficiente de asimetría, rompe el equilibrio en caso de distribución asimétrica.

Test basado en la asimetría

- Si la distribución de los datos muestrales es normal, entonces el coeficiente de asimetría tiene distribución asintótica normal de media cero y varianza $6/n$, por lo que se puede emplear como estadístico de contraste el siguiente:

$$\sqrt{\frac{n}{6}}\, A$$

- Se rechazará la normalidad, en base al coeficiente de asimetría, cuando el estadístico anterior sea muy grande (en positivo o negativo), en comparación con los cuantiles de la $N(0,1)$.

Test basado en la asimetría

- Como ejemplo vamos a presentar los resultados del contraste de normalidad basado en asimetría aplicado a datos simulados de distintas distribuciones.
- Mostramos en primer lugar los valores del coeficiente de asimetría, el estadístico de contraste y nivel crítico para datos simulados de la distribución normal.

```
> library(moments)
> n <- 30
> x <- rnorm(n)  # Generamos datos normales
> A <- skewness(x)  # Coeficiente de asimetria
> estad <- sqrt(n/6) * A
> estad  # Estadístico del contraste

[1] 0.6344

> pval <- 2 * (1 - pnorm(abs(estad)))  # p-valor
> pval

[1] 0.5259
```

Test basado en la asimetría

- Como ejemplo vamos a presentar los resultados del contraste de normalidad basado en asimetría aplicado a datos simulados de distintas distribuciones.
- Mostramos ahora los valores del coeficiente de asimetría, el estadístico de contraste y nivel crítico para datos simulados de la distribución exponencial.

```
> library(moments)
> n <- 30
> x <- rexp(n)  # Generamos datos esponenciales
> A <- skewness(x)  # Coeficiente de asimetria
> estad <- sqrt(n/6) * A
> estad  # Estadístico del contraste

[1] 1.99

> pval <- 2 * (1 - pnorm(abs(estad)))  # p-valor
> pval

[1] 0.04661
```

Test basado en la kurtosis

- Partimos de la muestra $x_1, \ldots, x_n$. El momento de orden cuatro de los valores estandarizados se calcula como:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4,$$

siendo $\overline{x}$ la media muestral y s^2 la varianza muestral.

- Las características de posición se calculan mediante un momento de orden uno (la media), las de dispersión a través de un momento de orden dos (la varianza) y la asimetría con un momento de orden tres (el coeficiente de asimetría). El momento de orden cuatro permite medir características de forma, que son un refinamiento todavía más detallado en el análisis de la distribución.

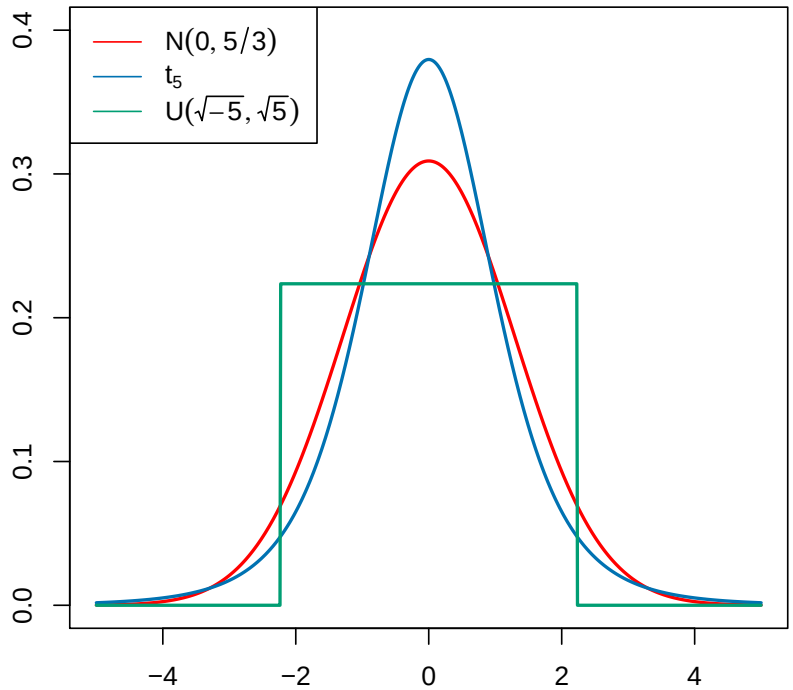
Test basado en la kurtosis

- Si bien en algunos textos la kurtosis se define como el momento de orden 4, es más habitual definir la kurtosis como:

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

  En esta definición se resta tres para que el coeficiente valga cero en la distribución normal. Así definida, K también se conoce como *exceso de kurtosis*.

- Si una distribución tiene un apuntamiento similar al de la normal decimos que es **mesocúrtica**, si el apuntamiento es mayor se dice **leptocúrtica**, y si es menor se dice **platicúrtica**.



Test basado en la kurtosis

- Si estamos ante una distribución simétrica, podemos valorar una posible desviación del modelo normal en función de su apuntamiento o kurtosis.
- Rechazaremos la normalidad si la kurtosis muestral es mucho mayor o mucho menor que cero, que es lo que corresponde a la distribución normal.
- Se puede demostrar que si la distribución de los datos muestrales es normal, entonces la kurtosis tiene distribución asintótica normal de media cero y varianza $24/n$, por lo que se puede emplear como estadístico de contraste el siguiente:

$$\sqrt{\frac{n}{24}}\, K$$

- Se rechazará la normalidad, en base a la kurtosis, cuando el estadístico anterior sea muy grande (en positivo o negativo), en comparación con los cuantiles de la $N(0,1)$.

Test basado en la kurtosis

- Como ejemplo vamos a presentar los resultados del contraste de normalidad basado en kurtosis aplicado a datos simulados de distintas distribuciones.
- Mostramos en primer lugar los valores del coeficiente de kurtosis, el estadístico de contraste y nivel crítico para datos simulados de la distribución normal.

```
> library(moments)
> n <- 100
> x <- rnorm(n)  # Generamos datos normales
> K <- kurtosis(x) - 3  # Exceso de kurtosis
> estad <- sqrt(n/24) * K  # Estadístico del contraste
> estad

[1] -0.287

> pval <- 2 * (1 - pnorm(abs(estad)))  # p-valor
> pval

[1] 0.7741
```

Test basado en la kurtosis

- Como ejemplo vamos a presentar los resultados del contraste de normalidad basado en kurtosis aplicado a datos simulados de distintas distribuciones.
- Mostramos ahora los valores del coeficiente de kurtosis, el estadístico de contraste y nivel crítico para datos simulados de la distribución t-de Student con 4 g.l.

```
> library(moments)
> n <- 100
> x <- rt(n, 4)  # Generamos datos de una t4
> K <- kurtosis(x) - 3  # Exceso de kurtosis
> estad <- sqrt(n/24) * K  # Estadístico del contraste
> estad

[1] 30.03

> pval <- 2 * (1 - pnorm(abs(estad)))  # p-valor
> pval

[1] 0
```

Test de Jarque-Bera

- También se puede elaborar un test conjunto para las dos características sin más que estandarizar cada una de las distribuciones y sumarlas.
- En este caso, se obtiene la siguiente expresión:

$$\frac{nA^2}{6} + \frac{nK^2}{24}$$

- Teniendo en cuenta que bajo la hipótesis nula de normalidad el estadístico resulta la suma de dos N(0,1) independientes, se tiene que sigue una distribución χ^2 con dos grados de libertad.
- Este test se conoce habitualmente con el nombre de Jarque-Bera.

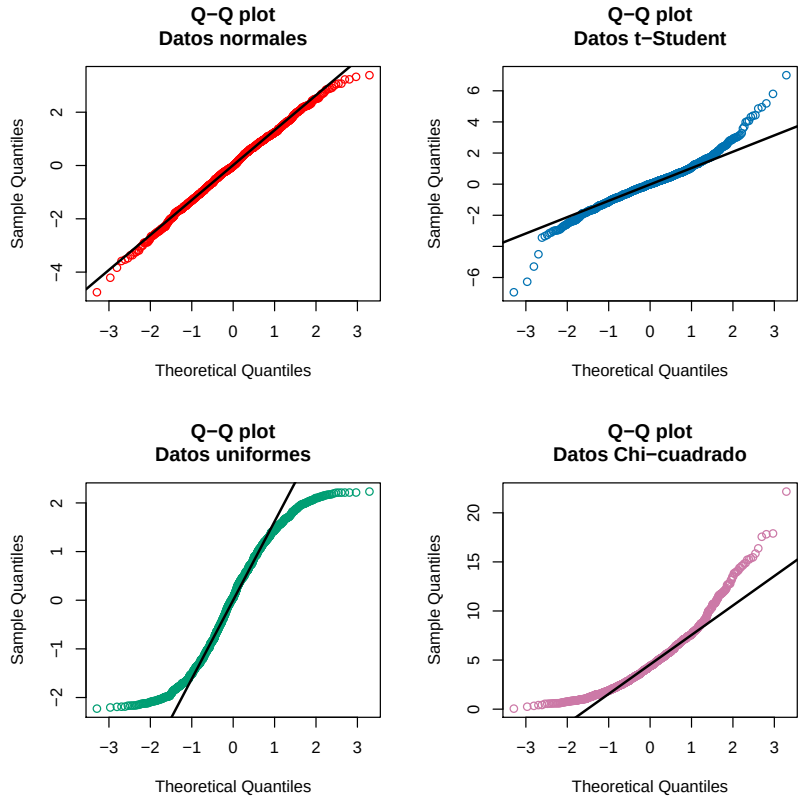
Test de Jarque-Bera

- Como ejemplo vamos a presentar los resultados del contraste de Jarque-Bera a datos simulados. Mostramos los valores del estadístico de contraste y nivel crítico para datos simulados de la distribución normal.

```
> library(moments)
> n <- 100
> x <- rnorm(n)  # Generamos datos normales
> A <- skewness(x)
> estadA <- sqrt(n/6) * A  # Estadístico de asimetría
> K <- kurtosis(x) - 3
> estadK <- sqrt(n/24) * K  # Estadístico de kurtosis
> estad <- estadA^2 + estadK^2  # Estadístico de Jarque-Bera
> estad

[1] 2.345

> pval <- 1 - pchisq(estad, 2)  # p-valor
> pval

[1] 0.3096
```

Gráfico de probabilidad normal

- Un procedimiento gráfico sencillo para evaluar la normalidad de un conjunto de observaciones de una variable es el gráfico de probabilidad normal[6], que compara los cuantiles de la muestra con los cuantiles de la distribución normal univariante.



[6] El gráfico de probabilidad normal es un caso particular de Q-Q plot

Test de Shapiro-Wilk

- El test de Shapiro -Wilk plantea como hipótesis nula que la muestra $x_1, \ldots, x_n$ proviene de una población normal.
- Este contraste mide el ajuste a una recta de la muestra representada en un gráfico de probabilidad normal. Se rechaza normalidad cuando el ajuste es malo que se corresponde a valores pequeños del estadístico.
- Este procedimiento es más general, pues permite detectar tanto defectos de simetría como de kurtosis, o incluso de otro tipo no considerado por los tests anteriores.

Test de Shapiro-Wilk

- Como ejemplo vamos a presentar los resultados del contraste de normalidad de Shapiro-Wilk aplicado a datos simulados de distintas distribuciones.

```
> shapiro.test(rnorm(100, mean = 5, sd = 3))


Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.9888, p-value = 0.5699

> shapiro.test(runif(100, min = 2, max = 4))


Shapiro-Wilk normality test

data:  runif(100, min = 2, max = 4)
W = 0.96, p-value = 0.004027
```

Transformaciones para obtener normalidad

- Como hemos visto, cuando hay dudas sobre la suposición de normalidad, podemos efectuar un contraste de bondad de ajuste. Si, como resultado del contraste, se acepta la hipótesis de normalidad podemos continuar con los métodos de inferencia clásicos, pero si se rechaza la hipótesis de normalidad, no podemos continuar con métodos que supongan tal hipótesis, y nos quedan las siguientes opciones:
  - Efectuar una transformación de los datos de manera que los datos transformados ya se adapten al modelo normal.
  - Buscar otro tipo de modelo de distribución que se ajuste mejor a los datos, contrastar este ajuste mediante los tests que hemos visto, y en caso de aceptar el modelo, aplicar métodos de inferencia específicos para el nuevo modelo.
  - Si no encontramos un modelo de distribución que se ajuste a los datos, habría que aplicar métodos de inferencia no paramétrica.

Transformaciones para obtener normalidad

- Existe una familia de transformaciones, conocida como **familia de Box-Cox**, que ha alcanzado mucha difusión como forma de acercar los datos al modelo normal.
- La causa más común de discrepancia respecto del modelo normal es la asimetría, la cual es muy habitual en variables que sean necesariamente positivas, como aquéllas que miden los tiempos de vida, o de supervivencia en ciertas circunstancias.
- Para casos donde el modelo normal no es aplicable por falta de simetría, están especialmente indicadas las transformaciones de Box-Cox.

Transformaciones para obtener normalidad

- La familia de transformaciones de Box-Cox viene dada por la siguiente expresión:

$$\forall y \in (0, +\infty) \qquad t_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y) & \text{si } \lambda = 0 \end{cases}$$

- Cada valor de λ produce una transformación diferente, de modo que escogiendo el valor de λ estamos escogiendo la transformación que más nos convenga en cada caso.
- Cualquiera de estas transformaciones está definida sobre el intervalo $(0, +\infty)$, y por tanto se aplica sobre una variable positiva.
- Para $\lambda > 1$ la transformación contrae los valores menores que uno y expande los valores mayores que uno según la potencia λ indicada.
- Para $\lambda \in (0, 1)$ produce el efecto contrario efectuando la raíz correspondiente.
- Con $\lambda = 0$ se lleva a cabo un logaritmo, que es una contracción muy severa de los valores grandes de la variable, mientras que los valores próximos a cero se expanden hacia $-\infty$.
- Por último, con $\lambda < 0$ se invierte la variable.

Ejercicios

- Se está estudiando el tiempo de ejecución de un algoritmo. Para ello, se han registrado los tiempos en 20 ejecuciones, con los resultados (en segundos) que figuran a continuación

  0.568, 0.893, 0.588, 1.805, 3.971, 2.302, 1.802, 0.778, 1.156, 0.470,
  0.465, 0.750, 0.715, 1.807, 1.342, 1.349, 1.299, 3.663, 0.955, 0.885

  Efectúa un test de Lilliefors para contrastar la normalidad de estos datos. Realiza, si es necesario, alguna transformación sobre los datos para conseguir normalidad.

- En las especificaciones técnicas de un programa para monotorizar la temperatura de la CPU se establece que los errores de medida siguen una distribución normal. Queremos contrastar esta hipótesis y para ello efectuamos treinta mediciones. Los datos se encuentran en el fichero `termometro.txt`. Resuelve el problema planteado utilizando los distintos procedimientos de contraste que hemos visto.

Introducción

- Veremos ahora algunos procedimientos de inferencia sencillos, en un contexto no paramétrico, es decir, sin suponer un modelo de distribución para las variables del problema.
- Veremos en primer lugar que es posible resolver un problema de inferencia sobre la posición central de una variable aleatoria sin suponer normalidad. Para ello, basaremos el procedimiento de inferencia en la mediana.
- Después veremos como comparar dos variables aleatorias tanto en el caso en que sean observadas sobre individuos independientes (muestras independientes) como en el caso en que sean observadas simultáneamente sobre los mismos individuos (muestras apareadas).
- Por último, haremos inferencia sobre la correlación entre dos variables aleatorias.

Contenidos: Contrastes de posición

Índice del curso

Test de los signos

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la mediana M difiere de un valor determinado M_0?

- Sea X una variable aleatoria continua.
- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución.
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste bilateral (hipótesis nula simple)

$$H_0 : M = M_0$$
$$H_1 : M \neq M_0$$

- Si la mediana fuera M_0, cada observación tendría una probabilidad $1/2$ de ser mayor que M_0, y por tanto cabría esperar que en torno a la mitad de las observaciones fueran mayores que M_0, y la otra mitad menores.

Test de los signos

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes?
- Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste bilateral (hipótesis nula simple)

$$H_0 : M = 29.3$$
$$H_1 : M \neq 29.3$$

Test de los signos

Contraste bilateral (hipótesis nula simple)

$$H_0 : M = M_0$$
$$H_1 : M \neq M_0$$

- Calcularemos el estadístico:

$$T = \text{ Número de observaciones muestrales mayores que } M_0$$

- Rechazaremos $H_0 : M = M_0$ si T es muy distinto de $n/2$.
- La distribución bajo la hipótesis nula de T es una es Binomial$(n, 1/2)$.
- Por tanto, podemos resolver este problema como el del contraste de una proporción ($H_0 : p = 1/2$, siendo $p = P(X > M_0)$).

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0, 1)$

Test de los signos

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes? Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x <- scan("ficheros.txt")
> n <- length(x)  # Tamaño muestral
> M0 <- 29.3  # Hipótesis nula
> T <- sum(x > M0)  # Nº de observaciones muestrales mayores que M0
> est <- (T/n - 0.5)/sqrt(0.5 * 0.5/n)  # Estadístico de contraste
> 2 * (1 - pnorm(abs(est)))  # p-valor

[1] 0.01963
```

- Por lo tanto, para una significación $\alpha = 0.05$, rechazamos la hipótesis nula.

Test de los signos

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes? Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Podríamos haber efectuado el test exacto:

```
> binom.test(T, n)


Exact binomial test

data:  T and n
number of successes = 25, number of trials = 36,
p-value = 0.02882
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5189 0.8365
sample estimates:
probability of success
                0.6944
```

Test de los signos

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la mediana M es mayor que un valor determinado M_0?

- Sea X una variable aleatoria continua.
- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución.
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste unilateral

$$H_0 : M \leq M_0$$
$$H_1 : M > M_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la mediana poblacional es menor o igual que M_0 si el número de observaciones por encima de M_0 es significativamente superior a $n/2$.

Rechazamos la hipótesis nula $H_0 : M \leq M_0$ frente a $H_1 : M > M_0$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Test de los signos

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la mediana M es menor que un valor determinado M_0?

- Sea X una variable aleatoria continua.
- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución.
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste unilateral

$$H_0 : M \geq M_0$$
$$H_1 : M < M_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la mediana poblacional es mayor o igual que M_0 si el número de observaciones por encima de M_0 es significativamente inferior a $n/2$.

Rechazamos la hipótesis nula $H_0 : M \geq M_0$ frente a $H_1 : M < M_0$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \leq -z_\alpha$$

z_α denota el punto tal que $P(Z > z_\alpha) = \alpha$ siendo Z una variable $N(0,1)$

Test de los rangos signados de Wilcoxon

Hipótesis: ¿Se puede concluir que una muestra de n individuos proviene de una población en la que la mediana M difiere de un valor determinado M_0?

- Sea X una variable aleatoria continua y simétrica.
- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución.
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste bilateral (hipótesis nula simple)

$$H_0 : M = M_0$$
$$H_1 : M \neq M_0$$

- Se define la variable $D = X - M_0$ que bajo la hipótesis nula H_0 es simétrica respecto al cero. Por tanto, si H_0 es cierta, debería tomar valores más o menos equidistantes respecto a cero.

Test de los signos

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes?
- Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`
- Se supone que la distribución es simétrica.
- Se desea contrastar una hipótesis relativa a la mediana M.

Contraste bilateral (hipótesis nula simple)

$$H_0 : M = 29.3$$
$$H_1 : M \neq 29.3$$

Test de los rangos signados de Wilcoxon

- El test de los rangos signados de Wilcoxon se basa en las diferencias:

$$|X_1 - M_0|, |X_2 - M_0|, \ldots, |X_n - M_0|$$

- Ordenamos las diferencias en orden creciente y asignamos a cada X_i su rango. Por ejemplo, si $n = 4$ y se tiene:

$$|X_2 - M_0| < |X_3 - M_0| < |X_1 - M_0| < |X_4 - M_0|$$

entonces, los rangos de X_i son:

$$\text{rg}(X_1) = 3; \text{rg}(X_2) = 1; \text{rg}(X_3) = 2; \text{rg}(X_4) = 4$$

- Definimos:

$$T^+ = \text{Suma de los rangos de los } X_i \text{ mayores que } M_0$$
$$T^- = \text{Suma de los rangos de los } X_i \text{ menores que } M_0$$

- Observamos que

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

- Además, si $H_0 : M = M_0$ es cierta, se debería tener $T^+ = \frac{n(n+1)}{4}$.
- El estadístico T^+ se denomina **estadístico de Wilcoxon**.

Test de los rangos signados de Wilcoxon

T^+ = Suma de los rangos de los X_i mayores que M_0

- Para muestras de tamaño grande el comportamiento de T^+ bajo H_0 es aproximadamente normal.

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Los valores iguales a cero se ignoran y si varias diferencias D_i son iguales se les asigna el rango promedio de los valores empatados.

Test de los rangos signados de Wilcoxon

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes? Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x <- scan("ficheros.txt")
> M0 <- 29.3 # Hipótesis nula
> D <- x - M0 # Diferencias X - M0
> if (any(D == 0)){D <- D[-which(D == 0)]} # Elimino Di = 0
> n <- length(D) # Tamaño muestral
> rangos <- rank(abs(D), ties.method = c("average"))
> T <- rangos * sign(D)
> Tmas <- sum(T[T>0]) # Estadístico T+
> Tmas

[1] 550.5

> est <- (Tmas - (n * (n + 1)/4))/sqrt(n * (n + 1) * (2 * n + 1)/24)
> 2 * (1 - pnorm(abs(est)))

[1] 0.000633
```

Test de los rangos signados de Wilcoxon

- Ejemplo: ¿Difiere el tamaño mediano de los ficheros de un sistema de archivos de 29.3 KBytes? Tomamos una muestra de 36 ficheros. Los datos se encuentran en `ficheros.txt`

Rechazamos la hipótesis nula $H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ si

$$\frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- También puedo utilizar la función `wilcox.test`

```
> library(MASS)
> x <- scan("ficheros.txt")
> M0 <- 29.3  # Hipótesis nula
> wilcox.test(x, mu = M0)

Warning:  cannot compute exact p-value with ties


Wilcoxon signed rank test with continuity correction

data:  x
V = 550.5, p-value = 0.0006513
alternative hypothesis: true location is not equal to 29.3
```

Tests para dos muestras

- Se trata ahora de verificar si dos poblaciones podrían ser homogéneas en cuanto a su posición.
- Básicamente estamos en la misma situación que el contraste de igualdad de medias vista en el apartado de inferencia paramétrica salvo que ahora la posición no se medirá con la media de la población sino con otras medidas.
- Como en aquel caso diferenciaremos los casos de muestras independientes y muestras apareadas.

Test de Wilcoxon-Mann-Whitney

Hipótesis: ¿Se puede concluir que dos muestras independientes proceden de dos poblaciones con medianas distintas?

- Disponemos de dos muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución.
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2).

Contraste bilateral (hipótesis nula simple)

$$H_0 : M_1 = M_2$$
$$H_1 : M_1 \neq M_2$$

- Observamos que, si es cierta H_0, entonces la distribución de $X_1 - X_2$ será simétrica respecto al origen y por tanto $P(X_1 > X_2) = P(X_2 > X_1) = 0.5$
- Una forma de comparar los resultados de X_1 con X_2 consistiría en hacer un recuento del tipo "Número de veces que X_2 es mayor que X_1".

Test de Wilcoxon-Mann-Whitney

- Ejemplo: ¿El número mediano de usuarios concurrentes a una aplicación A difiere del número mediano de usuarios concurrentes a otra aplicación B?
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : M_1 = M_2$$
$$H_1 : M_1 \neq M_2$$

siendo M_1 el número mediano de usuarios concurrentes a A y M_2 el número mediano de usuarios concurrentes a B.

Test de Wilcoxon-Mann-Whitney

- El estadístico de Mann-Whitney permite efectuar la comparación entre las dos muestras de la siguiente manera:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{I}_{\{X_{2j} > X_{1i}\}}$$

- Es decir, contamos todos los pares (X_{1i}, X_{2j}) en los que X_2 es mayor que X_1.
- Hay $n_1 n_2$ pares posibles, formados enfrentando cada valor de una variable con todos los valores de la otra.
- Teniendo en cuenta la definición del estadístico:
  - Consideraremos que las dos variables son semejantes si el estadístico de Mann- Whitney toma un valor moderado.
  - Pensaremos que X_2 tiende a ser mayor que X_1 para valores grandes del estadístico.
  - Pensaremos que X_2 es por lo general más pequeña que X_1 si el estadístico toma un valor pequeño

Test de Wilcoxon-Mann-Whitney

- El estadístico de Mann-Whitney permite efectuar la comparación entre las dos muestras de la siguiente manera:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{I}_{\{X_{2j} > X_{1i}\}}$$

- Como hemos comentado,la hipótesis nula que vamos a contrastar se formula como H_0 : $P(X_2 > X_1) = 1/2$.
- Tendremos presente que X_1 y X_2 son independientes.
- Si H_0 es cierta, el estadístico tiene distribución simétrica con

$$E(U) = \frac{n_1 n_2}{2}, \quad Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

- Su distribución exacta para muestras de tamaño pequeño se encuentra tabulada y para tamaños grandes se aproxima por una normal.

Rechazamos la hipótesis nula $H_0 : M_1 = M_2$ frente a $H_1 : M_1 \neq M_2$ si

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Test de Wilcoxon-Mann-Whitney

- El estadístico de Mann-Whitney se puede calcular fácilmente utilizando la siguiente formulación equivalente:
  - Se ordenan las dos muestras conjuntamente y se asignan rangos $1, 2, \ldots, n_1 + n_2$ de menor a mayor.
  - Se tiene:

$$U = W + \frac{n_2(n_2+1)}{2}$$

donde W es la suma de los rangos correspondientes de la segunda muestra.

Rechazamos la hipótesis nula $H_0 : M_1 = M_2$ frente a $H_1 : M_1 \neq M_2$ si

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Test de Wilcoxon-Mann-Whitney

- Ejemplo: ¿El número mediano de usuarios concurrentes a una aplicación A difiere del número mediano de usuarios concurrentes a otra aplicación B?
- Registramos en 15 ocasiones el nº de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el nº de usuarios concurrentes a la aplicación B.

Rechazamos la hipótesis nula $H_0 : M_1 = M_2$ frente a $H_1 : M_1 \neq M_2$ si

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x1 <- scan("UsuariosA.txt"); x2 <- scan("UsuariosB.txt")
> n1 <- length(x1); n2 <- length(x2)
> x <- c(x2,x1) # Junto las muestras
> rangos <- rank(x, ties.method = c("average"))
> r2 <- rangos[1:n2] # Rangos de la muestra 2
> U <- sum(r2) - n2 * (n2 + 1)/2
> U

[1] 128.5

> est <- (U - n1 * n2/2)/sqrt(n1 * n2 * (n1 + n2 + 1)/12)
> 2 * (1 - pnorm(abs(est))) # p-valor

[1] 0.4736
```

Test de Wilcoxon-Mann-Whitney

- Ejemplo: ¿El número mediano de usuarios concurrentes a una aplicación A difiere del número mediano de usuarios concurrentes a otra aplicación B?
- Registramos en 15 ocasiones el número de usuarios concurrentes a la aplicación A y, de manera independiente, registramos en 20 ocasiones el número de usuarios concurrentes a la aplicación B.

Rechazamos la hipótesis nula $H_0 : M_1 = M_2$ frente a $H_1 : M_1 \neq M_2$ si

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x1 <- scan("UsuariosA.txt")
> x2 <- scan("UsuariosB.txt")
> wilcox.test(x1, x2)

Warning:  cannot compute exact p-value with ties


Wilcoxon rank sum test with continuity correction

data:  x1 and x2
W = 171.5, p-value = 0.4835
alternative hypothesis: true location shift is not equal to 0
```

Test de los signos para muestras apareadas

Hipótesis: ¿Se puede concluir que dos muestras dependientes proceden de dos poblaciones distintas?

- En ocasiones nos interesará comparar dos métodos o tratamientos.
- En ese caso es natural que los individuos donde se aplican los tratamientos sean los mismos. Es decir, a un mismo individuo se le efectúa la medición de dos variables aleatorias, X_1 y X_2.
- Para tomar la decisión nos basamos en las observaciones de n individuos independientes, y por tanto podemos formalizarlo mediante una muestra aleatoria simple $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$.
- Si las muestras provienen de la misma población (si no hay efecto del tratamiento), la distribución de la variable $X_2 - X_1$ será simétrica respecto al origen y por tanto la $P(X_2 > X_1) = P(X_1 > X_2) = 0.5$.

Contraste bilateral (hipótesis nula simple)

$$H_0 : P(X_2 > X_1) = 1/2$$
$$H_1 : P(X_2 > X_1) \neq 1/2$$

Test de los signos para muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo de ejecución de nuestro código original difiere del tiempo de ejecución tras modificar la función?
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado. Queremos contrastar:

Contraste bilateral (hipótesis nula simple)

$$H_0 : H_0 : P(X_2 > X_1) = 1/2$$
$$H_1 : H_0 : P(X_2 > X_1) \neq 1/2$$

siendo X_1 el tiempo antes de la modificación y X_2 el tiempo después de la modificación.

Test de los signos para muestras apareadas

- Una forma de comparar los resultados de la variable X_1 con los de la variable X_2 consiste en hacer un recuento del tipo

$$T = \text{Número de veces que } X_2 \text{ es mayor que } X_1 = \sum_{i=1}^{n} \mathbb{I}_{\{X_{2i} > X_{1i}\}}$$

- Si H_0 es cierta, cabe esperar que T valga en torno a $n/2$. Es decir, rechazaremos H_0 si T es muy distinto de $n/2$.
- La distribución bajo la hipótesis nula de T es una es Binomial$(n, 1/2)$. Para tamaños muestrales grandes, el test se puede aproximar con la distribución normal.

Rechazamos la hipótesis nula $H_0 : P(X_2 > X_1) = 1/2$ si

$$\frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\frac{T}{n} - \frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0, 1)$

Test de los signos para muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo de ejecución de nuestro código original difiere del tiempo de ejecución tras modificar la función?
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado.

```
> x1 <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> x2 <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> n <- length(x1)
> D <- x2 - x1 # Diferencias X2 - X1
> T <- sum(D > 0) # Nº de observaciones muestrales mayores que M0
> est <- (T/n - 0.5)/sqrt(0.5 * 0.5/n)
> 2*(1-pnorm(abs(est))) # p-valor

[1] 0.001565
```

- Por lo tanto, rechazamos H_0

Test de los signos para muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo de ejecución de nuestro código original difiere del tiempo de ejecución tras modificar la función?
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado.
- Con la función `binom.test` realizamos el test exacto:

```
> x1 <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> x2 <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> n <- length(x1)
> D <- x2 - x1 # Diferencias X2 - X1
> T <- sum(D > 0) # Nº de observaciones muestrales mayores que M0
> binom.test(T, n)


Exact binomial test

data:  T and n
number of successes = 0, number of trials = 10,
p-value = 0.001953
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.0000 0.3085
sample estimates:
probability of success
                     0
```

Test de los rangos signados de Wilcoxon para muestras apareadas

- También se puede adaptar el test de los rangos signados de Wilcoxon al caso de muestras apareadas.
- Tomaremos las diferencias $D_i = X_{2i} - X_{1i}$ y las ordenamos de menor a mayor prescindiendo del signo.
- Les asignamos los rangos correspondientes $1, 2, \ldots, n$.
- Definimos:

$$T^+ = \text{Suma de los rangos de las } D_i \text{ mayores que } 0$$
$$T^- = \text{Suma de los rangos de las } D_i \text{ menores que } 0$$

$$T = \min(T^+, T^-)$$

- Para muestras de tamaño grande el comportamiento de T bajo H_0 es aproximadamente normal.

Rechazamos la hipótesis nula $H_0 : P(X_2 > X_1) = 1/2$ si

$$\frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Test de los rangos signados de Wilcoxon para muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo de ejecución de nuestro código original difiere del tiempo de ejecución tras modificar la función?
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado.

```
> x1 <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> x2 <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> n <- length(x1)
> D <- x2 - x1 # Diferencias x2 - x1
> if (any(D == 0)){D <- D[-which(D == 0)]} # Elimino Di = 0
> n <- length(D) # Tamaño muestral
> rangos <- rank(abs(D), ties.method = c("average"))
> T <- rangos * sign(D)
> Tmas <- sum(T[T>0]) # Estadístico T+
> Tmenos <- sum(T[T<0]) # Estadístico T-
> T <- min(Tmas, Tmenos)
> T

[1] -55

> est <- (T - (n * (n + 1)/4))/sqrt(n * (n + 1) * (2 * n + 1)/24)
> 2 * (1 - pnorm(abs(est)))

[1] 0
```

- Por lo tanto, rechazamos H_0

Test de los rangos signados de Wilcoxon para muestras apareadas

- Ejemplo: Estamos interesados en analizar el tiempo de ejecución de un determinado programa que hemos escrito con R. Tras analizar nuestro código, hemos decidido reescribir una parte y sustituir una función por otra equivalente pero que ha sido programada por un experto. ¿El tiempo de ejecución de nuestro código original difiere del tiempo de ejecución tras modificar la función?
- Registramos en 10 conjuntos de datos el tiempo de ejecución del código original y, para los mismos datos, registramos el tiempo de ejecución del código modificado.
- Utilizando la función `wilcox.test`

```
> x1 <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96, 18.15)
> x2 <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59, 16.57)
> n <- length(x1)
> wilcox.test(x1, x2, paired=TRUE)


Wilcoxon signed rank test

data:  x1 and x2
V = 55, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

Contenidos: Contrastes de asociación

Índice del curso

Contrastes de asociación

Hipótesis: ¿Es razonable admitir en base a la observación de 2 características en n individuos que dichas características están relacionadas?

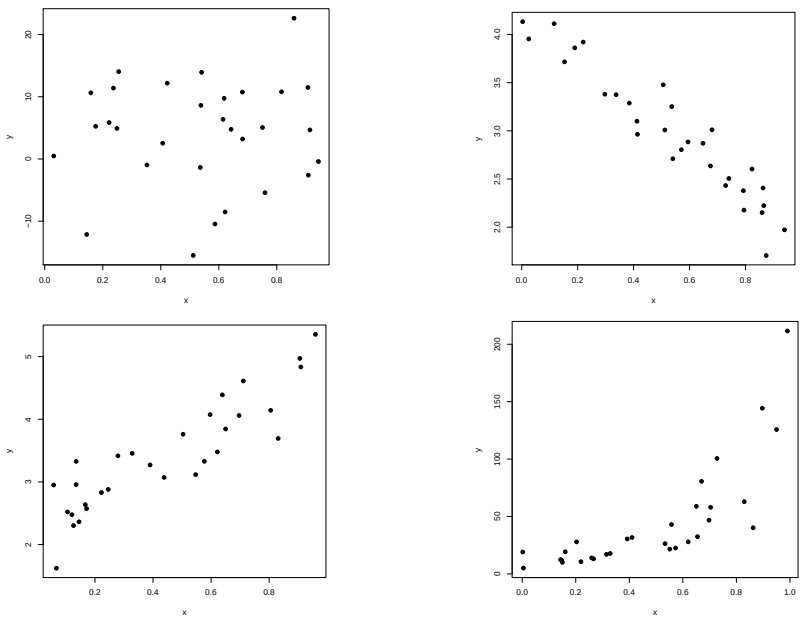
- Supongamos que de n elementos de una población se han observado dos características X e Y, obteniéndose una muestra aleatoria simple bidimensional $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
- Sobre la base de estas observaciones se desea contrastar si las características poblacionales X e Y están o no relacionadas.

Contraste de hipótesis

$H_0 : X$ e Y no están relacionadas
$H_1 : X$ e Y están relacionadas

Contrastes de asociación

- La representación gráfica más útil de dos variables continuas es el **diagrama de dispersión**.
- Consiste en representar en un eje de coordenadas los pares de observaciones (X_i, Y_i).
- La nube así dibujada refleja la posible relación entre las variables.
- A mayor relación entre las variables más estrecha y alargada será la nube.



Contrastes de asociación

- La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional.
- Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas.

Covarianza entre X e Y

$$\mathrm{Cov}(X, Y) = s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.
- La covarianza de (X, Y) es igual a la de (Y, X), es decir, $s_{xy} = s_{yx}$.
- La covarianza de (X, X) es igual a la varianza de X, es decir $s_{xx} = s_x^2$

Contrastes de asociación

- La covarianza cambia si modificamos las unidades de medida de las variables.
- Esto es un inconveniente porque no nos permite comparar la relación entre distintos pares de variables medidas en diferentes unidades.
- La solución es utilizar el **coeficiente de correlación lineal**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.
- Si toma valores cercanos a $+1$ diremos que hay una relación directa entre X e Y.
- Si toma valores cercanos a cero diremos que no existe relación lineal entre X e Y.

Coeficiente de correlación de Pearson

- Un coeficiente de correlación es una medida de si dos variables aleatorias presentan una relación creciente o decreciente, o lo que es lo mismo, indica si al aumentar una variable cabe esperar un aumento o una disminución de la otra.
- El coeficiente de correlación más conocido es el de Pearson, y es la mejor medida de correlación cuando las variables aleatorias siguen una distribución normal.
- Además, bajo esta suposición de normalidad, se puede construir un estadístico, con distribución conocida, para hacer inferencia sobre el coeficiente de correlación.

Coeficiente de correlación de Pearson

Hipótesis: ¿Es razonable admitir en base a la observación de 2 características en n individuos que dichas características están relacionadas?

- Supongamos que de n elementos de una población se han observado dos características X e Y, obteniéndose una muestra aleatoria simple bidimensional $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
- Suponemos que las variables X e Y son normales.
- Sobre la base de estas observaciones se desea contrastar si las características poblacionales X e Y están o no relacionadas.

Contraste de hipótesis

H_0 : X e Y no están relacionadas
H_1 : X e Y están relacionadas

- Entonces, si H_0 es cierta,

$$\sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \sim t_{n-2}$$

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \leq -t_{\alpha/2} \quad \text{ó} \quad \sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

Coeficiente de correlación de Pearson

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \leq -t_{\alpha/2} \quad \text{ó} \quad \sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

- Como ejemplo, generamos en primer lugar muestras independientes.

```
> set.seed(12345)
> n <- 50
> x <- rnorm(n)
> y <- rnorm(n)
> rxy <- cor(x, y)  # Coeficiente de correlación lineal
> rxy

[1] -0.001707

> est <- sqrt(rxy^2/(1 - rxy^2)) * sqrt(n - 2)  # Estadístico
> est

[1] 0.01183

> 2 * (1 - pt(abs(est), n - 2))  # p-valor

[1] 0.9906
```

- Observamos que no rechazamos la hipótesis nula.

Coeficiente de correlación de Pearson

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \leq -t_{\alpha/2} \quad \text{ó} \quad \sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

- Usando la función `cor.test`.

```
> set.seed(12345)
> n <- 50
> x <- rnorm(n)
> y <- rnorm(n)
> cor.test(x, y, method = "pearson")


Pearson's product-moment correlation

data:  x and y
t = -0.0118, df = 48, p-value = 0.9906
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2799  0.2768
sample estimates:
      cor
-0.001707
```

- Obtenemos el mismo resultado.

Coeficiente de correlación de Pearson

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \leq -t_{\alpha/2} \quad \text{ó} \quad \sqrt{\frac{r_{xy}^2}{1-r_{xy}^2}}\sqrt{n-2} \geq t_{\alpha/2}$$

$t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ g.l.

- Generamos ahora datos relacionados.

```
> set.seed(12345)
> n <- 50
> x <- rnorm(n)
> y <- 2 + 3 * x + rnorm(n)
> cor.test(x, y, method = "pearson")


Pearson's product-moment correlation

data:  x and y
t = 19.98, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9043 0.9685
sample estimates:
   cor
0.9448
```

Coeficiente de correlación por rangos de Spearman

- Si el modelo de distribución no es normal, podemos hacer un estudio de la correlación mediante coeficientes construidos sobre relaciones de orden entre los datos.
- Nos centraremos en el **coeficiente de correlación por rangos de Spearman**.
- Suponemos disponible una muestra bidimensional $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
- Queremos calcular un coeficiente que cuantifique la relación creciente o decreciente que pueda existir entre los valores X y los valores Y.
- Ordenamos los valores X de menor a mayor. Hacemos lo mismo con los valores Y.
- Denotamos:
  - $a1 = rango(X_1)$, como el rango o posición que ocupa el dato X_1 en la muestra ordenada de las X.
  - De igual modo, se denota $b1 = rango(Y_1)$ como el rango o posición que ocupa el dato Y_1 en la muestra ordenada de las Y.
- Haciendo lo mismo con cada par de datos de la muestra obtenemos

$$(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$$

- El coeficiente de correlación por rangos de Spearman no es más que el resultado de calcular el coeficiente de correlación de Pearson sobre los rangos.

Coeficiente de correlación por rangos de Spearman

Coeficiente de correlación de Spearman

$$R_S = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}}.$$

- Como coeficiente de correlación que es, este coeficiente tomará un valor próximo a cero cuando las variables X e Y sean incorrelacionadas, tomará un valor positivo cuando haya una relación creciente entre las variables, y será negativo cuando la relación sea decreciente. Además, los valores posibles para este coeficiente están entre -1 y 1.
- El coeficiente de Spearman también se puede usar para contrastar la hipótesis de incorrelación entre las variables X e Y.
- Rechazaremos la hipótesis nula de incorrelación cuando el coeficiente de Spearman sea "muy distinto" de cero.
- Para muestras grandes, usamos siguiente aproximación mediante la distribución normal:

$$\sqrt{n-1}R_S \sim N(0,1)$$

Coeficiente de correlación por rangos de Spearman

Contraste de hipótesis

H_0 : X e Y no están relacionadas
H_1 : X e Y están relacionadas

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas **si**

$$\sqrt{n-1}R_S \leq -z_{\alpha/2} \quad \text{ó} \quad \sqrt{n-1}R_S \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Coeficiente de correlación por rangos de Spearman

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas **si**

$$\sqrt{n-1}R_S \leq -z_{\alpha/2} \quad \text{ó} \quad \sqrt{n-1}R_S \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Como ejemplo, generamos en primer lugar muestras independientes.

```
> set.seed(12345)
> n <- 50
> x <- runif(n)
> y <- runif(n)
> ai <- rank(x)  # Rangos de muestra 1
> bi <- rank(y)  # Rangos de muestra 2
> Rs <- cor(ai, bi)  # Coeficiente de correlación lineal entre rangos [Spearman]
> Rs

[1] 0.07006

> est <- sqrt(n - 1) * Rs  # Estadístico
> 2 * (1 - pnorm(abs(est)))  # p-valor

[1] 0.6238
```

- Observamos que no rechazamos la hipótesis nula.

Coeficiente de correlación por rangos de Spearman

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{n-1}R_S \leq -z_{\alpha/2} \quad \text{ó} \quad \sqrt{n-1}R_S \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Usando la función `cor.test`.

```
> set.seed(12345)
> n <- 50
> x <- runif(n)
> y <- runif(n)
> cor.test(x, y, method = "spearman")


Spearman's rank correlation rho

data:  x and y
S = 19366, p-value = 0.6278
alternative hypothesis: true rho is not equal to 0
sample estimates:
    rho
0.07006
```

- Obtenemos el mismo resultado.

Coeficiente de correlación por rangos de Spearman

Rechazamos la hipótesis nula H_0 : X e Y no están relacionadas si

$$\sqrt{n-1}R_S \leq -z_{\alpha/2} \quad \text{ó} \quad \sqrt{n-1}R_S \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Generamos ahora datos relacionados.

```
> set.seed(12345)
> n <- 50
> x <- runif(n)
> y <- 2 + 3 * x + rnorm(n)
> cor.test(x, y, method = "spearman")


Spearman's rank correlation rho

data:  x and y
S = 8270, p-value = 5.77e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
   rho
0.6029
```

Contenidos: Contrastes de aleatoriedad

Índice del curso

Introducción

- Cuando las observaciones son dependientes, las expresiones obtenidas para los distintos estimadores cambian radicalmente.
- Esto es debido principalmente a que introduce un sesgo en los estimadores de las varianzas. En particular, la propiedad que asegura que la varianza de la media muestral de n observaciones $N(\mu,\sigma^2)$ independientes es σ^2/n ya no se cumple.
- Es por esto que la hipótesis de independencia se convierte entonces en una de las hipótesis más importantes a contrastar ya que de ella suele depender la fiabilidad de las estimaciones posteriores.
- En general, esta hipótesis debe contrastarse cuando los datos mediante procedan de una secuencia temporal o espacial.

Contraste de rachas

Hipótesis: ¿Se puede concluir que en una muestra de n individuos las observaciones son dependientes?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n observaciones.
- Definimos **racha** como una secuencia de valores que cumplen una condición lógica. Por ejemplo, que estén por encima o por debajo de un valor o que formen una secuencia monótona (creciente o decreciente).
- El número total de rachas en una muestra proporciona un indicio de si hay o no aleatoriedad en la muestra.

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

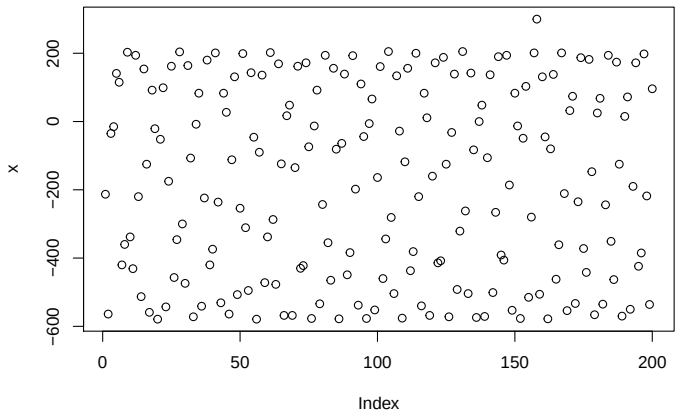
Contraste de rachas

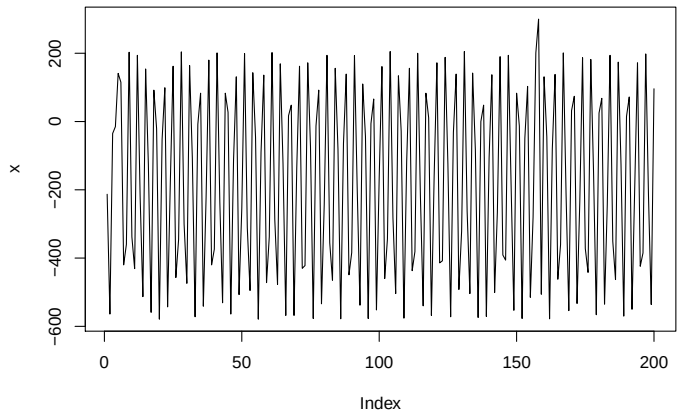
- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

```
> x <- scan("racha.txt")
> plot(x)  # Observaciones en función del índice
> plot(x, type = "l")
```





Contraste de rachas

- Supongamos que X toma dos valores $(+)$ y $(-)$.
- Observamos una muestra del tipo:

$$+ \; + \; + \; + \; - \; - \; - \; + \; + \; + \; - \; -$$

- Una racha es una secuencia de observaciones iguales. Por lo tanto, en este ejemplo tenemos 4 rachas.
- Para contar el número de rachas en la sucesión sólo debemos contar cuantas veces se cambia de valor y sumarle 1

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

- Una muestra con muchas o pocas rachas sugeriría que la muestra no es aleatoria

Contraste de rachas

- Consideramos la variable aleatoria:

$$R = \text{Número de rachas en la muestra.}$$

- Bajo la hipótesis nula de aleatoriedad:

$$R \sim N\left(1 + \frac{2n_1n_2}{n}, \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}\right)$$

siendo:

  - n_1 el número de signos $(+)$ en la muestra.
  - n_2 el número de signos $(-)$ en la muestra.
  - Por tanto $n = n_1 + n_2$.

- Para tamaños muéstrales pequeños, la aproximación anterior no es buena y conviene utilizar la distribución exacta (o utilizar corrección por continuidad).

Contraste de rachas

- Cuando aplicamos el test a variables continuas necesitamos dicotomizar los datos.
- Usualmente comparamos con la mediana muestral, ya que nos permite determinar cuanto valen n_1 y n_2 a partir de n.
- Definimos:

$$s_i = \begin{cases} 1 & \text{si } x_i < \text{mediana}, \\ 0 & \text{si } x_i > \text{mediana}. \end{cases}$$

- De nuevo, el número de rachas se cuenta como los cambios de valor de la secuencia s_i +1. Bajo independencia el número de rachas por encima o por debajo de la mediana sigue la siguiente distribución:

$$\text{n}^\text{o} \text{ rachas mediana} \sim N\left(n_1 + 1, \frac{n_1(n_1-1)}{2n_1-1}\right),$$

donde n_1 es el número de unos que tiene la secuencia (es igual al número de ceros).

Rechazamos la hipótesis nula H_0 : La muestra es aleatoria si

$$\frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Contraste de rachas

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula H_0 : La muestra es aleatoria si

$$\frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x <- scan("racha.txt")
> n <- length(x)  # Tamaño muestral
> si <- (x > median(x))  # Función si
> RM <- sum(diff(si) != 0) + 1  # Nº de rachas = Nº de cambios + 1
> n1 <- sum(x > median(x))
> est <- (RM - (n1 + 1))/sqrt(n1 * (n1 - 1)/(2 * n1 - 1))  # Estadístico
> est

[1] 2.694

> 2 * (1 - pnorm(abs(est)))  # p-valor

[1] 0.007065
```

Contraste de rachas

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula H_0 : La muestra es aleatoria si

$$\frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{R_M - (n_1+1)}{\sqrt{\frac{n_1(n_1-1)}{2n_1-1}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Podemos usar la función `runs.test` (paquete `tseries`)

```
> library(tseries)


Attaching package: 'tseries'
The following object is masked from 'package:chron':
    is.weekend

> x <- scan("racha.txt")
> n <- length(x)  # Tamaño muestral
> si <- (x > median(x))  # Función si
> runs.test(factor(si))


Runs Test
```

Contraste de autocorrelación

Hipótesis: ¿Se puede concluir que en una muestra de n individuos las observaciones son dependientes?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n observaciones.
- Se define el coeficiente de autocorrelación lineal de orden uno $r(1)$ como:

$$r(1) = \frac{\sum_{i=2}^{n} (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- Este coeficiente viene a ser aproximadamente el coeficiente de correlación lineal de las variables $(x_2, \ldots, x_n)$ y $(x_1, \ldots, x_{n-1})$ y mide la correlación lineal entre las observaciones y sus observaciones precedentes.
- Análogamente definimos el coeficiente de autocorrelación lineal de orden k como:

$$r(k) = \frac{\sum_{i=k+1}^{n} (x_i - \bar{x})(x_{i-k} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- Su interpretación es también análoga a la del coeficiente de orden uno.

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

- En caso de independencia es de esperar que las autocorrelaciones muestrales sean próximas a cero (valores grandes indicarían dependencia positiva o negativa según el signo).

Contraste de autocorrelación

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

```
> x <- scan("racha.txt")
> n <- length(x)
> cor(x[-1], x[-n])  # Coef. de autocorrelación de orden 1

[1] -0.3081

> cor(x[-c(1, 2)], x[-c(n, n - 1)])  # Coef. de autocorrelación de orden 2

[1] -0.749
```

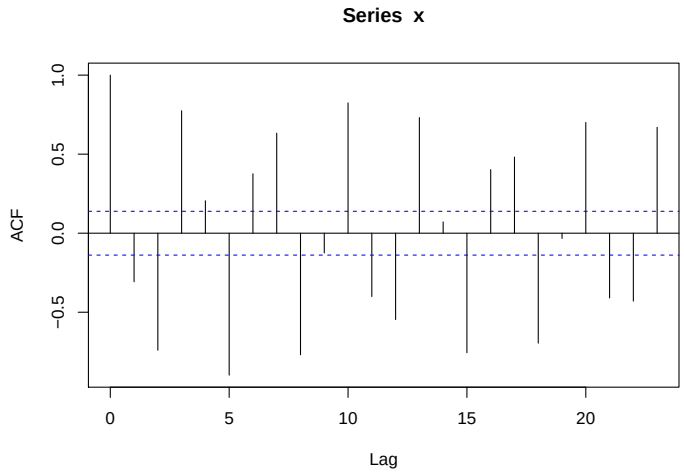
Contraste de autocorrelación

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Contraste

H_0 : La muestra es aleatoria
H_1 : La muestra no es aleatoria

```
> x <- scan("racha.txt")
> n <- length(x)
> acf(x)  # Coeficientes de autocorrelación [correlograma]
```



Contraste de autocorrelación

- Se denomina **correlograma** a la representación de los coeficientes de autocorrelación lineal en función de k (normalmente llamado *retardo*).
- Cuando las observaciones son independientes y proceden de una distribución normal, los coeficientes de autocorrelación muestrales siguen aproximadamente una distribución normal con media cero y varianza $1/n$.
- Por lo tanto, podemos considerar significativamente distintos de cero aquellos coeficientes que no estén en el intervalo

$$\left(-\frac{z_{\alpha/2}}{\sqrt{n}}, \frac{z_{\alpha/2}}{\sqrt{n}}\right).$$

Rechazamos la hipótesis nula $H_0 : \rho(k) = 0$ si

$$\frac{r(k)}{\sqrt{\frac{1}{n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{r(k)}{\sqrt{\frac{1}{n}}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Si n es grande ($n > 50$), podemos realizar un contraste conjunto de los primeros coeficientes de autocorrelación.

Contraste de autocorrelación

- Si H_0 es la hipótesis de independencia, entonces cada $r(k)$ se distribuye aproximadamente según $N(0, 1/\sqrt{n})$ y, por lo tanto,

$$Q = \sum_{k=1}^{m} \left(\frac{r(k)}{1/\sqrt{n}} \right)^2 = n \sum_{k=1}^{m} r(k)^2,$$

sigue, aproximadamente, una distribución χ^2 con $m-1$ grados de libertad (nótese que hay que estimar un parámetro: la media).

Contraste de Ljung y Box

- Este contraste ha sido mejorado por Ljung y Box que han demostrado que una aproximación más exacta es considerar el estadístico

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r(k)^2}{n-k}$$

que, como el anterior, si H_0 es cierta, se distribuye aproximadamente como una χ^2 con $m-1$ grados de libertad.

Rechazamos la hipótesis nula $H_0 : \rho(1) = \rho(2) = \ldots = \rho(m) = 0$ si

$$n(n+2)\sum_{k=1}^{m}\frac{r(k)^2}{n-k} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $m-1$ g.l.

Contraste de Ljung y Box

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula $H_0 : \rho(1) = \rho(2) = \ldots = \rho(m) = 0$ si

$$n(n+2)\sum_{k=1}^{m}\frac{r(k)^2}{n-k} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $m-1$ g.l.

```
> x <- scan("racha.txt")
> n <- length(x)
> rk <- acf(x)$acf[1:10]
> rk

 [1]  1.0000 -0.3073 -0.7404  0.7747  0.2052 -0.8982  0.3761
 [8]  0.6328 -0.7693 -0.1249

> m <- 4  # Contraste H0: rho1 = ... = rho4 = 0
> rk <- rk[2:(m + 1)]
> est <- n * (n + 2) * sum(rk^2/(n - 1:m))
> est

[1] 262.8

> 1 - pchisq(est, m - 1)

[1] 0
```

Contraste de Ljung y Box

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula $H_0 : \rho(1) = \rho(2) = \ldots = \rho(m) = 0$ si

$$n(n+2)\sum_{k=1}^{m}\frac{r(k)^2}{n-k} \geq \chi^2_\alpha$$

χ^2_α denota el punto tal que $P(J > \chi^2_\alpha) = \alpha$ siendo J una variable Ji-cuadrado con $m-1$ g.l.

- Podemos usar la función `Box.test`:

```
> x <- scan("racha.txt")
> n <- length(x)
> Box.test(x, lag = 4, type = "Ljung-Box")


Box-Ljung test

data:  x
X-squared = 262.8, df = 4, p-value < 2.2e-16
```

Test de Durbin-Watson

Hipótesis: ¿Se puede concluir que en una muestra de n individuos las observaciones son dependientes?

- Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n observaciones.
- El test se basa en calcular el estadístico

$$D_W = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Cuando la correlación es positiva toma valores cercanos a cero y cuando la correlación es negativa toma valores cercanos a 2.
- Este estadístico está indicado para detectar autocorrelación de primer orden en el conjunto de datos.

Test de Durbin-Watson

- Si la muestra es normal y $n > 20$ entonces

$$\frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \sim N(0,1),$$

lo que nos sirve para realizar el contraste.

Rechazamos la hipótesis nula $H_0 : \rho(1) = 0$ si

$$\frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

Test de Durbin-Watson

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula $H_0 : \rho(1) = 0$ si

$$\frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

```
> x <- scan("racha.txt")
> n <- length(x)
> DW <- sum(diff(x)^2)/(2 * sum((x - mean(x))^2))  # Estadístico DW
> DW

[1] 1.305

> est <- (DW - 1)/sqrt(1/(n + 1) * (1 - 1/(n - 1)))  # Estadístico de contraste
> 2 * (1 - pnorm(est))  # p-valor

[1] 1.474e-05
```

Test de Durbin-Watson

- Ejemplo: Disponemos de un conjunto de 200 medidas con las que planeamos llevar a cabo un contraste sobre la media.
- Las observaciones se encuentran en el fichero `racha.txt`. Nos gustaría contrastar la hipótesis de aleatoriedad de la muestra.

Rechazamos la hipótesis nula $H_0 : \rho(1) = 0$ si

$$\frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{D_W - 1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \geq z_{\alpha/2}$$

$z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo Z una variable $N(0,1)$

- Podemos usar la función `durbinWatsonTest` (paquete `car`)

```
> library(car)

Loading required package:  nnet

> x <- scan("racha.txt")
> n <- length(x)
> durbinWatsonTest(lm(x ~ 1))  # Debemos introducir un modelo lineal

 lag Autocorrelation D-W Statistic p-value
   1         -0.3073          2.61       0
 Alternative hypothesis: rho != 0
```

Contenidos: Introducción a las técnicas de remuestreo

Índice del curso

Introducción

- **Muestra natural**: un conjunto de observaciones de la población de estudio obtenidas mediante labor de campo.
- **Muestra artificial**: un conjunto de observaciones de dicha población no obtenidas mediante labor de campo (muestra de laboratorio).

La obtención de muestras artificiales requiere
el conocimiento del modelo de la población

Introducción

Aplicaciones:

- Aproximación por Monte Carlo de la distribución o características de un estadístico (sesgo y precisión: error cuadrático medio).
- Comparación de dos intervalos de confianza aproximando por Monte Carlo sus errores de recubrimiento.
- Comparación de dos contrastes de hipótesis aproximando por Monte Carlo sus funciones de potencia.

Contrastes de hipótesis usando Montecarlo

- Supongamos que tenemos un modelo completamente especificado y un estadístico T para el cual valores pequeños o grandes implican alejamientos del modelo.
- Para llevar a cabo un contraste de hipótesis habitual se necesita conocer la distribución de T.
- Si la distribución de T no está perfectamente determinada, se puede recurrir a la simulación.
- Simulando el modelo bajo la hipótesis nula H_0, se puede estimar el punto crítico al nivel α.

Ejemplo: Se quiere llevar a cabo un experimento para estudiar el tiempo medio de ejecución de un algoritmo. Se van a realizar $n = 300$ ejecuciones. El algoritmo propuesto se considerará aceptable si el tiempo medio es superior a 12 segundos pero en ningún caso se debería tener un tiempo medio superior a 12.5 segundos. Determina mediante simulación la región crítica y potencia del contraste teniendo en cuenta el tamaño muestral y suponiendo que el tiempo de ejecución es normal y que su desviación típica es de 3 segundos.

Bootstrap

"to pull oneself up by one's bootstrap"

- La idea central de este método Bootstrap (Efron, 1979) es simple.
- El método bootstrap se basa en la analogía entre la muestra y la población de la cual la muestra es extraída.
- Dada una muestra con n observaciones el estimador no paramétrico de máxima verosimilitud de la distribución poblacional es la función de densidad de probabilidad que asigna una masa de probabilidad $1/n$ a cada una de las observaciones.
- Dada una muestra aleatoria con n observaciones se trata a dicha muestra como si fuera toda la población y de ella extraeremos B muestras **con reemplazamiento**.
- Este enfoque tiene su antecedente inmediato en las técnicas de simulación Monte Carlo, las cuales consisten en extraer un número elevado de muestras aleatorias de una población conocida, para calcular a partir de ellas el valor del estadístico cuya distribución muestral pretende ser estimada.
- Los estudios teóricos han demostrado que el enfoque Bootstrap proporciona una buena aproximación de la **distribución** de los estimadores, lo cual permitirá describir algunas de sus propiedades muestrales, así como el cálculo de intervalos de confianza y la realización de contrastes de hipótesis.

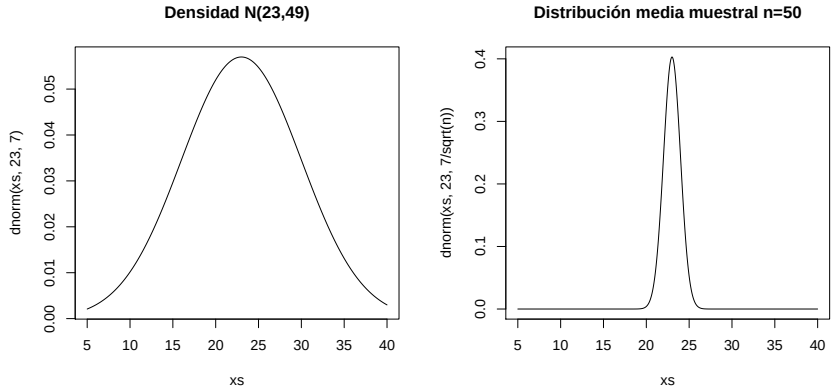
Bootstrap

Dada una muestra $X_1, \ldots, X_n$, el proceso se desarrolla mediante los pasos siguientes:

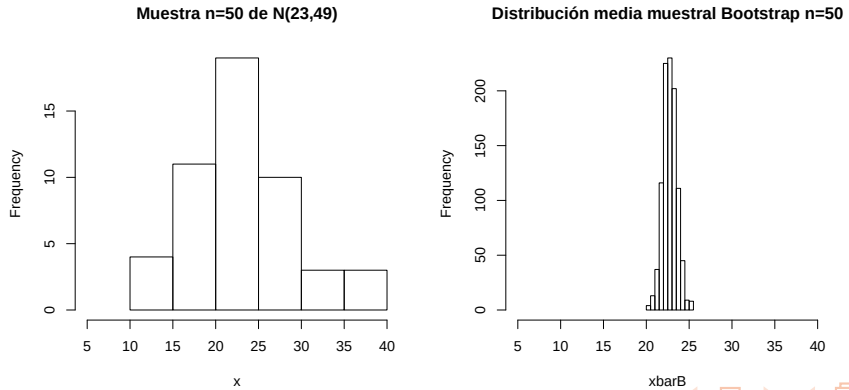
1. Se construye una distribución de probabilidad empírica, a partir de la muestra asignando una probabilidad de 1/n a cada punto, $X_i, i = 1, \ldots, n$.
2. Simular una muestra aleatoria simple de tamaño n con reemplazamiento (remuestra bootstrap).
3. Se calcula el estadístico de interés a partir de la remuestra.
4. Repetimos los pasos 2 y 3 B veces (B grande).
5. Construimos la distribución de probabilidad a partir de los B valores del estadístico de interés obtenidos de las B muestras.

Bootstrap

- Ejemplo: Consideramos una muestra de tamaño $n = 50$ obtenida de una N(23,49). Sabemos entonces que $\bar{X}$ se distribuye como una normal N(23,49/n). Mostramos en el siguiente gráfico ambas distribuciones.



- Ahora, a partir de la muestra original construimos 1000 muestras de tamaño $n = 50$ y calculamos la media muestral para cada una de ellas. En la gráfica aparece representado el histograma de una de las muestras y el histograma de las medias muestrales de las 1000 muestras (distribución Bootstrap). Observamos que la distribución Bootstrap presenta **aproximadamente la misma variabilidad y forma** que la distribución teórica de $\bar{X}$ aunque **los centros son diferentes**.



Bootstrap

- Ejemplo: Se obtuvieron los siguientes datos a partir de un estudio piloto:

$$56, 48, 44, 62, 50, 47, 49, 57, 48, 55, 96, 47, 46, 47, 49, 72, 46, 61$$

Determina el tamaño muestral requrido para obtener una potencia del 90% para rechazar $H_0 : \mu = 50$ cuando $\mu = 52$. Utiliza la metodología Bootstrap.

Contenidos: Contrastes en más dos poblaciones

Índice del curso

Contraste sobre la igualdad de medias en más dos poblaciones normales: muestras independientes

Hipótesis: ¿Se puede concluir que más de dos muestras independientes proceden de poblaciones con medias distintas?

- Disponemos de muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
  - ...
  -
  - $\{X_{k1}, X_{k2}, \ldots, X_{kn_k}\}$, n_k variables independientes y con la misma distribución $N(\mu_k, \sigma_k^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de las distintas poblaciones son distintos).
- Suponemos que las **varianzas $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, son iguales** ($\sigma_1^2 = \sigma_2^2 = \ldots, = \sigma_k^2$).

Contraste

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
$H_1 : \mu_i \neq \mu_j$ para algún $i \neq j$

Contraste sobre la igualdad de medias en más dos poblaciones normales: muestras independientes

- Ejemplo: Consideramos la siguiente muestra correspondiente a tiempos de respuesta (en milisegundos) de monitores de 3 modelos distintos

1.3	1.5	1.4	1.7	1.6		A
4.7	4.5	4.9	4.0			B
6.0	5.1	5.9	5.6	5.8	6.6	C

- Tenemos así una muestra de $n = 15$ elementos que se diferencian en un factor (modelo del monitor). En cada elemento de la muestra observamos una característica continua (tiempo de respuesta), que varía aleatoriamente de un elemento a otro.
- Nos interesa determinar si existen diferencias significativas en el tiempo de respuesta medio en los modelos de monitor.

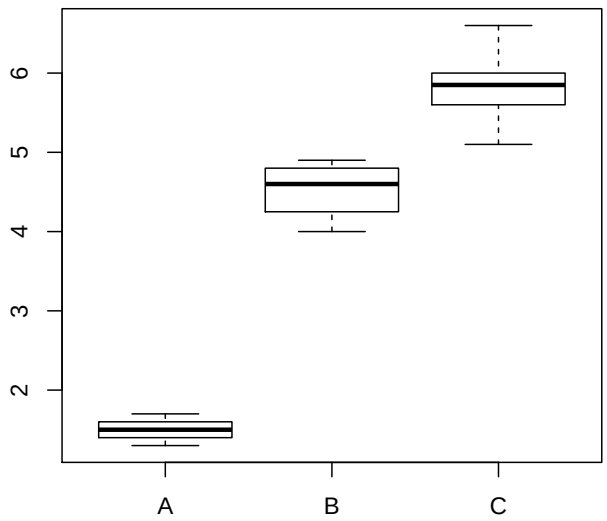
Contraste

$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_1 : \mu_i \neq \mu_j$ para algún $i \neq j$

Contraste sobre la igualdad de medias en más dos poblaciones normales: muestras independientes

- Ejemplo: Consideramos la siguiente muestra correspondiente a tiempos de respuesta (en milisegundos) de monitores de 3 modelos distintos

```
> A <- c(1.3, 1.5, 1.4, 1.7, 1.6)
> B <- c(4.7, 4.5, 4.9, 4)
> C <- c(6, 5.1, 5.9, 5.6, 5.8, 6.6)
> n1 <- length(A)
> n2 <- length(B)
> n3 <- length(C)
> ni <- c(n1, n2, n3)  # Tamaños muestrales de cada nivel
> n <- sum(ni)  # Tamaño muestral total
> tiempo <- c(A, B, C)
> modelo <- rep(c("A", "B", "C"), ni)
> x <- data.frame(tiempo, modelo)
> boxplot(tiempo ~ modelo, data = x)
```



Análisis de la varianza

- El objetivo del Análisis de la Varianza (ANOVA) es estudiar si existe relación entre el valor medio de una variable respuesta o característica (por ejemplo el tiempo de respuesta medio) y una o varias variables cualitativas o **factores** (por ejemplo el modelo de monitor).
  - Se denomina Análisis de la Varianza de un solo factor (one-way ANOVA) cuando se trabaja con un único factor.

Análisis de la varianza

- El Análisis de la Varianza nos permitirá determinar si un factor influye en la respuesta de un fenómeno que nos interesa estudiar.
- Entre los objetivos del Análisis de la Varianza está contrastar si las medias de la variable de interés en los distintos grupos determinados por el factor son iguales.
- Así, el ANOVA es una generalización del contraste para dos medias de la t de Student, cuando el número de muestras a contrastar es mayor que dos.

Análisis de la varianza

- Disponemos entonces de k muestras independientes:

$$\begin{array}{cccccc} X_{11} & X_{12} & \cdots & X_{1\,n_1} & \text{de una población} & N\left(\mu_1, \sigma^2\right) \\ X_{21} & X_{22} & \cdots & X_{2\,n_2} & \text{de una población} & N\left(\mu_2, \sigma^2\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{k\,n_k} & \text{de una población} & N\left(\mu_k, \sigma^2\right) \end{array}$$

- Cada una de las k muestras está formada por variables independientes y con la misma distribución (k muestras aleatorias simples).
- Se supone que las k muestras son, entre sí, independientes.
- A las medias se les permite ser distintas, pero las varianzas se suponen todas iguales (modelo **homocedástico**).
- n_i representa el número de observaciones de la respuesta para el nivel i del factor. Si $n_1 = n_2 = \ldots = n_k$ se dice que el diseño es equilibrado.
- $n = \sum_{i=1}^{k} n_i$ representa el número total de observaciones.

Análisis de la varianza

- El modelo ANOVA con un factor depende de $k+1$ parámetros desconocidos:
  - las medias $\mu_1, \ldots, \mu_k$
  - la varianza común σ^2.
- Estimamos la media μ_i de la población i mediante:

$$\bar{X}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

- Denotaremos

$$\bar{X}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{X}_{i\bullet}$$

a la media global, que se puede expresar como media de las $\bar{X}_{i\bullet}$ ponderadas por los tamaños muestrales n_i.

Análisis de la varianza

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos distintos.

```
> mean(tiempo)  # Media muestral global

[1] 4.04

> tapply(tiempo, modelo, mean)  # Medias muestrales de cada población

    A     B     C
1.500 4.525 5.833
```

$$\bar{X}_{1\bullet} = \frac{1.3 + 1.5 + 1.4 + 1.7 + 1.6}{5} = 1.5$$

$$\bar{X}_{2\bullet} = \frac{4.7 + 4.5 + 4.9 + 4.0}{4} = 4.525$$

$$\bar{X}_{3\bullet} = \frac{6.0 + 5.1 + 5.9 + 5.6 + 5.8 + 6.6}{6} = 5.833$$

$$\bar{X}_{\bullet\bullet} = \frac{1.3 + 1.5 + \ldots + 5.8 + 6.6}{15} = 4.04$$

Análisis de la varianza

Ejemplo 1

					$\bar{X}_{i\bullet}$
G1:	45	0	10	25	20
G2:	15	44	2	35	24
G3:	8	30	38	12	22
					$\bar{X}_{\bullet\bullet} = 22$

Ejemplo 2

					$\bar{X}_{i\bullet}$
G1:	20	20	21	19	20
G2:	24	25	23	24	24
G3:	22	21	23	22	22
					$\bar{X}_{\bullet\bullet} = 22$

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Ejemplo 3

					$\bar{X}_{i\bullet}$
G1:	50	20	0	-50	5
G2:	60	10	30	-20	20
G3:	40	50	0	90	45
					$\bar{X}_{\bullet\bullet} = 23.3$

Ejemplo 4

					$\bar{X}_{i\bullet}$
G1:	5	10	0	5	5
G2:	30	10	22	18	20
G3:	45	52	43	40	45
					$\bar{X}_{\bullet\bullet} = 23.3$

Análisis de la varianza

Ejemplo 1

					$\bar{X}_{i\bullet}$
G1:	45	0	10	25	20
G2:	15	44	2	35	24
G3:	8	30	38	12	22
					$\bar{X}_{\bullet\bullet} = 22$

Ejemplo 2

					$\bar{X}_{i\bullet}$
G1:	20	20	21	19	20
G2:	24	25	23	24	24
G3:	22	21	23	22	22
					$\bar{X}_{\bullet\bullet} = 22$

¿Las diferencias entre las medias son grandes comparadas con las diferencias entre los datos dentro de cada grupo?
La idea del test ANOVA es comparar la variabilidad entre las medias con la variabilidad dentro de cada grupo.

Ejemplo 3

					$\bar{X}_{i\bullet}$
G1:	50	20	0	-50	5
G2:	60	10	30	-20	20
G3:	40	50	0	90	45
					$\bar{X}_{\bullet\bullet} = 23.3$

Ejemplo 4

					$\bar{X}_{i\bullet}$
G1:	5	10	0	5	5
G2:	30	10	22	18	20
G3:	45	52	43	40	45
					$\bar{X}_{\bullet\bullet} = 23.3$

Análisis de la varianza

- Para determinar si hay diferencias significativas entre las respuestas medias a distintos niveles del factor, el ANOVA descompone la variabilidad de un experimento en componentes independientes que se asignan a causas distintas.
- ¿Cómo medimos la variabilidad total?

$$VT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\bullet\bullet})^2$$

- La variabilidad, medida a través de la desviación cuadrática de los datos a la media global, admite la siguiente descomposición:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})^2 + \sum_{i=1}^{k} n_i (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2$$

Análisis de la varianza

Fuente de variación	Suma de cuadrados	Grados de libertad
Entre poblaciones (VE)	$\sum_{i=1}^{k} n_i \left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2$	$k-1$
Error (VNE)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2$	$n-k$
Total (VT)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)^2$	$n-1$

- VE representa las desviaciones de las medias muestrales de cada población respecto a la media global.
  - Sirve como medición de la variabilidad entre poblaciones.
  - Es la variabilidad explicada por el modelo o por las diferencias entre niveles del factor.
- VNE representa las desviaciones de cada dato respecto a la media muestral de la población de la que procede.
  - Es útil como medida de la variabilidad interna, presente entre los individuos de la misma población.
  - Servirá para estimar σ^2 y por eso le llamamos error del modelo o varianza residual.

Análisis de la varianza

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos distintos.

```
> media <- mean(tiempo)  # Media muestral global
> mediai <- tapply(tiempo, modelo, mean)  # Media muestral de cada población
> VEaux <- ni * (mediai - media)^2  # Variabilidades intergrupo
> VEaux

      A       B       C
32.2580  0.9409 19.2963

> VNEaux <- tapply(tiempo, modelo, var) * (ni - 1)  # Variabilidades intragrupo
> VNEaux

     A      B      C
0.1000 0.4475 1.2133
```

Nivel i	n_i	$\bar{X}_{i\bullet}$	$n_i\left(\bar{X}_{i\bullet}-\bar{X}_{\bullet\bullet}\right)^2$	$\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\bullet}\right)^2$
$i=1$	5	1.5	$5\cdot(1.5-4.04)^2=32.258$	0.1
$i=2$	4	4.525	$4\cdot(4.525-4.04)^2=0.9409$	0.4475
$i=3$	6	5.833	$6\cdot(5.833-4.04)^2=19.29$	1.2133

Análisis de la varianza

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos.

```
> VE <- sum(VEaux)  # Variabilidas explicada
> VE

[1] 52.5

> VNE <- sum(VNEaux)  # Variabilidad no explicada
> VNE

[1] 1.761

> VT <- VE + VNE  # Variabilidad total
> VT

[1] 54.26

> var(tiempo) * (n - 1)

[1] 54.26
```

Nivel i	n_i	$\bar{X}_{i\bullet}$	$n_i\,(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2$	$\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2$
$i = 1$	5	1.5	$5 \cdot (1.5 - 4.04)^2 = 32.258$	0.1
$i = 2$	4	4.525	$4 \cdot (4.525 - 4.04)^2 = 0.9409$	0.4475
$i = 3$	6	5.833	$6 \cdot (5.833 - 4.04)^2 = 19.29$	1.2133
			$VE = 52.4951$	$VNE = 1.7608$

$$VT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = (1.3 - 4.04)^2 + \ldots + (6.6 - 4.04)^2 = 54.256$$

$$VE + VNE = 52.4951 + 1.7608 = 54.256$$

Análisis de la varianza

- Se verifica

$$\frac{VNE}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})^2 \in \chi^2_{n-k}$$

- De esta expresión obtenemos como estimador insesgado de σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\bullet})^2$$

Análisis de la varianza

- Para efectuar el contraste de igualdad de las medias

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

debemos considerar un estadístico que mida la discrepancia respecto a la hipótesis nula de igualdad.

- Si las medias fueran iguales, las desviaciones entre poblaciones no deberían ser muy grandes, comparadas con las desviaciones dentro de cada población.

Fuente de variación	Suma de cuadrados	Grados de libertad
Entre poblaciones (VE)	$\sum_{i=1}^{k} n_i \left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2$	$k-1$
Error (VNE)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2$	$n-k$
Total (VT)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)^2$	$n-1$

Análisis de la varianza

- Si la hipótesis nula $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ es cierta, se verifica

$$\frac{VE}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2 \in \chi^2_{k-1}$$

- Xa hemos dicho que

$$\frac{VNE}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2 \in \chi^2_{n-k}$$

- Un estadístico razonable para el contraste podría ser entonces

$$F = \frac{VE/(k-1)}{VNE/(n-k)} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2 /(k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2 /(n-k)}$$

Análisis de la varianza

$$F = \frac{VE/(k-1)}{VNE/(n-k)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_{i\bullet}\right)^2/(n-k)}$$

- Si la hipótesis nula $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ es cierta, F presenta una distribución F de Snédecor $F \in F_{(k-1),(n-k)}$

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ si

$$\frac{VE/(k-1)}{VNE/(n-k)} \geq f_\alpha$$

f_α denota el punto tal que $P(F > f_\alpha) = \alpha$ siendo F una variable F de Snedecor con $k-1, n-k$ g.l.

Análisis de la varianza

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos.

```
> k <- 3  # Número de muestras
> est <- (VE/(k - 1))/(VNE/(n - k))  # Estadístico F de contraste
> est

[1] 178.9

> 1 - pf(est, k - 1, n - k)  # p-valor

[1] 1.168e-09
```

- Por lo tanto, rechazamos la hipótesis nula de igualdad de medias.

Análisis de la varianza

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos.
- Podemos utilizar la función `lm`

```
> z <- lm(tiempo ~ modelo, data = x)
> summary(z)


Call:
lm(formula = tiempo ~ modelo, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-0.733 -0.150  0.000  0.171  0.767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.500      0.171    8.76  1.5e-06 ***
modeloB        3.025      0.257   11.77  6.0e-08 ***
modeloC        4.333      0.232   18.68  3.1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.383 on 12 degrees of freedom
Multiple R-squared:  0.968, Adjusted R-squared:  0.962
F-statistic:  179 on 2 and 12 DF,  p-value: 1.17e-09
```

Comparaciones múltiples

- Si se rechaza $H_0 : \mu_1 = \ldots = \mu_k$, nos podemos preguntar qué medias son diferentes
- Queremos determinar qué parejas de medias son distintas entre si y estimar las diferencias $\mu_i - \mu_j$.
- En principio, podríamos usar el pivote

$$\frac{\bar{X}_{i\bullet} - \bar{X}_{j\bullet} - (\mu_i - \mu_j)}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

Así, rechazaríamos $H_0 : \mu_i = \mu_j$ si

$$\frac{|\bar{X}_{i\bullet} - \bar{X}_{j\bullet}|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > t_{n-k,\alpha/2}.$$

Análisis de la varianza. Comparaciones múltiples

- ¿Cuántos contrastes del tipo $H_0 : \mu_i = \mu_j$ se pueden realizar? Si tenemos k niveles, es posible formar $k(k-1)/2$ comparaciones, es decir, podemos realizar $k(k-1)/2$ contrastes $H_0 : \mu_i = \mu_j$.
- Surge un problema (relacionado con el nivel de significación) si aplicamos simultáneamente en todas las comparaciones el contraste descrito previamente.
- Si todas las medias son iguales, cada comparación tiene una probabilidad α de rechazar la hipótesis nula $H_0 : \mu_i = \mu_j$, pero al hacer muchas comparaciones la probabilidad de que alguna de ellas produzca un rechazo es mucho mayor que α.
- No es que las comparaciones anteriores sean incorrectas, simplemente se trata de que respetan el nivel de significación si se consideran individualmente, pero no respetan un nivel de significación conjunto o múltiple.
- Como solución, utilizamos el método de Bonferroni para calcular intervalos de confianza o realizar contrastes múltiples

Método de Bonferroni

- Consiste en rectificar el nivel de significación empleado en cada comparación, para que el nivel conjunto siga respetando un cierto valor α.
- Así, si hacemos m comparaciones, en cada una de ellas usaremos como nivel de significación α/m.
- En nuestro caso, $m = k(k-1)/2$. Entonces el contraste simultáneo consistirá en rechazar $H_0 : \mu_i = \mu_j$ si

$$\frac{|\bar{X}_{i\bullet} - \bar{X}_{j\bullet}|}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} > t_{n-k,\alpha/(2m)}.$$

Método de Bonferroni

En términos de intervalos de confianza, y usando los mismos pivotes ya indicados, los intervalos adoptan la forma:

$$\left(\bar{X}_{i\bullet} - \bar{X}_{j\bullet} - t_{n-k,\alpha/(2m)}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}, \bar{X}_{i\bullet} - \bar{X}_{j\bullet} + t_{n-k,\alpha/(2m)}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right)$$

Así, se considera que hay discrepancia significativa entre las medias de dos poblaciones si el intervalo de confianza correspondiente a su diferencia de medias no contiene al cero. Este criterio es equivalente al que hemos planteado con anterioridad.

Contraste sobre la igualdad de varianzas en más dos poblaciones normales

Hipótesis: ¿Se puede concluir que más de dos muestras independientes proceden de poblaciones con varianzas distintas?

- Disponemos de muestras:
  - $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$, n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1^2)$.
  - $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$, n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2^2)$
  - ...
  - $\{X_{k1}, X_{k2}, \ldots, X_{kn_k}\}$, n_k variables independientes y con la misma distribución $N(\mu_k, \sigma_k^2)$
- Suponemos que las muestras son **independientes** (los individuos donde se han obtenido las mediciones de las distintas poblaciones son distintos).
- La suposición de igualdad de varianzas u homocedasticidad se puede contrastar mediante el test de Levene

Contraste

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algún } i \neq j$$

Contraste sobre la igualdad de varianzas en más dos poblaciones normales

- Ejemplo: Consideramos la siguiente muestra correspondiente a tiempos de respuesta (en milisegundos) de monitores de 3 modelos distintos

1.3	1.5	1.4	1.7	1.6		A
4.7	4.5	4.9	4.0			B
6.0	5.1	5.9	5.6	5.8	6.6	C

- Tenemos así una muestra de $n = 15$ elementos que se diferencian en un factor (modelo del monitor). En cada elemento de la muestra observamos una característica continua (tiempo de respuesta), que varía aleatoriamente de un elemento a otro.
- Nos interesa determinar si existen diferencias significativas en la varianza del tiempo de respuesta en los modelos de monitor.

Contraste

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
$H_1 : \sigma_i^2 \neq \sigma_j^2$ para algún $i \neq j$

Contraste sobre la igualdad de varianzas en más dos poblaciones normales

- Para efectuar el contraste, vamos a considerar las desviaciones absolutas de cada dato a la media de su grupo:

$$Z_{ij} = \left|X_{ij} - \bar{X}_{i\bullet}\right| \qquad j \in \{1, \ldots, n_i\} \quad i \in \{1, \ldots, k\}$$

- El test de Levene consiste en efectuar un test F sobre los valores Z_{ij}.

$$L = \frac{\sum_{i=1}^{k} \sum_{j=i}^{n_i} \left(\bar{Z}_{i\bullet} - \bar{Z}_{\bullet\bullet}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=i}^{n_i} \left(Z_{ij} - \bar{Z}_{i\bullet}\right)^2 / (n-k)}$$

  - $\bar{Z}_{i\bullet} = n_i^{-1} \sum_{j=1}^{n_i} Z_{ij}$ la media de los Z_{ij} en el grupo i-ésimo,
  - $\bar{Z}_{\bullet\bullet} = n^{-1} \sum_{i=1}^{I} \sum_{j=i}^{n_i} Z_{ij}$ la media global de todos los Z_{ij}.

Rechazamos la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ si

$$\frac{\sum_{i=1}^{k} \sum_{j=i}^{n_i} \left(\bar{Z}_{i\bullet} - \bar{Z}_{\bullet\bullet}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=i}^{n_i} \left(Z_{ij} - \bar{Z}_{i\bullet}\right)^2 / (n-k)} \geq f_\alpha$$

f_α denota el punto tal que $P(F > f_\alpha) = \alpha$ siendo F una variable F de Snedecor con $k-1, n-k$ g.l.

Contraste sobre la igualdad de varianzas en más dos poblaciones normales

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos.
- Utilizamos los residuos del ANOVA:

```
> z <- lm(tiempo ~ modelo, data = x)
> zij <- abs(z$residuals)
> lev <- lm(zij ~ x$modelo)
> summary(lev)


Call:
lm(formula = zij ~ x$modelo)

Residuals:
   Min     1Q Median     3Q    Max
-0.300 -0.143 -0.020  0.090  0.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.120      0.110    1.10     0.29
x$modeloB      0.155      0.164    0.94     0.36
x$modeloC      0.213      0.148    1.44     0.18

Residual standard error: 0.245 on 12 degrees of freedom
Multiple R-squared:  0.152, Adjusted R-squared:  0.0101
F-statistic: 1.07 on 2 and 12 DF,  p-value: 0.373
```

- No rechazamos la hipótesis nula de igualdad de varianzas

Contraste sobre la igualdad de varianzas en más dos poblaciones normales

- Ejemplo: Volvemos al ejemplo sobre tiempos de respuesta (en milisegundos) de monitores de 3 modelos.
- Podemos utilizar la función `levene.test` (paquete `lawstat`)

```
> library(lawstat)
> levene.test(x$tiempo, x$modelo, location = "mean")


classical Levene's test based on the absolute
deviations from the mean ( none not applied because
the location is not set to median )

data:  x$tiempo
Test Statistic = 1.071, p-value = 0.3732
```

Variable	Número de muestras		Hipótesis evaluada	Test
Una variable cuantitativa	Una muestra		Hipótesis sobre la media poblacional u otra medida de posición central	▸ z-test para una media (Distribución normal, σ conocida) ▸ t-test para una media (Distribución normal, σ desconocida) ▸ Tests de los signos para la mediana ▸ Test de los rangos signados de Wilcoxon
			Hipótesis sobre la varianza poblacional	▸ Contraste χ^2 sobre la varianza (Distribución normal)

			Hipótesis sobre la distribución	▸ Test de Kolmogorov-Smirnov para una muestra
				▸ Tests específicos para contrastar normalidad:
				  ▸ Test de Lilliefors
				  ▸ Test de asimetria
				  ▸ Test de curtosis
				  ▸ Test de Jarque-Bera
				  ▸ Test de Shapiro-wilk

			Hipótesis sobre la aleatoriedad de la muestra	▸ Test de rachas ▸ Test de autocorrelación ▸ Test de Durbin-Watson
	Dos muestras	Muestras independientes	Hipótesis sobre la diferencia de medias o de otras medidas de tendencia central	▸ z-test para la diferencia de medias (Distribución normal, varianzas conocidas) ▸ t-test para la diferencia de medias (Distribución normal, varianzas desconocidas pero iguales) ▸ Test de Welch para diferencias de medias (Distribución normal, varianzas desconocidas y desiguales) ▸ Test de Wilcoxon-Mann-Whitney
			Hipótesis sobre dos varianzas	▸ Contraste F para dos varianzas (Distribución normal)
			Contrastar la distribución	▸ Test de Kolmogorov-Smirnov para dos muestras
		Muestras apareadas	Hipótesis sobre la diferencia de medias	▸ t-test para muestras apareadas (Distribución normal) ▸ Test de los signos para muestras apareadas ▸ Test de los rangos signados de Wilcoxon para muestras apareadas
	Más de dos muestras	Muestras independientes	Hipótesis sobre la igualdad de medias o de otras medidas de tendencia central	▸ ANOVA de un factor (poblaciones normales con varianzas iguales) ▸ Test de Kruskal-Wallis
			Hipótesis sobre la igualdad de varianzas	▸ Test de Levene
		Muestras dependientes	Hipótesis sobre la igualdad de medias o de otras medidas de tendencia central	▸ ANOVA con medidas repetidas ▸ Test de Friedman

Variable	Número de muestras		Hipótesis evaluada	Test
Una variable cualitativa	Una muestra		Hipótesis sobre una proporción	▸ z-test para una proporción
			Hipótesis sobre la distribución	▸ Test Chi-cuadrado de bondad de ajuste
	Dos muestras	Muestras independientes	Hipótesis sobre la diferencia de proporciones	▸ z-test para la diferencia de proporciones
			Hipótesis sobre la distribución de dos poblaciones independientes	▸ Test Chi-cuadrado de homogeneidad

Variable	Número de muestras	Hipótesis evaluada	Test
Dos variables cuantitativas	Una muestra bidimensional	Hipótesis sobre la asociación entre las variables	▸ Test sobre el coeficiente de correlación de Pearson ▸ Test de correlación por rangos de Spearman
Dos variables cualitativas	Una muestra bidimensional	Hipótesis sobre la asociación entre las variables	▸ Test Chi-cuadrado de independencia para tablas de contingencia