TEMA 1. ALGUNOS MODELOS CONTINUOS DE VARIABLES ALEATORIAS. INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

1. Algunos modelos continuos de variable aleatoria:
   - Distribución uniforme
   - Distribución normal
   - Distribución chi-cuadrado
   - Distribución t-student
   - Distribución F-snedecor
2. Concepto de muestra y estadístico
3. Valor esperado y varianza de la media muestral
4. Valor esperado de la varianza y cuasivarianza muestral

DISTRIBUCIÓN UNIFORME

Una variable aleatoria continua X sigue una distribución uniforme $(U(a,b))$ cuando su función de densidad viene dada por

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{, si } x \in (a,b) \\ 0 & \text{, en otro caso} \end{cases}$$

Características:

Función de distribución

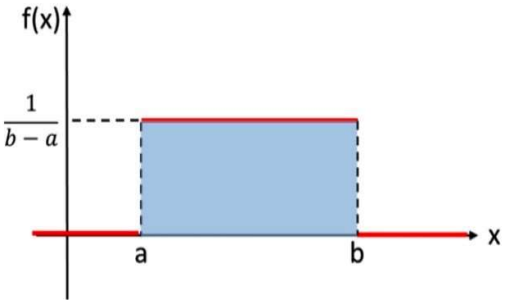
$$F(x) = \begin{cases} 0 & \text{, si } x < a \\ \frac{x-a}{b-a} & \text{, si } a \leq x < b \\ 1 & \text{, si } x \geq b \end{cases}$$

Esperanza matemática

$$E[X] = \frac{a+b}{2}$$

Varianza

$$Var(X) = \frac{(b-a)^2}{12}$$



La probabilidad de que tome un valor dentro de ese intervalo es la misma para cualquier sub-intervalo de igual longitud

R: función de distribución
punif(**cuantil**,min=**número**,max=**número**, lower.tail=**T/F**)
Cuantiles
qunif(**probabilidad**,min=**número**,max=**número**, lower.tail=**T/F**)

Ejemplo. El número de automóviles que circulan por una plaza se considera que sigue una distribución uniforme. Si al cabo de la tarde circulan como máximo 200 coches

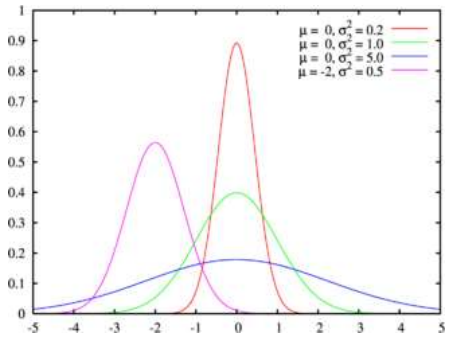
a) halle la probabilidad de que en una tarde pasen como mínimo 100 coches

b) Hallar el percentil 20

DISTRIBUCIÓN NORMAL

Una variable aleatoria continua X sigue una distribución normal de parámetros μ y σ $(N(\mu, \sigma))$ cuando su función de densidad viene dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},\ \sigma > 0,\ \mu \in \mathbb{R}.$$



Características:

Esperanza matemática $E[X] = \mu$

Varianza $Var(X) = \sigma^2$

La distribución normal o de Gauss modeliza gran cantidad de variables. El Teorema Central del Límite (que veremos más adelante) hace que sea la distribución más importante y más utilizada.

R: función de distribución
pnorm(**cuantil**, mean = $\boldsymbol{\mu}$, sd = $\boldsymbol{\sigma}$,lower.tail = **T/F**)
Cuantiles
qnorm(**probabilidad**, mean = $\boldsymbol{\mu}$, sd = $\boldsymbol{\sigma}$,lower.tail = **T/F**)

En un examen las notas siguen una distribución Normal de media 78 y varianza 100. Halle la probabilidad de que, en ese examen, un estudiante obtenga entre 76 y 80 puntos

¿Cuál es la nota máxima para el 20% de los peores alumnos?

Propiedad de reproductividad. La combinación lineal de distribuciones normales sigue siendo una distribución normal, cuya media es la combinación lineal de las medias de cada distribución y cuya varianza es la combinación lineal de las varianzas de cada distribución con los coeficientes al cuadrado

$$\left.\begin{array}{l} Y = \sum_{i=1}^{n} a_i \cdot X_i \\ X_i \sim N(\mu_i, \sigma_i) \end{array}\right\} \quad Y \sim N\left(\sum_{i=1}^{n} a_i \cdot \mu_i, \sum_{i=1}^{n} a_i^2 \cdot \sigma_i^2\right)$$

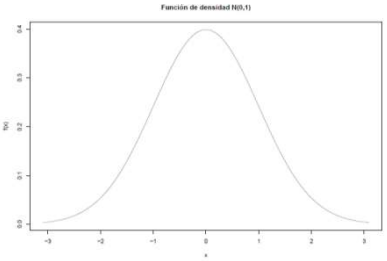
Una variable aleatoria continua Z sigue una distribución **normal de parámetros 0 y 1** $(N(0,1))$ si se distribuye como una normal de media 0 y varianza 1

Características:

Función de densidad $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad \forall z \in \mathbb{R}.$

Esperanza matemática $E[Z] = 0.$

Varianza $Var(Z) = 1$



Tipificación: Sea X una variable aleatoria con distribución normal de parámetros μ y σ. La **variable aleatoria tipificada**

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

se distribuye como una normal de parámetros 0 y 1.

Dadas tres variables independientes con medias y varianzas respectivas: $X_1 \sim N(5,2), X_2 \sim N(3,1), X_3 \sim N(6,4)$ y dada otra variable $Y \sim 3X_1 + 2X_2 + X_3$, calcular la probabilidad del suceso $Y \geq 16$.

En un examen las notas siguen una distribución Normal de media 78 y varianza 100. Halle la probabilidad de que, en ese examen, un estudiante obtenga entre 76 y 80 puntos (TIPIFICANDO)

DISTRIBUCIÓN CHI- CUADRADO

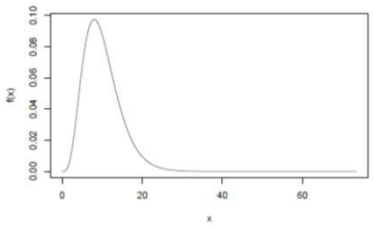
Una variable aleatoria continua *X* sigue una distribución chi-cuadrado con *n* grados de libertad (χ_n^2) si se construye como la suma de *n* variables independientes e idénticamente distribuidas según una *N(0,1)*:

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2, \quad X_i \sim N(0,1), \forall i.$$

Al parámetro *n* se le denomina **grados de libertad** y coincide con el número de variables independientes usadas.

Propiedad de reproductividad. La distribución chi-cuadrado es reproductiva en sus grados de libertad:

$$X_i \sim \chi_{n_i}^2, i = 1, \dots, k \Rightarrow \sum_{i=1}^{k} X_i \sim \chi_{n_1+\cdots+n_k}^2$$



R: función de distribución
pchisq(**cuantil**, **grados de libertad**, lower.tail = **T/F**)
Cuantiles
qchisq(**probabilidad**, **grados de libertad** ,lower.tail = **T/F**)

Sea una variable que sigue una distribución chi-cuadrado con 6 grados de libertad. ¿cuánto vale el percentil 90?

¿Cuál es la probabilidad de que la variable tome valores menores o iguales a 2.2?

DISTRIBUCIÓN t-STUDENT

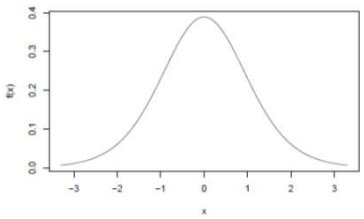
Una variable aleatoria continua T sigue una t-student con *n* grados de libertad (t_n) si se construye como el cociente entre una variable aleatoria distribuida como una N(0,1) y la raíz cuadrada de una chi-cuadrado con n grados de libertad dividida entre sus grados de libertad, siendo ambas variables independientes:

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}, \quad X \sim N(0,1),\ Y \sim \chi^2_n.$$

Características:

Es una distribución simétrica

Para grados de libertad altos, la distribución t-student se aproxima a la normal tipificada



R: función de distribución
pt(**cuantil**, **grados de libertad**, lower.tail = **T/F**)
Cuantiles
qt(**probabilidad**, **grados de libertad** ,lower.tail = **T/F**)

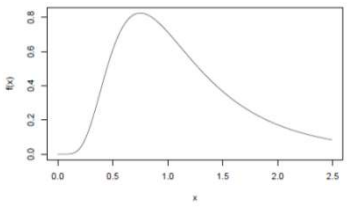
Sea T una variable aleatoria que sigue una distribución t de Student con 12 grados de libertad. Determine el percentil 90

¿Cuál es la probabilidad de que la variable se mayor que 1.782?

DISTRIBUCIÓN F-SNEDECOR

Una variable aleatoria continua F sigue una F-Snedecor con *n* y m grados de libertad $(F_{n,m})$ si se construye como el cociente de dos variables aleatorias independientes y distribuidas según chi-cuadrados con n y m grados de libertad, respectivamente, divididas entre sus correspondientes grados de libertad.

$$F = \frac{\frac{X}{n}}{\frac{Y}{m}}, \qquad X \sim \chi^2_n,\ Y \sim \chi^2_m.$$



R: función de distribución
pf(**cuantil, grados de libertad 1, grados de libertad 2**, lower.tail = **T/F**)
Cuantiles
qf(**probabilidad**, **grados de libertad 1, grados de libertad 2** ,lower.tail = **T/F**)

Sea X una variable que se distribuye según una F-Snedecor de 3 y 7 grados de libertad. Obtenga el percentil 90

¿Cuál es la probabilida de que la variable sea menor que 4.35?

1.2 CONCEPTO DE MUESTRA Y ESTADÍSTICO

Un problema fundamental en la estadística consiste en estudiar alguna característica desconocida de una población (variable aleatoria). Dicha cuestión será estudiada en los próximos temas, se presentan a continuación conceptos necesarios para dicho estudio.

Población: Conjunto de elementos sobre los que se quieres estudiar una o más características. El objetivo es inferir conclusiones sobre valores numéricos de la población, como la media, la proporción, …,etc. Como suele ser muy grande (coste en tiempo y dinero) se selecciona una muestra.

Muestra: Subconjunto representativo (con características similares a la población) de elementos de la población. Es fundamental elegir adecuadamente los elementos de la muestra y el tamaño muestral. Normalmente trabajaremos con lo que se conoce como **muestreo aleatorio simple** (**m.a.s.**; todos los elementos de la población tienen la misma probabilidad de estar en la muestra y son independientes entre ellos).

Hay que distinguir entre muestra y población. Así hablaremos de:

Parámetro poblacional: característica numérica de la población que sirve para distinguir total o parcialmente la distribución de probabilidad de la variable aleatoria estudiada. Los más usados son la esperanza matemática y la varianza.

Estadístico muestral: cualquier función real de las v.a. que forman la muestra, es decir, es una función de las observaciones muestrales y no contiene ningún valor o parámetro desconocido

Resumiendo, se entiende por parámetro poblacional a las características desconocidas de las distribuciones de probabilidad, mientras que el estadístico muestral es un resumen de la información existente en la muestra.

Parámetros poblacionales	Estadísticos muestrales
Tamaño población: *N*	Tamaño muestra: *n*
Media poblacional: $E[X]$	Media muestral: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$
Varianza poblacional: $Var(X)$	Varianza muestral: $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$ Cuasivarianza muestral: $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$
Proporción poblacional: *p*	Proporción muestral: $\hat{p} = \sum_{i=1}^{n} \frac{A_i}{n}$ $A_i = \begin{cases} 1 & \text{, ocurre } i \\ 0 & \text{, en otro caso} \end{cases}$

1.3 MEDIA MUESTRAL: SU ESPERANZA Y SU VARIANZA

Dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ de tamaño *n* de una variable aleatoria *X*, se define la media muestral como:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Ejemplo. Sea X la variable aleatoria definida como tiempo (en horas) transcurridos en realizar el examen final de cierta asignatura. Durante la ultima convocatoria se anotó el tiempo que tardó cada alumno en realizar dicho examen, obteniéndose la siguiente muestra: 1, 1.7, 1.8, 1.5, 1.85, 1.3, 2, 1.9, 1.6, 1.75, 1.8, 1.9.

$$\overline{X} = \frac{1 + 1'7 + 1'8 + 1'5 + 1'85 + \cdots + 1'75 + 1'8 + 1'9}{12} = \frac{20'1}{12} = 1'675$$

La esperanza de la media muestral es igual a la media poblacional: $\boxed{E[\overline{X}] = E[X]}$

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X] = \frac{1}{n}\cdot n \cdot E[X] = E[X]$$

La varianza de la media muestral es igual a la varianza poblacional dividida por el tamaño muestral:

$$\boxed{Var(\overline{X}) = \frac{Var(X)}{n}}$$

1.4 VARIANZA Y CUASIVARIANZA MUESTRALES: SU ESPERANZA

Dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ de tamaño *n* de una variable aleatoria *X*, se define la **varianza muestral** como:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Y la **cuasivarianza muestral** será:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

La varianza y la cuasivarianza muestral se relacionan mediante la siguiente igualdad:

$$n \cdot S_n^2 = (n-1) \cdot S_{n-1}^2$$

La esperanza de la cuasivarianza muestral es igual a la varianza poblacional: $E[S_{n-1}^2] = Var(X)$

La esperanza de la varianza muestral es igual a la varianza poblacional multiplicada por el factor $\frac{n-1}{n}$

$$E[S_n^2] = \frac{n-1}{n} \cdot Var(X)$$

Ejemplo

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	-0'675	0'4556
1'7	0'025	0'0006
1'8	0'125	0'0156
1'5	-0'175	0'0306
1'85	0'175	0'0306
1'3	-0'375	0'1406
2	0'325	0'1056
1'9	0'225	0'0506
1'6	-0'075	0'0056
1'75	0'075	0'0056
1'8	0'125	0'0156
1'9	0'225	0'0506
		0'9075

$$S_n^2 = \frac{0'9075}{12} = 0'0756$$

$$S_{n-1}^2 = \frac{0'9075}{11} = 0'0825$$

Técnicas Cuantitativas II

TEMA 2. ESTIMACIÓN PUNTUAL DE PARÁMETROS

Concepto de estimador paramétrico
Obtención de estimadores puntuales
Propiedades deseables de un estimador paramétrico

Índice

Introducción a la estimación

Método de máxima verosimilitud

Método de los momentos

Propiedades deseables para un estimador paramétrico

Concepto de estimador de un parámetro

Supongamos que el número medio de mensajes multimedia recibidos en el móvil semanalmente por los estudiantes de la Universidad de Granada puede considerarse como una variable aleatoria con distribución Poisson de parámetro λ. Un problema fundamental en estadística es conocer el valor del parámetro desconocido λ, para lo cual existen distintas técnicas.

En el presente tema se presenta la estimación puntual, que consiste en obtener un único número calculado a partir de las observaciones muestrales y que es utilizado como estimación del valor del parámetro desconocido. Así, se entiende por **estimador** un estadístico muestral usado para calcular una aproximación numérica de un parámetro desconocido de la población. Esto es:

Definición 1 Sea θ el parámetro desconocido de una determinada distribución de probabilidad. Un estimador de θ, que se denota como $\hat{\theta}(x_1, \ldots, x_n)$, será un estadístico muestral que aproxime el verdadero valor de θ a partir de la muestra.

Téngase en cuenta que el estimador depende sólo de la muestra y, en ningún caso, del parámetro que se desea estimar. Por tanto, el estimador será un valor numérico distinto dependiendo de la muestra considerada. Además, para cada parámetro pueden existir varios estimadores diferentes. En general, escogeremos aquel que posea mejores propiedades.

Método de máxima verosimilitud

En esta sección empezamos a estudiar los métodos de obtención de estimadores puntuales. Aunque se pueden obtener estimadores de forma simultánea para todos los parámetros poblacionales desconocidos de una distribución, nos centraremos en estudiar el caso en el que la distribución de probabilidad depende de un único parámetro desconocido.

El **método de máxima verosimilitud** consiste en escoger como estimador puntual aquel valor del parámetro que maximice la probabilidad de aparición de los valores muestrales obtenidos, es decir, aquel que maximice la función de probabilidad.

Ejemplo 1 Supongamos que el número de inundaciones por año sigue una distribución de Poisson con parámetro desconocido λ. En una determinada provincia han ocurrido dos inundaciones (valor observado) durante un año, ¿cómo se puede estimar el valor de λ si solo conocemos una observación?

Si los posibles valores de λ fueran 1, 1'5, 2, 2'5 y 3, la probabilidad de obtener la observación dos sería, respectivamente, 0'1839, 0'2510, 0'2770, 0'2565 y 0'2250.

Observando las tablas correspondientes a la distribución de Poisson se aprecia que el valor de λ que maximiza la probabilidad de aparición del valor observado es 2. ♯

Método de máxima verosimilitud

Sea X una variable aleatoria (discreta o continua) cuya distribución de probabilidad depende de un único parámetro desconocido θ. Dada una muestra aleatoria simple, $X_1, \ldots, X_n$, de X, se define la función de verosimilitud como

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n} P[X_i = x_i; \theta] & \text{, caso discreto} \\ \prod_{i=1}^{n} f(x_i; \theta) & \text{, caso continuo} \end{cases},$$

donde $p_i = P[X_i = x_i; \theta]$ denota la función de cuantía y f la función de densidad, dependiendo de la naturaleza de la variable.
Entonces, el método para encontrar el estimador de θ se convierte en el problema matemático de maximizar la función $L(\theta; x_1, x_2, \ldots, x_n)$. Es decir, hallar aquellos valores que anulan la primera derivada de $L(\theta; x_1, x_2, \ldots, x_n)$ y hacen negativa a la segunda.

Método de máxima verosimilitud

Índice
Introducción a la estimación
Método de máxima verosimilitud
Método de los momentos
Propiedades deseables para un estimador paramétrico

Luego $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$ será el estimador máximo verosímil del parámetro θ si verifica que:

$$\begin{aligned} \frac{\partial}{\partial \theta} L(\widehat{\theta}; x_1, x_2, \ldots, x_n) &= 0, \\ \frac{\partial^2}{\partial \theta^2} L(\widehat{\theta}; x_1, x_2, \ldots, x_n) &< 0. \end{aligned}$$

Puesto que al derivar directamente la función de verosimilitud se suelen obtener expresiones complicadas, normalmente se maximiza el logaritmo neperiano de la función de verosimilitud, ya que la solución obtenida es la misma al ser monótona creciente la función logaritmo.
Por tanto, el estimador máximo verosímil será aquel valor $\widehat{\theta}$ que verifique que:

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln L(\widehat{\theta}; x_1, x_2, \ldots, x_n) &= 0, \\ \frac{\partial^2}{\partial \theta^2} \ln L(\widehat{\theta}; x_1, x_2, \ldots, x_n) &< 0. \end{aligned}$$

A modo de resumen, para obtener un estimador por máxima verosimilitud, los pasos a seguir son los siguientes:

- Calcular la función de verosimilitud de la muestra teniendo en cuenta la distribución de probabilidad de los datos obtenidos.
- Determinar el logaritmo de la función de verosimilitud. Se puede demostrar que si una función alcanza un máximo en un punto, su logaritmo neperiano alcanza el máximo en ese mismo punto. Esto nos facilitará buscar el estimador máximo-verosímil, ya que generalmente es más fácil determinar el máximo del logaritmo de una función que directamente el de la función.
- Derivar el logaritmo neperiano con respecto al parámetro que se pretende estimar e igualar a cero para determinar el valor que maximiza dicha función. De esta expresión despejaremos el estimador, comprobando que la segunda derivada evaluada en dicho punto es negativa.

Método de máxima verosimilitud: Ejemplo 1

A continuación vamos a obtener el estimador máximo verosímil del parámetro λ de la distribución de Poisson. Después se obtendrá su valor para las siguientes muestras correspondientes al número de visitas semanales recibidas por un profesor en tutorías:

Muestra 1: $3, 3, 6, 2, 1, 4, 0, 2, 2, 5.$

Muestra 2: $1, 3, 1, 4, 5, 3, 4, 6, 1, 3.$

Dada una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de una $P(\lambda)$, sabemos que la función de probabilidad en este caso es

$$P[X_i = x_i; \lambda] = e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!},$$

por lo que la función de verosimilitud se expresará como:

$$\begin{aligned} L(\lambda; x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= e^{-\lambda} \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} \cdots e^{-\lambda} \frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda} \cdot \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$

Método de máxima verosimilitud: Ejemplo 1

Entonces, tomando logaritmos

$$\ln L(\lambda; x_1, x_2, \ldots, x_n) = \left(\sum_{i=1}^{n} x_i\right) \cdot \ln \lambda - \ln\left(\prod_{i=1}^{n} x_i!\right) - n\lambda,$$

y derivando la expresión anterior

$$\frac{\partial}{\partial \lambda} \ln L(\lambda; x_1, x_2, \ldots, x_n) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n.$$

Igualando a cero la derivada anterior se obtiene la igualdad

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0,$$

cuya solución es el posible estimador puntual buscado

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{X}.$$

Método de máxima verosimilitud: Ejemplo 1

Para comprobar que es un máximo tendríamos que ver que la segunda derivada es negativa cuando se sustituye λ por el estimador obtenido.
Para obtener el valor del estimador máximo verosímil para cada una de las muestras consideradas tan sólo hay que calcular la media aritmética de cada colección de datos:

$$\widehat{\lambda}_1 = \frac{3+3+\cdots+2+5}{10} = 2'8, \quad \widehat{\lambda}_2 = \frac{1+3+\cdots+1+3}{10} = 3'1.$$

Método de máxima verosimilitud: Ejemplo 2

Una variable aleatoria con distribución gamma, con parámetro α conocido, tiene función de densidad:

$$f(x;\theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \cdot x^{\alpha-1} \cdot \exp\left\{-\frac{x}{\theta}\right\}, \quad x > 0, \quad \alpha > 0$$

con valor esperado $E[X] = \alpha\theta$ y varianza $Var(X) = \alpha\theta^2$.
Para obtener el estimador puntual para θ mediante el método de máxima verosimilitud hay que obtener en primer lugar la función de verosimilitud muestral. Es decir

$$L(\theta; x_1, x_2, \ldots, x_n) = \frac{1}{\Gamma(\alpha)^n \theta^{n\alpha}} \cdot \prod_{i=1}^{n} x_i^{\alpha-1} \cdot \exp\left\{-\frac{\sum_{i=1}^{n} x_i}{\theta}\right\}.$$

Tomando logaritmo en la verosimilitud

$$\ln L(\theta; x_1, x_2, \ldots, x_n) = -n\alpha \ln\theta - n \ln\Gamma(\alpha) + (\alpha - 1)\sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i}{\theta},$$

Método de máxima verosimilitud: Ejemplo 2

derivando con respecto a θ

$$\frac{\partial}{\partial\theta} \ln L(\theta; x_1, x_2, \ldots, x_n) = -\frac{n\alpha}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2},$$

e igualando a cero

$$\frac{-n\alpha\theta + \sum_{i=1}^{n} x_i}{\theta^2} = 0,$$

se obtiene la solución

$$\widehat{\theta}(x_1, \ldots, x_n) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\alpha} = \frac{\overline{x}}{\alpha}.$$

Para confirmar que este valor es un máximo, habría que comprobar que la derivada segunda es negativa cuando se sustituye en ella θ por la solución obtenida.

Método de los momentos

Este método se basa en la relación existente entre los parámetros poblacionales desconocidos y los momentos poblacionales. Más concretamente, consiste en trasladar esa relación sustituyendo los momentos poblacionales por los muestrales, para así estimar los primeros.
Dada una variable aleatoria, X, cuya distribución de probabilidad depende de un único parámetro desconocido, el método de los momentos consiste en enfrentar características poblacionales, $E[X^r]$, a sus homólogas en la muestra

$$\frac{1}{n}\sum_{i=1}^{n} x_i^r \cdot n_i.$$

En el caso que se estudia un único parámetro desconocido, habrá que enfrentar un único parámetro poblacional a un único parámetro muestral ($r = 1$). Es decir, $E[X] = \bar{X}$, de forma que se plantea una ecuación, cuya solución es el estimador puntual buscado.

Método de los momentos: Ejemplo 1

A continuación se van a obtener los estimadores puntuales por el método de los momentos de las distribuciones de Poisson de parámetro λ y binomial de parámetro p.
Adviértase que la igualdad a plantear es $E[X] = \bar{X}$.
En el caso de la binomial, $E[X] = np$, luego entonces $np = \bar{X}$, y por tanto, el estimador puntual buscado es

$$\widehat{p} = \frac{\bar{X}}{n}.$$

Mientras que para la Poisson, $E[X] = \lambda$, luego directamente se obtiene que $\widehat{\lambda} = \bar{X}$.
Obsérvese que los estimadores obtenidos coinciden con los del método anterior. Esto se debe a que, bajo ciertas condiciones, ambos métodos coinciden.

Método de los momentos: Ejemplo 2

Para estimar el parámetro θ de una variable aleatoria con distribución gamma, con parámetro α conocido, por el método de los momentos, simplemente habría que igualar la media poblacional a la media muestral y despejar θ. Esto es

$$E[X] = \overline{X} \Rightarrow \alpha\theta = \overline{X} \Rightarrow \theta = \frac{\overline{X}}{\alpha}.$$

Puede parecer que este método es más sencillo de aplicar que el anterior, pero nada más lejos de la realidad. Supongamos que se ignora en este caso que $E[X] = \alpha\theta$, entonces tendría que calcular dicho valor esperado y, por tanto, se empieza a complicar este método.

Propiedades deseables para un estimador

Tras estudiar métodos enfocados a la obtención de estimadores, a continuación vamos a preguntarnos por la calidad de las estimaciones que estos proporcionan. Necesitamos alguna medida que nos permita seleccionar el mejor estimador, ya que recordemos que para un mismo parámetro desconocido es posible que exista más de un estimador puntual.

A continuación se estudiará como comprobar si un estimador verifica cada una de estas propiedades.

Dada una variable aleatoria, X, cuya distribución de probabilidad depende de un parámetro desconocido, θ, se dice que el estimador $\widehat{\theta}(X_1, \ldots, X_n)$ de θ es insesgado si $E[\widehat{\theta}] = \theta$. Si no se verifica la condición anterior, se dice que no es insesgado o que es sesgado, y que tiene un sesgo $E\left[\widehat{\theta}\right] - \theta$.

Ejemplo 2 Sea X una variable aleatoria cuya distribución de probabilidad depende de un parámetro desconocido θ tal que $E[X] = \theta$ y sea $\widehat{\theta}(X_1, \ldots, X_n) = \bar{X}$ un estimador puntual de dicho parámetro.
Puesto que se ha demostrado que $E[\overline{X}] = E[X]$, entonces se verifica que $E[\widehat{\theta}] = E[\overline{X}] = \theta$, por lo que $\widehat{\theta}$ es un estimador insesgado de θ, es decir, que **la media aritmética siempre es estimador insesgado de la media poblacional**.

♯

Por tanto, en el caso en el que X se distribuya según una Normal, una Poisson o una Bernoulli ($\theta = \mu$, $\theta = \lambda$ y $\theta = p$, respectivamente), puesto que en estos casos se verifica que $\widehat{\theta}(X_1, \ldots, X_n) = \bar{X}$, entonces $\widehat{\theta}$ es un estimador puntual insesgado de θ de forma inmediata.

Ejemplo 3 Dada una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de una variable aleatoria, X, distribuida según una normal de parámetros μ y σ^2, entonces se verifica que la media poblacional es $E[X] = \mu$ y la varianza poblacional es $Var(X) = \sigma^2$.
Por tanto, en tal caso hemos demostrado que se verifica que $E[S^2_{n-1}] = \sigma^2$ y $E[S^2_n] = \frac{n-1}{n}\sigma^2$, es decir, la cuasivarianza muestral es un estimador insesgado de σ^2, mientras que la varianza muestral lo es sesgado. ♯

Ejemplo 4 Una vez obtenido el estimador para el parámetro θ de una distribución gamma con α conocido, $\widehat{\theta}(x_1, \ldots, x_n) = \frac{\overline{x}}{\alpha}$, para estudiar si es insesgado hemos de comprobar que se verifica que $E[\widehat{\theta}] = \theta$. En efecto,

$$E[\widehat{\theta}] \quad = \quad E\left[\frac{\overline{X}}{\alpha}\right] = \frac{1}{\alpha}E[\overline{X}] = \frac{1}{\alpha}E[X] = \frac{1}{\alpha}\alpha\theta = \theta.$$

♯

Eficiencia

Dada una variable aleatoria, X, cuya distribución de probabilidad depende de un parámetro desconocido, θ, y siendo $\widehat{\theta}$ un estimador puntual de θ. Los estimadores de un parámetro θ verifican siempre la siguiente desigualdad

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{n \cdot E\left[\left(\frac{\partial}{\partial\theta} \ln L(\theta; x)\right)^2\right]}.$$

donde $L(\theta; x)$ es la función de verosimilitud para una muestra de tamaño 1 (es decir, la función de probabilidad para variables discretas o la función de densidad para variables continuas)

Un **estimador es eficiente** si es insesgado[1] y se verifica la igualdad en la expresión anterior.

La expresión

$$\frac{1}{n \cdot E\left[\left(\frac{\partial}{\partial\theta} \ln L(\theta; x)\right)^2\right]},$$

recibe el nombre de cota de Frechet-Cramer-Rao.

[1] Por tanto, los estimadores sesgados no son eficientes.

Eficiencia: Ejemplo 1

Para estudiar si el estimador máximo verosímil del parámetro θ de una distribución gamma con α conocido es eficiente hay que comprobar que $Var\left(\widehat{\theta}\right)$ coincide con la cota de Frechet-Cramer-Rao.

Calculamos la varianza del estimador:

$$Var\left(\widehat{\theta}\right) = Var\left(\frac{\overline{X}}{\alpha}\right) = \frac{1}{\alpha^2} Var(\overline{X}) = \frac{1}{\alpha^2} \cdot \frac{Var(X)}{n} = \frac{1}{\alpha^2} \cdot \frac{\alpha\theta^2}{n} = \frac{\theta^2}{\alpha n}.$$

Calculamos la cota de Frechet-Cramer-Rao:

- $L(\theta; x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} \cdot \exp\left\{-\frac{x}{\theta}\right\}$
- $\ln L(\theta; x) = -\ln(\Gamma(\alpha)) - \alpha \ln\theta + \ln x^{\alpha-1} - \frac{x}{\theta}$
- $\frac{\partial}{\partial\theta} \ln L(\theta; x) = -\frac{\alpha}{\theta} + \frac{x}{\theta^2} = \frac{x-\alpha\theta}{\theta^2}$

Eficiencia: Ejemplo 1

- $E\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta;x)\right)^2\right] = E\left[\left(\frac{X-\alpha\theta}{\theta^2}\right)^2\right] = \frac{1}{\theta^4}E\left[(X-\alpha\theta)^2\right] =$

  $= \frac{1}{\theta^4}E\left[(X-E[X])^2\right] = \frac{1}{\theta^4}Var(X) = \frac{\alpha\theta^2}{\theta^4} = \frac{\alpha}{\theta^2}$

- $\frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta;x)\right)^2\right]} = \frac{\theta^2}{\alpha n}$

La cota coincide con la varianza del estimador, luego el estimador es eficiente.

Eficiencia: Ejemplo 2

Estudiemos a continuación si el estimador puntual $\widehat{\lambda}(X_1,\ldots,X_n) = \bar{X}$ del parámetro desconocido λ de una variable aleatoria discreta, X, distribuida según una Poisson, que sabemos que es insesgado, es también eficiente.
En este caso, se tiene que la función de probabilidad es:

$$P[X=x;\lambda] = e^{-\lambda}\cdot\frac{\lambda^x}{x!}, \quad x=0,1,2,\ldots, \quad \lambda>0,$$

con $E[X] = \lambda = Var(X)$.
Con dicha información, calculamos la varianza del estimador:

$$Var\left(\widehat{\lambda}\right) = Var\left(\overline{X}\right) = \frac{Var\,(X)}{n} = \frac{\lambda}{n}.$$

- $L(\lambda;x) = e^{-\lambda}\cdot\frac{\lambda^x}{x!}$

- $\ln L(\lambda;x) = \ln\left(e^{-\lambda}\cdot\frac{\lambda^x}{x!}\right) = \ln e^{-\lambda} + \ln\frac{\lambda^x}{x!}$

  $= -\lambda\ln e + \ln\lambda^x - \ln x! = -\lambda + x\ln\lambda - \ln x!$

Eficiencia: Ejemplo 2

- $\frac{\partial}{\partial\lambda}\ln L(\lambda;x) = -1 + \frac{x}{\lambda} = \frac{x-\lambda}{\lambda}$
- $E\left[\left(\frac{\partial}{\partial\lambda}\ln L(\lambda;x)\right)^2\right] = E\left[\left(\frac{X-\lambda}{\lambda}\right)^2\right] = \frac{E[(X-E[X])^2]}{\lambda^2} = \frac{Var(X)}{\lambda^2} = \frac{1}{\lambda}$
- $\frac{1}{nE\left[\left(\frac{\partial}{\partial p}\ln L(\lambda;x)\right)^2\right]} = \frac{\lambda}{n}.$

El estimador $\widehat{\lambda}$ es eficiente ya que

$$Var\left(\widehat{\lambda}\right) = \frac{\lambda}{n} = \frac{1}{n\cdot E\left[\left(\frac{\partial}{\partial\theta}\ln L(\lambda;x)\right)^2\right]}.$$

Consistencia

Dada una variable aleatoria, X, cuya distribución de probabilidad depende de un parámetro desconocido, θ, y sea $\widehat{\theta}(X_1,\ldots,X_n)$ un estimador puntual de dicho parámetro. Se dice que dicho estimador es consistente cuando

$$\lim_{n\to\infty} P\left[|\widehat{\theta}(X_1,\ldots,X_n)-\theta| > \epsilon\right] = 0, \qquad \forall\epsilon > 0.$$

Puesto que la condición dada anteriormente no siempre es fácil de comprobar, en la práctica se trabajará con la siguiente condición suficiente:

- El estimador es asintóticamente insesgado: $\lim_{n\to\infty} E[\widehat{\theta}] = \theta$.
- El estimador tiene asintóticamente varianza cero: $\lim_{n\to\infty} Var[\widehat{\theta}] = 0$.

Consistencia: Ejemplo 1

Dada una variable aleatoria continua con función de densidad

$$f(x,\lambda) = \frac{2}{\lambda^2}(\lambda - x), \qquad 0 < x < \lambda,$$

obtener el estimador del parámetro λ por el método de los momentos y comprobar que es consistente.
Usando el método de los momentos, el estimador para una muestra de tamaño n saldrá de la igualdad que se obtiene al igualar el momento poblacional de orden uno con su homólogo en la muestra

$$E[X] = \overline{x},$$

donde

$$E[X] = \int_0^\lambda x \cdot f(x)dx = \cdots = \frac{\lambda}{3}.$$

Por tanto,

$$\frac{\lambda}{3} = \overline{x} \Longrightarrow \widehat{\lambda} = 3\overline{x}.$$

Consistencia: Ejemplo 1

Para estudiar la consistencia del estimador obtenido veremos si es asintóticamente insesgado y con varianza cero en el límite:

- $E[\widehat{\lambda}] = 3E[\overline{X}] = 3E[X] = 3\frac{\lambda}{3} = \lambda \Longrightarrow \lim_{n\to\infty} E[\widehat{\lambda}] = \lim_{n\to\infty} \lambda = \lambda,$
- $Var(\widehat{\lambda}) = 9Var(\overline{X}) = \frac{9}{n}Var(X) = \frac{9}{n} \cdot \frac{\lambda^2}{18} = \frac{\lambda^2}{2n}$

$$\Longrightarrow \lim_{n\to\infty} Var(\widehat{\lambda}) = \lim_{n\to\infty} \frac{\lambda^2}{2n} = 0,$$

donde se ha usado que

$$E[X^2] = \int_0^\lambda x^2 \cdot f(x)dx = \cdots = \frac{\lambda^2}{6},$$

con lo que

$$Var(X) = E[X^2] - (E[X])^2 = \frac{\lambda^2}{6} - \frac{\lambda^2}{9} = \frac{\lambda^2}{18}.$$

Luego se ha demostrado que el estimador obtenido es consistente.

Técnicas Cuantitativas II

Tema 3. Distribuciones de los estadísticos muestrales

Distribuciones de los estadísticos muestrales

Como se ha visto en el tema anterior, los estadísticos muestrales se pueden utilizar para la estimación puntual de los correspondientes parámetros poblacionales. Pero además de la estimación puntual, existen otros métodos de estimación de parámetros poblacionales como son los intervalos de confianza y los contrastes de hipótesis (que se estudiaran en temas posteriores).

Para el estudio de estos métodos será fundamental tener en cuenta el carácter aleatorio de los estadísticos muestrales y conocer su distribución.

Destacar que este tema se centra en estadísticos muestrales cuyas distribuciones de probabilidad son obtenidas a partir de poblaciones con distribución normal. Esta característica marcará también los siguientes temas de estimación paramétrica mediante intervalos de confianza y contraste de hipótesis.

Distribución para la cuasivarianza muestral

Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de tamaño n procedente de una población $N(\mu, \sigma)$. Se verifica entonces que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

A partir de ahora por S^2 **denotaremos a la cuasivarianza muestral**.

Ejemplo 1 Si suponemos que la variable altura de los jugadores de una determinado equipo de baloncesto sigue una distribución normal, calcular la probabilidad de que en una muestra de 4 jugadores seleccionados al azar, la cuasi-desviación típica muestral supere el valor de 0.42 cm si la desviación típica poblacional es 1.8 cm.

$$\begin{aligned} P\left[S > 0.42\right] &= P\left[S^2 > (0.42)^2\right] = P\left[S^2 > 0.1764\right] \\ &= P\left[\frac{(n-1)}{(\sigma)^2} \cdot S^2 > \frac{3 \cdot 0.1764}{(1.8)^2}\right] = P\left[\chi^2_3 > 0.1633\right] \\ &= 0.9833 \qquad [pchiq(0.1633, 3, lower.tail = F)] \end{aligned}$$

♯

Distribución para la media muestral

Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de tamaño n procedente de una población $N(\mu, \sigma)$, donde σ es desconocida. Se verifica entonces que

$$\frac{\left(\overline{X} - \mu\right)\sqrt{n}}{S} \sim t_{n-1}.$$

Ejemplo 2 Si consideramos que el beneficio semanal de un determinado comercio es una variable aleatoria cuya distribución es aproximadamente Normal de media 1500 euros, calcular la probabilidad de que una muestra de 5 semanas de lugar a una media muestral superior a 1525 sabiendo que la cuasi-desviación típica muestral es 36.
Se tiene que

$$t = \frac{\left(\overline{X} - 1500\right)\sqrt{5}}{36} \sim t_4,$$

de forma que

$$\begin{aligned} P\left[\overline{X} > 1525\right] &= P\left[t_4 > \frac{(1525 - 1500)\sqrt{5}}{36}\right] \\ &= P\left[t_4 > 1.5528\right] = 0.0977 \qquad [pt(1.5528, 4, lower.tail = F)], \end{aligned}$$

♯

Distribución para la proporción muestral

Sea $X_1, \ldots, X_n$ una muestra aleatoria simple procedente de una variable aleatoria distribuida según una Bernouilli de parámetro p, entonces la variable aleatoria dada por la proporción muestral

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} A_i,$$

tiene distribución aproximadamente Normal

$$\widehat{p} \sim N\left(p, \sigma^2 = \frac{p(1-p)}{n}\right),$$

si el tamaño muestral es suficientemente elevado y donde

$$A_i = \begin{cases} 1 & \text{, si el individuo } i \text{ presenta la característica} \\ & \text{en estudio con probabilidad } p \\ 0 & \text{, en otro caso} \end{cases} .$$

Adviértase, que a efectos prácticos, para el análisis mediante intervalos de confianza y contrastes de hipótesis se suele admitir que

$$\widehat{p} \sim N\left(p, \sigma^2 = \frac{\widehat{p}(1-\widehat{p})}{n}\right),$$

Distribución para la proporción muestral

Distribución para la varianza muestral
Distribución para la media muestral
Distribución para la proporción muestral
Distribución para el cociente de varianzas
Distribucion para la diferencia de medias muestrales
Distribución para la diferencia de proporciones

Ejemplo 3 Si el 20% de las empresas están dirigidas por mujeres, calcular, para una muestra aleatoria de tamaño 300, la probabilidad de que el porcentaje de empresas dirigidas por mujeres sea inferior al 15%.
Estamos ante un experimento de Bernouilli donde el éxito, $p = 0.2$, consiste en que una empresa esté dirigida por una mujer. Luego sabemos que

$$\widehat{p} \sim N\left(0.2, \sigma^2 = \frac{0.2 \cdot 0.8}{300}\right) \equiv N(0.2, \sigma^2 = 0.00053).$$

.

Entonces: $P[\widehat{p} \leq 0.15] = 0.0149$

$[pnorm(0.15, mean = 0.2, sd = sqrt(0.00053), lower.tail = T)]$ ♯

Distribución para el cociente de varianzas

Dadas $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos muestras aleatorias simples independientes procedentes de sendas poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, se verifica entonces que

$$F = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,m-1}$$

Ejemplo 4 Se sabe que la calificación media de la asignatura Métodos Cuantitativos durante el curso 2007/2008 fue de 4.3 y la desviación típica 1.1. Durante el curso 2008/2009 se llevaron a cabo ciertos cambios metodológicos y la calificación media aumentó a 5.6 siendo la desviación típica de 0.98. Si se seleccionan 20 alumnos al azar del curso 2007/2008 y otros 25 alumnos del curso 2008/2009, calcule la probabilidad de que la primera muestra tenga una cuasivarianza muestral superior al triple de la segunda. ♯

Distribución para el cociente de varianzas

Distribución para la varianza muestral
Distribución para la media muestral
Distribución para la proporción muestral
Distribución para el cociente de varianzas
Distribucion para la diferencia de medias muestrales
Distribución para la diferencia de proporciones

En este caso partimos de las distribuciones

$$X \sim N(4.3, 1.1^2), Y \sim N(5.6, 0.98^2)$$

y además sabemos que

$$F = \frac{S_1^2}{S_2^2} \cdot \frac{0.98^2}{1.1^2} \sim F_{19,24}.$$

Entonces la probabilidad deseada se obtiene como sigue

$$P[S_1^2 > 3 \cdot S_2^2] = P\left[\frac{S_1^2}{S_2^2} > 3\right] = P\left[\left(\frac{0.98}{1.1}\right)^2 \cdot \frac{S_1^2}{S_2^2} > 3 \cdot \left(\frac{0.98}{1.1}\right)^2\right]$$

$$= P[F_{19,24} > 2.381] = 0.0231 \qquad [pf(2.381, 19, 24, lower.tail = F)]$$

Distribucion para la diferencia de medias muestrales: Varianzas poblacionales desconocidas e iguales

Dadas $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos muestras aleatorias simples independientes procedentes de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, siendo σ_1 y σ_2 desconocidas e iguales $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$. Entonces,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{m+n}{mn}}} \sim t_{n+m-2},$$

donde

$$S_p^2 = \frac{(n-1)\cdot S_1^2 + (m-1)\cdot S_2^2}{n+m-2}.$$

Ejemplo 5 El tiempo, medido en segundos, que se tarda en realizar cierta tarea administrativa por un empleado con más de dos años de experiencia sigue una distribución normal de media 200 segundos. Esa misma tarea llevada a cabo por un empleado recién contratado tiene un tiempo medio de realización de 215 segundos. Tomando 15 empleados al azar con más de dos años de experiencia y 5 empleados recién contratados, y sabiendo que las cuasivarianzas muestrales son 180 y 250 respectivamente, se pide calcular la probabilidad de que el tiempo medio de realización sea más de 10 segundos superior en el caso de los empleados recién contratados, sabiendo que ambas varianzas poblacionales son iguales. ♯

Distribucion para la diferencia de medias muestrales: Varianzas poblacionales desconocidas e iguales

Si definimos $\bar{X}$ como *tiempo medio que tardan los empleados con más de dos años de experiencia* y $\bar{Y}$ como *tiempo medio que tardan los empleados recién contratados*. La probabilidad que se desea calcular es $P[\bar{Y} > \bar{X} + 10]$, es decir, $P[\overline{X} - \overline{Y} \leq -10]$. Suponiendo que el tiempo de realización de la tarea es una variable normal y que las desviaciones típicas son desconocidas pero iguales, usaremos el estadístico:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{20}{75}}} \sim t_{18},$$

donde

$$S_p = \sqrt{\frac{(n-1)\cdot S_1^2 + (m-1)\cdot S_2^2}{n+m-2}} = \sqrt{\frac{14\cdot 180 + 4\cdot 250}{18}} = 13.984,$$

luego

Distribucion para la diferencia de medias muestrales: Varianzas poblacionales desconocidas e iguales

$$
\begin{aligned}
P[\overline{X}-\overline{Y}\leq -10] &= P\left[\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{m+n}{mn}}}\leq \frac{-10-(\mu_1-\mu_2)}{S_p\sqrt{\frac{m+n}{mn}}}\right] \\
&= P\left[t_{18}<\frac{-10-(200-215)}{13.984\sqrt{\frac{1}{15}+\frac{1}{5}}}\right] \\
&= P\left[t_{18}<\frac{5}{13.984\cdot 0.516}\right]=P[t_{18}<0.6929]=0.7514,
\end{aligned}
$$

$$[pt(0.6929, 18, lower.tail = T)]$$

Distribucion para la diferencia de medias muestrales: Varianzas poblacionales desconocidas pero diferentes

Dadas $X_1,\ldots,X_n$ e $Y_1,\ldots,Y_m$ dos muestras aleatorias simples independientes procedentes de dos poblaciones normales $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$, siendo σ_1 y σ_2 desconocidas y diferentes. Entonces,

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}}\sim t_v$$

donde

$$v=\frac{\left(\frac{S_1^2}{n}+\frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n-1}+\frac{\left(\frac{S_2^2}{m}\right)^2}{m-1}}$$

Cuando los tamaños muetrales son grandes, se puede demostrar que:

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}}\sim N(0,1)$$

Distribución para la diferencia de proporciones

Dadas $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos muestras aleatorias simples independientes con tamaños n y m, procedentes de variables aleatorias de Bernouilli, con parámetros p_1 y p_2. Se verifica entonces que la variable aleatoria

$$\frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim N(0,1).$$

Adviértase, que a efectos prácticos, para el análisis mediante intervalos de confianza y contrastes de hipótesis se suele admitir que

$$\frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} \sim N(0,1).$$

Distribución para la diferencia de proporciones

Distribución para la varianza muestral
Distribución para la media muestral
Distribución para la proporción muestral
Distribución para el cociente de varianzas
Distribucion para la diferencia de medias muestrales
Distribución para la diferencia de proporciones

Ejemplo 6 Según los resultados publicados en el informe del 2008 realizado por Infoadex, el 47'6% de la inversión publicitaria en el 2008 fue destinada a medios convencionales mientras que en el 2007 se destinó el 49'5%. Si se selecciona una muestra de 10 empresas en cada uno de los años, ¿cuál es la probabilidad de que la proporción muestral de inversión en medios convencionales en el 2007 sea inferior a la del 2008?
Siendo $\widehat{p}_1$ la proporción de inversión publicitaria en el año 2007 y $\widehat{p}_2$ en el 2008 y usando que $Z = \frac{\widehat{p}_1 - \widehat{p}_2 - (0.495 - 0.476)}{\sqrt{\frac{0.495(1-0.495)}{10} + \frac{0.476(1-0.476)}{10}}} \sim N(0,1)$, se tiene

$$P[\widehat{p}_1 < \widehat{p}_2] = P[\widehat{p}_1 - \widehat{p}_2 < 0] = P\left[Z < -\frac{0.019}{0.223}\right] = P[Z < -0.086] = 0.4657$$

$$[pnorm(0.086, mean = 0, sd = 1, lower.tail = T)]$$

♯

Técnicas Cuantitativas II

Tema 4. Estimación de parámetros mediante intervalos de confianza

Estimación mediante intervalos de confianza

Como ya hemos visto, la estimación puntual aproxima mediante un número el valor de una característica poblacional o parámetro desconocido (la calificación media de una asignatura, proporción de alumnos con coche,...) pero no nos indica el error que se comete en dicha estimación.
En la práctica, además de realizar la estimación puntual de un parámetro, lo razonable sería obtener una herramienta que mida el margen de error de esa estimación puntual. Esa herramienta son los intervalos de confianza.

Un **intervalo de confianza** para un parámetro con un **nivel de confianza** $1-\alpha$ donde $(0<\alpha<1)$ es un conjunto de valores entre los que esperamos que esté el verdadero valor del parámetro con una confianza de $1-\alpha$:

$$P[\text{parámetro } \in (a,b)] = 1-\alpha$$

$$P[a \leq \text{ parámetro } \leq b] = 1-\alpha$$

Los extremos del intervalo de confianza se construirán a partir de datos muestrales y, como veremos, también dependerán del nivel de confianza.
Los valores más habituales para el nivel de confianza $1-\alpha$ son 0.9, 0.95 y 0.99. A α se le denomina **nivel de significación**.

Estimación mediante intervalos de confianza

Al igual que los estimadores puntuales eran variables aleatorias (tomaban diferentes valores con distintas probabilidades, según la muestra recogida), para cada muestra diferente, los intervalos de confianza también serán diferentes.

Los intervalos de confianza se interpretan se la siguiente manera:
Un intervalo de confianza al 95% garantiza que, si tomamos 100 muestras y construimos 100 intervalos, el verdadero valor del parámetro estará dentro del intervalo en aproximadamente el 95 de los intervalos construidos.

No es correcto decir "la probabilidad de que el parámetro pertenezca al intervalo (a, b) es $1 - \alpha$ " porque el parámetro NO es una variable aleatoria. El intervalo es aleatorio ya que sus extremos son funciones de la muestra y por lo tanto, debemos decir "**la probabilidad de que el intervalo (a, b) contenga al parámetro es $1 - \alpha$** " .

Una vez construído el intervalo a partir de una muestra dada, ya no tiene sentido hablar de probabilidad. En todo caso, tenemos "confianza" de que el intervalo contenga al parámetro. La confianza recae en el método de construcción de los intervalos, que nos asegura que $(1 - \alpha)100\%$ de las muestras producirán intervalos que contienen al parámetro.

Intervalo de confianza para la varianza poblacional

Sea X una variable aleatoria con distribución $N(\mu, \sigma^2)$. Se toma una muestra aleatoria simple de dicha población: $X_1, \ldots, X_n$.
El intervalo de confianza para la varianza poblacional (σ^2) se construye de la siguiente manera (Método pivote):

- Seleccionar un estadístico que contenga al parametro que se desea estimar (σ^2), cuya distribución sea conocida y no dependa de dicho parámetro:

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

- Como la distribución esta variable aleatoria es independiente del valor del parámetro, se pueden encontrar dos cuantiles que definen un intervalo que contendrá a esa variable con probabilidad $1 - \alpha$:

$$P\left[\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2} S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

  donde

  - $\chi^2_{n-1,p}$ es el valor de una χ^2 con $n-1$ grados de libertad que deja a la izquierda (por debajo) una probabilidad de p

Intervalo de confianza para la varianza poblacional

- Despejar el parámetro desconocido en la desigualdad anterior:

$$P\left[\frac{\chi^2_{n-1,\frac{\alpha}{2}}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1,1-\frac{\alpha}{2}}}{(n-1)S^2}\right] = 1-\alpha$$

$$\boxed{P\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = 1-\alpha}$$

Ejemplo 1 El gasto mensual (en euros) en clases particulares de los alumnos de la Facultad de Ciencias Económicas y Empresariales sigue una distribución normal. Se toma una muestra aleatoria de ocho alumnos y se obtienen los siguientes datos: 90, 87, 95, 105, 101, 100, 99, 97.
Se pide obtener un intervalo de confianza al 99% para la varianza poblacional. ♯

Intervalo de confianza para la varianza poblacional

- $\overline{X} = 96.75$
- $S^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - \overline{X})^2 = 245.5$
- $\chi^2_{8-1,1-\frac{0.001}{2}} = \chi^2_{7,0.995} = 20.2777$
- $\chi^2_{8-1,\frac{0.001}{2}} = \chi^2_{7,0.005} = 0.9893$

Por lo que el intervalo que contiene a la varianza poblacional con una confianza del 99% es $(12.11,\ 248.17)$

Script de R:

```
gasto=c(90, 87, 95, 105, 101, 100, 99, 97)
alpha=0.01
n = 8
media_muestral=mean(gasto)
cuasi_varianza=var(gasto)
cuantil_inf= qchisq(1-(alpha/2),n-1,lower.tail = T)
cuantil_sup= qchisq((alpha/2),n-1,lower.tail = T)
lim_inf=(n-1)*cuasi_varianza/cuantil_inf
lim_sup=(n-1)*cuasi_varianza/cuantil_sup
```

Intervalo de confianza para la media poblacional

Sea X una variable aleatoria con distribución $N(\mu, \sigma^2)$. Se toma una muestra aleatoria simple de dicha población: $X_1, \ldots, X_n$. El intervalo de confianza para la media poblacional (μ) se construye utilizando el método pivote:

- Seleccionar un estadístico que contenga al parametro que se desea estimar (μ), cuya distribución sea conocida y no dependa de dicho parámetro:

$$\frac{\left(\overline{X} - \mu\right)\sqrt{n}}{S} \sim t_{n-1}$$

- Como la distribución esta variable aleatoria es independiente del valor del parámetro, se pueden encontrar dos cuantiles que definen un intervalo que contendrá a esa variable con probabilidad $1 - \alpha$:

$$P\left[t_{n-1,\frac{\alpha}{2}} \leq \frac{\left(\overline{X} - \mu\right)\sqrt{n}}{S} \leq t_{n-1,1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

donde

  - $t_{n-1,p}$ es el valor de una t-student con $n - 1$ grados de libertad que deja a la izquierda (por debajo) una probabilidad de p

Intervalo de confianza para la media poblacional

- Despejar el parámetro desconocido en la desigualdad anterior:

$$P\left[t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \overline{X} - \mu \leq t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[-\overline{X} + t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq -\mu \leq -\overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

Y como la distribución t-student es simétrica:

$$\boxed{P\left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] = 1 - \alpha}$$

Ejemplo 2 Supongamos que la variable *edad de los funcionarios españoles* es una variable aleatoria con distribución Normal. Construir un intervalo de confianza al 95% de confianza para la media poblacional sabiendo que se ha recogido una muestra de 10 funcionarios obteniéndose los siguientes datos: 38, 38, 50, 40, 40, 43, 58, 35, 44, 40. ♯

Intervalo de confianza para la media poblacional

- $\overline{X} = 42.6$
- $S^2 = \frac{1}{10-1}\sum_{i=1}^{10}(x_i - \overline{X})^2 = 46.04$
- $t_{10-1,1-\frac{0.05}{2}} = t_{9,0.975} = 2.262$

Por lo que el intervalo que contiene a la media poblacional con una confianza del 95% es

$$P\left[42.6 - 2.262 \cdot \frac{\sqrt{46.04}}{\sqrt{10}} < \mu < 42.6 + 2.262 \cdot \frac{\sqrt{46.04}}{\sqrt{10}}\right] = 0'95$$

$$P[37.75 \leq \mu \leq 47.45] = 0.95.$$

Script de R:

```
edad=c(38, 38, 50, 40,40,43, 58, 35, 44, 40)
alpha=0.05
n = 10
media_muestral=mean(edad)
cuasi_varianza=var(edad)
cuantil= qt(1-(alpha/2),n-1,lower.tail = T)
lim_inf=media_muestral-cuantil*sqrt(cuasi_varianza)/sqrt(n)
lim_sup=media_muestral+cuantil*sqrt(cuasi_varianza)/sqrt(n)
```


Intervalo de confianza para la proporción

Dada $X_1, \ldots, X_n$ una muestra aleatoria procedente de una distribución de Bernoulli con probabilidad de obtener exito p (proporción).
El intervalo de confianza para la proporción poblacional (p) se construye de la siguiente manera:

- Seleccionar un estadístico que contenga al parámetro que se desea estimar cuya distribución sea conocida y no dependa de dicho parámetro:

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1) (n \text{ es suficientemente grande})$$

- Encontrar dos cuantiles que definan un intervalo que contenga a esa variable con probabilidad $1 - \alpha$:

$$P\left[-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha.$$

donde $z_{1-\frac{\alpha}{2}}$ es el valor de una $N(0,1)$ que deja a la izquierda una probabilidad de $1 - \frac{\alpha}{2}$

Intervalo de confianza para la proporción

- Despejar el parámetro desconocido:

$$P\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1-\alpha.$$

Ejemplo 3 Según la encuesta publicada por "The Washington Post" el 5 de noviembre de 2008, de los 17834 encuestados, el 47.3% afirmaba no haber votado por el candidato Barack Obama. Se pide construir un intervalo de confianza al 99% para la verdadera proporción de votantes que lo hicieron por Obama.

- $\hat{p} = 1 - 0.473 = 0.527$
- $n = 17834$
- $z_{1-\frac{0.01}{2}} = z_{0.995} = 2.575$

Por lo que el intervalo que contiene a la proporción poblacional con una confianza del 99% es $(0.5173,\ 0.5366)$.
Si para ganar el candidato necesita contar con el 51% de votantes, ¿hay alguna evidencia para creer en su éxito?
Atendiendo al intervalo de confianza obtenido, hay fuerte evidencia a favor del triunfo del candidato puesto que el intervalo al 99% está por encima de 0.51. ♯

Intervalo de confianza para la proporción

Script de R:

```
alpha=0.01
n = 17834
p_muestral=1-0.473
cuantil= qnorm(1-(alpha/2),mean=0, sd=1,lower.tail = T)
lim_inf=p_muestral-cuantil*sqrt(p_muestral*(1-p_muestral)/n)
lim_sup=p_muestral+cuantil*sqrt(p_muestral*(1-p_muestral)/n)
```

Intervalo de confianza para el cociente de varianzas

Dadas $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, dos muestras aleatorias simples independientes procedentes de dos poblaciones normales, $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$.

El intervalo de confianza para el cociente de varianzas poblacionales $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$ se construye de la siguiente manera:

- $\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,m-1}$
- $P\left[F_{\frac{\alpha}{2}} \leq \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\frac{\alpha}{2}}\right] = 1-\alpha$

  donde $F_{n-1,m-1;p}$ es el valor de una F-Snedecor con $n-1$ y $m-1$ grados de libertad que deja a la izquierda una probabilidad p
- $\boxed{P\left[F_{\frac{\alpha}{2}} \cdot \frac{S_2^2}{S_1^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\frac{\alpha}{2}} \cdot \frac{S_2^2}{S_1^2}\right] = 1-\alpha}$

Intervalo de confianza para el cociente de varianzas

IC para la varianza poblacional
IC para la media poblacional
IC para la proporción
IC para el cociente de varianzas
IC para diferencia de medias
IC para la diferencia de proporciones

Ejemplo 4 Supongamos que las notas de Métodos Cuantitativos siguen una distribución normal en los grupos existentes. Se selecciona una muestra aleatoria simple de 31 alumnos matriculados en los grupos de la mañana y otra de 46 alumnos de los grupos de la tarde, ambas independientes, y se obtienen como cuasivarianzas muestrales 49 y 36, respectivamente. Se pide obtener un intervalo de confianza para el cociente de varianzas poblacionales al nivel de confianza del 90%.

A partir de los datos muestrales del enunciado se tiene que

$$n = 31, \quad S_1^2 = 49, \qquad m = 46, \quad S_2^2 = 36.$$

Por otro lado, para $\alpha = 0.1$, $F_{31-1,46-1;0.05} = 0'563$ y $F_{31-1,46-1;0.95} = 1.713$.

El intervalo de confianza que contiene al cociente $\frac{\sigma_2^2}{\sigma_1^2}$ con una confianza del 90% es $(0.41,\ 1.26)$ ♯

Intervalo de confianza para el cociente de varianzas

Script de R:

```
alpha=0.1
n = 31
Cuasi_varianza_1=49
m=46
Cuasi_varianza_2=36
cuantil_inf= qf(alpha/2,n-1,m-1 ,lower.tail = T)
cuantil_sup= qf(1-(alpha/2),n-1,m-1 ,lower.tail = T)
lim_inf=cuantil_inf*Cuasi_varianza_2/Cuasi_varianza_1
lim_sup=cuantil_sup*Cuasi_varianza_2/Cuasi_varianza_1
```


Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas pero iguales

Dadas dos muestras aleatorias simples independientes, $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, procedentes de dos poblaciones normales, $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, con varianzas desconocidas e iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

El intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$ se construye de la siguiente manera:

- $\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{m+n}{nm}}} \sim t_{n+m-2}$ donde $S_p = \sqrt{\frac{(n-1)\cdot S_1^2+(m-1)\cdot S_2^2}{n+m-2}}$.
- $P\left[-t_{n+m-2;1-\frac{\alpha}{2}} \leq \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{m+n}{nm}}} \leq t_{n+m-2;1-\frac{\alpha}{2}}\right] = 1-\alpha$ donde $t_{n+m-2;1-\frac{\alpha}{2}}$ es el cuantil de una t-Student con $n+m-2$ grados de libertad que deja a la izquierda una probabilidad de $1-\frac{\alpha}{2}$
- El intervalo de confianza para $\mu_1 - \mu_2$ al nivel de confianza $1-\alpha$ es:

$$\left[(\overline{X}-\overline{Y}) - t_{1-\frac{\alpha}{2}} S_p\sqrt{\frac{m+n}{nm}}, (\overline{X}-\overline{Y}) + t_{1-\frac{\alpha}{2}} S_p\sqrt{\frac{m+n}{nm}}\right]$$

Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas pero iguales

Ejemplo 5 A partir de la encuesta de Infoadex, la inversión publicitaria en medios convencionales, en millones de euros, en una muestra de 6 empresas del sector automovilístico ha sido: 86.6, 69.7, 45.1, 44, 42, 40.1. Mientras que tomando una muestra de tres empresas dedicadas a la comunicación se ha obtenido: 173.8, 87.5, 58.4. Suponiendo normalidad y que las varianzas poblacionales son iguales, ¿se puede considerar que coinciden las medias poblacionales de ambas variables?
A partir del enunciado se puede obtener:

- $n = 6, \overline{X} = 54.58, S_1^2 = 364.73$
- $m = 3, \overline{Y} = 106.57, S_2^2 = 3601.94$
- $S_p = \sqrt{\frac{(6-1)\cdot 364.75+(3-1)\cdot 3601.94}{6+3-2}} = 35.91$
- $t_{7;0.975} = 2.36$

Sustituyendo en el intervalo:

$$(54.58 - 106.56) \pm 2.36 \cdot 35.91\sqrt{\frac{6+3}{6\cdot 3}} = [-112.03, 8.06].$$

Como el intervalo de confianza construido contiene el 0, se puede afirmar, con un nivel de confianza del 95%, que la inversión publicitaria media de ambos sectores puede ser igual.
♯

Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas pero iguales

Script de R

```
inversion_1=c(86.6, 69.7, 45.1, 44, 42, 40.1)
inversion_2=c(173.8, 87.5, 58.4)
alpha=0.05
n = 6
media_muestral_1=mean(inversion_1)
Cuasi_varianza_1=var(inversion_1)
m=3
media_muestral_2=mean(inversion_2)
Cuasi_varianza_2=var(inversion_2)
S_p=sqrt(((n-1)*Cuasi_varianza_1+(m-1)*Cuasi_varianza_2)/(n+m-2))
cuantil= qt(1-(alpha/2),n+m-2,lower.tail = T)
lim_inf=(media_muestral_1-media_muestral_2)-
         cuantil*S_p*sqrt((m+n)/(n*m))
lim_sup=(media_muestral_1-media_muestral_2)+
         cuantil*S_p*sqrt((m+n)/(n*m))
```

Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas y diferentes

Dadas dos muestras aleatorias simples independientes, $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, procedentes de dos poblaciones normales, $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, con varianzas desconocidas y diferentes ($\sigma_1^2 \neq \sigma_2^2$).

El intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$ se construye de la siguiente manera:

- $\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \sim t_v$ donde $v = \frac{\left(\frac{S_1^2}{n}+\frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n-1}+\frac{\left(\frac{S_2^2}{m}\right)^2}{m-1}}$
- $P\left[-t_{v;1-\frac{\alpha}{2}} \leq \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \leq t_{v;1-\frac{\alpha}{2}}\right] = 1-\alpha$ donde $t_{v;1-\frac{\alpha}{2}}$ es el cuantil de una t-Student con v grados de libertad que deja a la izquierda una probabilidad de $1-\frac{\alpha}{2}$
- El intervalo de confianza para $\mu_1 - \mu_2$ al nivel de confianza $1-\alpha$ es:

$$\left[(\overline{X}-\overline{Y}) - t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}, (\overline{X}-\overline{Y}) + t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}\right]$$

Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas y diferentes

Ejemplo 6 A partir de la encuesta de Infoadex, la inversión publicitaria en medios convencionales, en millones de euros, en una muestra de 6 empresas del sector automovilístico ha sido: 86.6, 69.7, 45.1, 44, 42, 40.1. Mientras que tomando una muestra de tres empresas dedicadas a la comunicación se ha obtenido: 173.8, 87.5, 58.4. Suponiendo normalidad y que las varianzas poblacionales son **diferentes**, ¿se puede considerar que coinciden las medias poblacionales de ambas variables?

A partir del enunciado se puede obtener:

- $n = 6, \overline{X} = 54.58, S_1^2 = 364.73$
- $m = 3, \overline{Y} = 106.57, S_2^2 = 3601.94$
- $v \simeq 2$
- $t_{2;0.975} = 3.94$

El intervalo que contiene a la diferencia de medias poblacionales con una probabilidad de 0.95 es $(-191.94,\ 87.97)$

Como el intervalo de confianza construido contiene el 0, se puede afirmar, con un nivel de confianza del 95%, que la inversión publicitaria media de ambos sectores puede ser igual.

♯

Intervalo de confianza para la diferencia de medias con varianzas poblacionales desconocidas y diferentes

Script de R:

```
inversion_1=c(86.6, 69.7, 45.1, 44, 42, 40.1)
inversion_2=c(173.8, 87.5, 58.4)
alpha=0.05
n = 6
media_muestral_1=mean(inversion_1)
Cuasi_varianza_1=var(inversion_1)
m=3
media_muestral_2=mean(inversion_2)
Cuasi_varianza_2=var(inversion_2)
v=((Cuasi_varianza_1/n)+ (Cuasi_varianza_2/m))^2/
  ((Cuasi_varianza_1/n)^2/(n-1)+(Cuasi_varianza_2/m)^2/(m-1))
cuantil= qt(1-(alpha/2),v,lower.tail = T)
lim_inf=(media_muestral_1-media_muestral_2)-
        cuantil*sqrt((Cuasi_varianza_1/n)+(Cuasi_varianza_2/m))
lim_sup=(media_muestral_1-media_muestral_2)+
        cuantil*sqrt((Cuasi_varianza_1/n)+(Cuasi_varianza_2/m))
```


Intervalo de confianza para la diferencia de medias

Ejemplo 7 Se desea comparar el tiempo que se necesita para completar la inscripción en dos aplicaciones informáticas de búsqueda de empleo que se piensan lanzar al mercado. Se toman dos muestras de 101 individuos cada una y se les pide que accedan a la aplicación y completen la ficha de inscripción, obteniendo unas medias de 50.2 y 52.9 segundos en la primera y segunda aplicación, respectivamente, y unas **varianzas** de 4.75 y 5.35, respectivamente. Si se supone que las poblaciones están normalmente distribuidas, ¿existe diferencia entre las medias del tiempo de inscripción en ambas aplicaciones? ♯

Puesto que piden responder si las medias poblacionales son iguales, la herramienta a usar será un intervalo de confianza de diferencia de medias (y ver si contiene al cero). En tal caso, se tienen dos opciones: varianzas poblacionales iguales o diferentes.
Para ver cómo considerar a las varianzas, se puede obtener el intervalo de confianza para el cociente de varianzas y ver si contiene al uno. Si contiene al 1 las varianzas se considerarán iguales y si no se creerá que son diferentes.

Intervalo de confianza para la diferencia de medias

En primer lugar obtendremos el intervalo de confianza para el cociente de varianzas, esto es:

$$\left[F_{n-1,m-1;\frac{\alpha}{2}} \cdot \frac{S_2^2}{S_1^2}, F_{n-1,m-1;1-\frac{\alpha}{2}} \cdot \frac{S_2^2}{S_1^2}\right].$$

A partir del enunciado se tiene que:

- $n = 101, S_1^2 = \frac{n}{n-1} \cdot 4.75 = \frac{101}{100} \cdot 4.75 = 4.7975$
- $m = 101, S_2^2 = \frac{m}{m-1} \cdot 5.35 = \frac{101}{100} \cdot 5.35 = 5.4035$
- $F_{n-1,m-1;\frac{\alpha}{2}} = F_{100,100;0.025} = 0.67$
- $F_{n-1,m-1;1-\frac{\alpha}{2}} = F_{100,100;0.975} = 1.48$

El intervalo de confianza que contiene al cociente de varianzas con una confianza del 95% es $(0.759,\ 1.671)$
Puesto que dicho intervalo contiene al 1, se puede afirmar, a un nivel de confianza del 95%, que las varianzas poblacionales son iguales.

Intervalo de confianza para la diferencia de medias

Una vez comprobado que las varianzas poblacionales desconocidas son iguales, se puede calcular el intervalo de confianza para la diferencia de medias:

$$\left[(\overline{X} - \overline{Y}) - t_{n+m-2;1-\frac{\alpha}{2}} S_p \sqrt{\frac{m+n}{nm}}, (\overline{X} - \overline{Y}) + t_{n+m-2;1-\frac{\alpha}{2}} S_p \sqrt{\frac{m+n}{nm}}\right].$$

Puesto que:

- $\overline{x} = 50.2, \overline{y} = 52.9$
- $t_{200;0.975} = 1.972$
- $S_p = 2.2584$

El intervalo de confianza que contiene a la diferencia de medias poblacionales a un nivel de confianza del 95% es $(-3.327,\ -2'073)$.
Puesto que el intervalo es negativo, se puede afirmar al 95% de confianza, que $\mu_1 < \mu_2$.
Es decir, el tiempo medio de inscripción es menor en la primera aplicación.

Intervalo de confianza para la diferencia de medias

Script de R

```
alpha=0.05
n = 101
Cuasi_varianza_1=(n/(n-1))*4.75
m=101
Cuasi_varianza_2=(m/(m-1))*5.35
cuantil_inf= qf(alpha/2,n-1,m-1 ,lower.tail = T)
cuantil_sup= qf(1-(alpha/2),n-1,m-1 ,lower.tail = T)
lim_inf=cuantil_inf*Cuasi_varianza_2/Cuasi_varianza_1
lim_sup=cuantil_sup*Cuasi_varianza_2/Cuasi_varianza_1

media_muestral_1=50.2
media_muestral_2=52.9
S_p=sqrt(((n-1)*Cuasi_varianza_1+(m-1)*Cuasi_varianza_2)/(n+m-2))
cuantil= qt(1-(alpha/2),n+m-2,lower.tail = T)
lim_inf=(media_muestral_1-media_muestral_2)-cuantil*S_p*sqrt((m+n)/
lim_sup=(media_muestral_1-media_muestral_2)+cuantil*S_p*sqrt((m+n)/
```


Intervalo de confianza para la diferencia de proporciones

Dadas dos muestras aleatorias simples de tamaño m y n procedentes de dos variables aleatorias independientes distribuidas según dos distribuciones de Bernouilli con probabilidades de éxito p_1 y p_2, respectivamente.
El intervalo de confianza para la diferencia de proporciones poblacionales $p_1 - p_2$ se construye de la siguiente manera:

- $\frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{m}}} \sim N(0,1).$
- $P\left[-z_{1-\frac{\alpha}{2}} \leq \frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{m}}} \leq z_{1-\frac{\alpha}{2}}\right] = 1-\alpha$ donde $z_{1-\frac{\alpha}{2}}$ es el cuantil de una $N(0,1)$ que deja a la izquierda una probabilidad de $1-\frac{\alpha}{2}$
- El intervalo de confianza para $p_1 - p_2$ al nivel de confianza $1-\alpha$ es:

$$\boxed{\left[(\hat{p_1}-\hat{p_2}) - z_{1-\frac{\alpha}{2}} \cdot \delta, (\hat{p_1}-\hat{p_2}) + z_{1-\frac{\alpha}{2}} \cdot \delta\right]}$$

donde $\delta = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n}+\frac{\hat{p_2}(1-\hat{p_2})}{m}}$

Ejemplo 8 Según la encuesta publicada por "The Washington Post" el 5 de noviembre de 2008, el 95% de los encuestados de raza negra y el 43% de los encuestados de raza blanca votaron por Obama. Sabiendo que de los 17834 total de encuestados 13.197 eran de raza blanca, 2318 de raza negra, 1.427 eran latinos y el resto de otras razas. Construya de un intervalo de confianza al 95% para la diferencia de proporciones entre los votantes de raza blanca y de raza negra que votaron por Obama.
Dados los datos se tiene que:

- $n = 13197, \hat{p}_1 = 0.43$
- $m = 2318, \hat{p}_2 = 0.95$
- $z_{1-\frac{\alpha}{2}} = 1.96$

El intervalo de confianza que contiene a la diferencia de proporciones poblacionales es $(-0.532,\ -0.508)$.

♯

Script de R:

```
alpha=0.05
n = 13197
p_muestral_1=0.43
m=2318
p_muestral_2=0.95
cuantil= qnorm(1-(alpha/2),mean=0, sd=1,lower.tail = T)
lim_inf=(p_muestral_1-p_muestral_2)-
        cuantil*sqrt((p_muestral_1*(1-p_muestral_1)/n)+
        p_muestral_2*(1-p_muestral_2)/m)
lim_sup=(p_muestral_1-p_muestral_2)+
        cuantil*sqrt((p_muestral_1*(1-p_muestral_1)/n)+
        p_muestral_2*(1-p_muestral_2)/m)
```

Introducción al contraste de hipótesis

En el presente tema se aborda el problema de inferencia sobre los parámetros desconocidos de una distribución desde un nuevo enfoque. En este caso desarrollaremos un procedimiento, conocido como contraste de hipótesis, que va a permitir discernir si una propuesta sobre los posibles valores que puede tomar un parámetro puede considerarse o no como cierta. Dicha decisión será tomada a partir de unas reglas basadas en la información muestral.

Ejemplo 1 Supongamos que el número de horas semanales dedicado por los empleados de cierta empresa a navegar por internet tiene una distribución normal. Se toma una muestra de seis empleados y se obtiene: 12.2, 18.4, 23.1, 11.7, 8.2, 24. El dueño de la empresa quiere sabe si el tiempo medio semanal dedicado por los empleados a navegar por internet es superior a 10 horas. ♯

Introducción al contraste de hipótesis

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Nuestro método de contraste comienza con una hipótesis sobre el parámetro llamada **Hipótesis Nula ($\mathbf{H_0}$)**, que mantendremos a menos que existan pruebas contundentes en contra de ella.
Si rechazamos la hipótesis nula, entonces aceptaremos la segunda hipótesis llamada **Hipótesis Alternativa ($\mathbf{H_1}$)**.
Si no rechazamos la hipótesis nula, o bien es correcta la hipótesis nula, o bien es correcta la alternativa pero nuestro método de contraste no es suficientemente fuerte para rechazar la hipótesis nula.

Ejemplo 2 Siguiendo el ejemplo 1, el contraste que quiere resolver el empresario es:

$$\left.\begin{array}{l} H_0: \ \mu \leq 10 \\ H_1: \ \mu > 10 \end{array}\right\}$$

Por el momomento, lo único que sabemos es los que nos cuenta la muestra: $\overline{X} = 16.27$ y $S = 6.53$. ♯

Tipos de hipótesis

Hay dos tipos de hipótesis:

- Hipótesis simples: Si especifica un único valor del parámetro estudiado.
- Hipótesis compuestas: Si especifica un rango de valores para el parámetro estudiado.

Ejemplo 3 Sea X una variable aleatoria que modeliza el peso de los alumnos de la UGR y que se distribuye según una normal de media μ y varianza σ^2. Nos planteamos contrastar si el verdadero valor de la media es o no 68 kg. Se han establecido distintas posibilidades clasificadas en simples y compuestas (observar que el signo = siempre aparece en la hipótesis nula):

$$a)\ \left.\begin{array}{l} H_0:\ \mu = 68 \\ H_1:\ \mu \neq 68 \end{array}\right\}, \qquad b)\ \left.\begin{array}{l} H_0:\ \mu \geq 68 \\ H_1:\ \mu < 68 \end{array}\right\}, \qquad c)\ \left.\begin{array}{l} H_0:\ \mu \leq 68 \\ H_1:\ \mu > 68 \end{array}\right\}$$

Tipos de error

Los posibles errores que se pueden cometer al realizar un contraste de hipótesis son:

	H_0 cierta	H_0 falsa
Se rechaza H_0	Error tipo I	Decisión correcta
No se rechaza H_0	Decisión correcta	Error tipo II

Resumiendo

- Se comete error de tipo I cuando se rechaza la hipótesis nula siendo cierta.
  A la probabilidad de cometer el error tipo I se le denomina **nivel de significación** y se le denota por α
- Se comete error de tipo II cuando no rechazamos la hipótesis nula siendo falsa.
  A la probabilidad de cometer el error tipo II se le denota por β y a $1-\beta$ se le denomina **potencia**

Tipos de error

Lo ideal sería que las probabilidades de ambos errores fueran lo más pequeñas posible a la hora de realizar un contraste pero esto es imposible dado que un reducción en la probabilidad de cometer un error de Tipo I conlleva un aumento en la probabilidad de cometer un error tipo II (y una disminución de la potencia del test).

En la práctica se selecciona una probabilidad de cometer un error Tipo I pequeña (**nivel de significación**)y se utiliza esa probabilidad para formular la regla de decisión del contraste.

Cuando realizamos un contraste la hipótesis nula se mantiene como verdadera a menos que los datos contengan pruebas contundentes para rechazarla. Fijando un nivel de significación bajo, tenemos una pequeña probabilidad de rechazar una hipótesis nula verdadera. Cuando la rechazamos, la probabilidad de cometer un error es el nivel de significación α.

Si no rechazamos la hipótesis nula, o bien es verdadera o bien nuestro metodo para detectar una hipótesis nula falsa no tiene suficiente potencia. Cuando rechazamos la hipótesis nula, tenemos pruebas contundentes de que no es verdadera y, por tanto, de que la hipótesis alternativa es verdad.

Introducción al contraste de hipótesis

El procedimiento de contrastación tiene las siguientes etapas:

- Planteamiento de hipótesis nula y alternativa
- Elección del nivel de significación α (habitualmente: 0.01, 0.05 y 0.10)
- Selección de un estadístico de contraste. El estadístico tendrá una distribución basada en la muestra y en el parámetro especificado por la hipótesis nula.
- Se calcula el conjunto de valores del estadístico de contraste que no tienen la probabilidad de darse si la hipótesis nula es verdad. A este conjunto de valores se le denomina **Región de Rechazo**.
- Calculamos el valor del estadístico según la muestra y si ese valor está en la región de rechazo, rechazamos la hipótesis nula y aceptamos la alternativa.
- Si el valor del estadístico NO está en la región de rechazo, **NO rechazamos** la hipótesis nula.

Contraste para la media de una población normal

Sea X una variable aleatoria cuya distribución de probabilidades es normal $N(\mu, \sigma^2)$, donde σ^2 es desconocida. Dada una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de X, los posibles test, al nivel de significación α, para la media de una población normal con varianza desconocida, junto a su correspondiente región de rechazo, son los siguientes:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}\right\}$	$\dfrac{\lvert\overline{X} - \mu_0\rvert\sqrt{n}}{S} > t_{n-1,1-\alpha/2}$
$\left.\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}\right\}$	$\dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} > t_{n-1,1-\alpha}$
$\left.\begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}\right\}$	$\dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} < t_{n-1,\alpha}$

Donde $t_{n-1,a}$ es el punto de una t-Student de $n-1$ grados de libertad que deja por debajo suya una probabilidad igual a a, es decir, $P[t < t_{n-1,a}] = a$, donde $t \sim t_{n-1}$.

Contraste para la media de una población normal

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Ejemplo 4 Supongamos que el número de horas semanales dedicado por los empleados de cierta empresa a navegar por internet tiene una distribución normal. Se toma una muestra de seis empleados y se obtiene: 12.2, 18.4, 23.1, 11.7, 8.2, 24. Contraste, a un nivel de significación de 0.01, si el tiempo medio semanal dedicado por los empleados a navegar por internet es superior a 10 horas.

A partir de los datos se obtiene que $\overline{X} = 16.27$ y $S = 6.53$. Para contrastar, al 1%, las hipótesis

$$\left.\begin{array}{l} H_0 : \mu \leq 10 \\ H_1 : \mu > 10 \end{array}\right\},$$

Valor del estadístico experimental: $T_{exp} = \dfrac{(16.27 - 10)\sqrt{6}}{6.53} = 2.35.$

Valor crítico: $t_{5,0.99} = 3.36$

Dado que el estadístico experimental no está en la región de rechazo, la conclusión sería que **no hay evidencias suficientes para rechazar la hipótesis nula.** ♯

Contraste para la varianza de una población normal

Sea X una variable aleatoria cuya distribución de probabilidades es normal $N(\mu, \sigma^2)$. Entonces, dada una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de X, los posibles test, al nivel de significación α, para la varianza de una población normal, junto a su correspondiente región de rechazo, son los siguientes:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}\right\}$	$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2}$ ó $\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2}$
$\left.\begin{array}{l} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}\right\}$	$\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}$
$\left.\begin{array}{l} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{array}\right\}$	$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,\alpha}$

Donde $\chi^2_{n-1,a}$ es el punto de una chi cuadrado de $n-1$ grados de libertad que deja por debajo suya una probabilidad igual a a, es decir, $P[\chi < \chi^2_{n-1,a}] = a$, donde $\chi \sim \chi^2_{n-1}$.

Contraste para la varianza de una población normal

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Ejemplo 5 Una empresa de ventas de muebles toma una muestra del almacén de 25 muebles elegidos al azar y calcula que el tiempo medio en meses que lo tiene en stock es 1.5 con una cuasivarianza muestral de 0.8. Contraste, con un nivel de significación del 5%, si la varianza poblacional puede tomar el valor 1.

Las hipótesis a contrastar, al 5%, serán:

$$\left.\begin{array}{l} H_0 : \sigma^2 = 1 \\ H_1 : \sigma^2 \neq 1 \end{array}\right\}.$$

Valor del estadístico experimental: $\chi^2_{exp} = \frac{(25-1)0.8}{1} = 19.20,$

Valores críticos: $\chi_{0.025} = 12.40$ y $\chi_{0.975} = 39.36$.

Por tanto, la conclusión sería que **no hay evidencias suficientes para rechazar la hipótesis nula**. ♯

Contraste de hipótesis para la proporción

Sea X una variable aleatoria distribuida según una Bernouilli de parámetro p. Dada una muestra aleatoria simple $X_1, \ldots, X_n$ procedente de X, los posibles test, al nivel de significación α, para la proporción muestral, junto a su correspondiente región de rechazo, son los siguientes:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array}\right\}$	$\dfrac{\lvert\hat{p} - p_0\rvert}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > Z_{1-\alpha/2}$
$\left.\begin{array}{l} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{array}\right\}$	$\dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > Z_{1-\alpha}$
$\left.\begin{array}{l} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{array}\right\}$	$\dfrac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < Z_{\alpha}$

Donde Z_a es el punto de una normal de media cero y varianza uno que deja por debajo suya una probabilidad igual a a, es decir, $P[Z < Z_a] = a$, donde $Z \sim N(0,1)$.

Contraste de hipótesis para la proporción

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Ejemplo 6 Según el artículo publicado en La voz de Galicia, el 29 de noviembre de 2009, antes de la crisis la edificación daba empleo al 14% de los extranjeros. Si se tomó una muestra de 110 habitantes extranjeros y 85 no trabajaban en la construcción, ¿Para un nivel de significación del 0.05 puede considerarse que la afirmación del periódico es cierta?

Contraste de hipótesis: $\left.\begin{array}{l} H_0 : p = 0.14 \\ H_1 : p \neq 0.14 \end{array}\right\}$,

Proporción muestral: $\hat{p} = \frac{25}{110} = 0.23$.

Valor del estadístico muestral: $Z_{exp} = \dfrac{0.23 - 0.14}{\sqrt{\frac{0.23 \cdot 0.77}{110}}} = 2.18$.

Valores críticos: $z_{0.025} = -1.96$ y $z_{0.975} = 1'96$.

Dado que el valor del estadístico muestral está en la región de rechazo, se **rechaza la hipótesis nula**. En consecuencia, la afirmación del periódico no es cierta. ♯

Contraste de diferencias de medias con varianzas poblacionales desconocidas e iguales

Sean X e Y dos variables aleatorias independientes distribuidas según una $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente, donde las varianzas son desconocidas e iguales. Dadas dos muestras aleatorias simples $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ procedentes de X e Y, respectivamente, los posibles test, al nivel de significación α, para la comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales son:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 = \mu_0 \\ H_1 : \mu_1 - \mu_2 \neq \mu_0 \end{array}\right\}$	$\dfrac{\lvert\overline{X} - \overline{Y}\rvert - \mu_0}{S_p\sqrt{\frac{m+n}{mn}}} > t_{n+m-2,1-\alpha/2}$
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 \leq \mu_0 \\ H_1 : \mu_1 - \mu_2 > \mu_0 \end{array}\right\}$	$\dfrac{(\overline{X} - \overline{Y}) - \mu_0}{S_p\sqrt{\frac{m+n}{mn}}} > t_{n+m-2,1-\alpha}$
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 \geq \mu_0 \\ H_1 : \mu_1 - \mu_2 < \mu_0 \end{array}\right\}$	$\dfrac{(\overline{X} - \overline{Y}) - \mu_0}{S_p\sqrt{\frac{m+n}{mn}}} < t_{n+m-2,\alpha}$

Contraste de diferencias de medias con varianzas poblacionales desconocidas e iguales

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

donde $t_{n+m-2,a}$ es el punto de una t-Student de $n+m-2$ grados de libertad que deja por debajo suya una probabilidad igual a a y

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

Ejemplo 7 Se desea comparar el tiempo que se necesita para completar la inscripción en dos aplicaciones informáticas de búsqueda de empleo que se piensan lanzar al mercado. Se toman dos muestras de 101 individuos cada una y se les pide que accedan a la aplicación y completen la ficha de inscripción, obteniendo una media de 50.2 y 52.9 segundos en la primera y segunda aplicación, respectivamente, y una cuasivarianza de 4.75 y 5.35, respectivamente. Si se supone que las poblaciones están normalmente distribuidas y las **varianzas poblacionales son iguales**, ¿existe diferencia entre las medias del tiempo de inscripción en ambas aplicaciones? $\alpha = 0.05$ ♯

Contraste de diferencias de medias con varianzas poblacionales desconocidas e iguales

Contraste de hipótesis: $\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{array}\right\}$

Valor del estadístico muestral: $T_{exp} = \frac{(50.2-52.9)-0}{S_p\sqrt{\frac{101+101}{101*101}}} = -8.538$ donde

$S_p^2 = \frac{(101-1)4.75+(101-1)5.35}{101+101-2} = 5.05$

Valores críticos: $t_{200,0.975} = 1.972$ y $t_{200,0.025} = -1.972$

Como el estadístico pertenece a la región de rechazo, **se rechaza la hipótesis nula**. Existen diferencias en el tiempo medio de inscripción.

Contraste de diferencias de medias con varianzas poblacionales desconocidas y diferentes

Sean X e Y dos variables aleatorias independientes distribuidas según una $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente, donde las varianzas son desconocidas y diferentes. Dadas dos muestras aleatorias simples $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ procedentes de X e Y, respectivamente, los posibles test, al nivel de significación α, para la comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales son:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 = \mu_0 \\ H_1 : \mu_1 - \mu_2 \neq \mu_0 \end{array}\right\}$	$\frac{\lvert\overline{X} - \overline{Y}\rvert - \mu_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} > t_{v,1-\alpha/2}$
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 \leq \mu_0 \\ H_1 : \mu_1 - \mu_2 > \mu_0 \end{array}\right\}$	$\frac{(\overline{X} - \overline{Y}) - \mu_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} > t_{v,1-\alpha}$
$\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 \geq \mu_0 \\ H_1 : \mu_1 - \mu_2 < \mu_0 \end{array}\right\}$	$\frac{(\overline{X} - \overline{Y}) - \mu_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} < t_{v,\alpha}$

donde $t_{v,a}$ es el punto de una t-Student de v grados de libertad que deja por debajo suya una probabilidad igual a a y

$$v = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_2^2}{m}\right)^2}{m-1}}$$

Ejemplo 8 Se desea comparar el tiempo que se necesita para completar la inscripción en dos aplicaciones informáticas de búsqueda de empleo que se piensan lanzar al mercado. Se toman dos muestras de 101 individuos cada una y se les pide que accedan a la aplicación y completen la ficha de inscripción, obteniendo una media de 50.2 y 52.9 segundos en la primera y segunda aplicación, respectivamente, y una cuasivarianza de 4.75 y 5.35, respectivamente. Si se supone que las poblaciones están normalmente distribuidas y las **varianzas poblacionales son diferentes**, ¿existe diferencia entre las medias del tiempo de inscripción en ambas aplicaciones? $\alpha = 0.05$

Contraste de diferencias de medias con varianzas poblacionales desconocidas y diferentes

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Contraste de hipótesis: $\left.\begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{array}\right\}$

Valor del estadístico muestral: $T_{exp} = \frac{(50.2-52.9)-0}{\sqrt{\frac{4.75}{101}+\frac{5.35}{101}}} = -8.538$

$$v = \frac{\left(\frac{4.75}{101}+\frac{5.35}{101}\right)^2}{\frac{\left(\frac{4.75}{101}\right)^2}{101-1}+\frac{\left(\frac{5.35}{101}\right)^2}{101-1}} = 199.297$$

Valores críticos: $t_{v,0.975} = 1.972$ y $t_{v,0.025} = -1.972$

Como el estadístico pertenece a la región de rechazo, **se rechaza la hipótesis nula**. Existen diferencias en el tiempo medio de inscripción.

Sean X e Y dos variables aleatorias independientes distribuidas según una $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente. Dadas dos muestras aleatorias simples $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ procedentes de X e Y, respectivamente, los posibles test, al nivel de significación α, para la comparación de varianzas de dos poblaciones normales con medias desconocidas son:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array}\right\}$	$\frac{S_1^2}{S_2^2} < F_{n-1,m-1,\alpha/2}$ ó $\frac{S_1^2}{S_2^2} > F_{n-1,m-1,1-\alpha/2}$
$\left.\begin{array}{l} H_0 : \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{array}\right\}$	$\frac{S_1^2}{S_2^2} > F_{n-1,m-1,1-\alpha}$
$\left.\begin{array}{l} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{array}\right\}$	$\frac{S_1^2}{S_2^2} < F_{n-1,m-1,\alpha}$

Donde $F_{n-1,m-1,a}$ es el punto de una F-Snedecor de $n-1$ y $m-1$ grados de libertad que deja por debajo suya una probabilidad igual a a.

Ejemplo 9 Dos profesores están interesados en comparar el nivel académico de dos grupos en una misma asignatura. Tomando una muestra de 81 alumnos para el primer grupo, se obtiene una nota media de 5.7 con una cuasivarianza de 1.2. Mientras que para una muestra de 61 alumnos del segundo grupo, se obtiene una nota media de 7 con cuasivarianza igual a 2.2. ¿Existe diferencia entre la dispersión de ambos grupos? ($\alpha = 0.01$)

Contraste de hipótesis: $\left.\begin{array}{l} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{array}\right\}$.

Valor del estadístico experimental: $F_{exp} = \dfrac{1.2}{2.2} = 0.545$

Valores críticos: $F_{80,60,0.005} = 0.539$ y $F_{80,60,0.995} = 1.899$

Como el valor del estadístico NO está en la región de rechazo, **no existe evidencia empírica para rechazar la hipótesis nula**. Es decir, no existe evidencia empírica para rechazar que la dispersiones de ambos grupos son iguales. ♯

Contraste para la diferencia de proporciones

Sean X e Y dos variables aleatorias independientes que se distribuyen según dos distribuciones de Bernouilli con probabilidades de éxito p_1 y p_2, respectivamente. Dadas dos muestras aleatorias simples $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ procedentes de X e Y, respectivamente, los posibles test, al nivel de significación α, para la diferencia de proporciones muestrales son:

Casos	Región de rechazo
$\left.\begin{array}{l} H_0 : p_1 - p_2 = p_0 \\ H_1 : p_1 - p_2 \neq p_0 \end{array}\right\}$	$\dfrac{\lvert(\hat{p}_1 - \hat{p}_2) - p_0\rvert}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}} > z_{1-\alpha/2}$
$\left.\begin{array}{l} H_0 : p_1 - p_2 \leq p_0 \\ H_1 : p_1 - p_2 > p_0 \end{array}\right\}$	$\dfrac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}} > z_{1-\alpha}$
$\left.\begin{array}{l} H_0 : p_1 - p_2 \geq p_0 \\ H_1 : p_1 - p_2 < p_0 \end{array}\right\}$	$\dfrac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}} < z_{\alpha}$

Donde z_a es el punto de una normal de media cero y varianza uno que deja por debajo suya una probabilidad igual a a.

Contraste para la diferencia de proporciones

Introducción
Contraste para la media
Contraste para la varianza
Contraste para la proporción
Contraste para diferencias de medias
Contraste para el cociente de varianzas
Contraste para la diferencia de proporciones

Ejemplo 10 Se encuestaron a 200 estudiantes de la UGR, de los cuales 20 manifestaron estar en contra del denominado "Plan Bolonia". En la UMA se encuestaron a 180 estudiantes, siendo 30 los que se mostraron disconformes con el nuevo plan. Al nivel de significación del 5%, ¿puede decirse que la proporción de alumnos que no están de acuerdo con Bolonia es distinta en cada una de las provincias?

Del enunciado se obtiene que $\hat{p}_1 = \frac{20}{200} = 0.1$ y $\hat{p}_2 = \frac{30}{180} = 0.16$.

Contraste de hipótesis: $\left.\begin{array}{l} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \end{array}\right\}$,

Valor del estadístico muestral:

$$Z_{exp} = \frac{0.1 - 0.17}{\sqrt{\frac{0.1(1-0.1)}{200} + \frac{0.17(1-0.17)}{180}}} = -1.907$$

Valores críticos: $z_{0.025} = -1.96$ y $z_{0.975} = 1.96$

Como el valor del estadístico NO está dentro de la región de rechazo, **NO existe evidencia empírica para rechazar la hipótesis nula.** ♯

Técnicas Cuantitativas II

CONTRASTES NO PARAMÉTRICOS

Introducción

Los contrastes no paramétricos son el método adecuado para extraer conclusiones estadísticas sobre datos cualitativos (nominales u ordinales) o sobre datos numéricos cuando no se puede aceptar el supuesto de normalidad de la distribución de probabilidad de la población.

- Contrastes de bondad de ajuste
  - Contraste de bondad de ajuste χ^2.
  - Contraste de bondad de ajuste Kolmogorov-Smirnov.
  - Contraste de normalidad de Lilliefors.
- Contraste de Kolmogorov-Smirnov para dos muestras independientes
- Contraste de Kruskal-Wallis para k muestras independientes

Contraste de bondad de ajuste χ^2

El contraste χ^2 se emplea para decidir si un conjunto de datos provienen de una distribución de probabilidad dada. En general, planteamos:

$$\begin{cases} H_0 : X \text{ sigue una distribución } F(X) \\ H_1 : X \text{ no sigue una distribución } F(X) \end{cases}$$

Ejemplo 1 Se observó una muestra de 300 sujetos que compraron una bebida refrescante. De estos sujetos, 75 seleccionaron la marca A, 110 seleccionaron la marca B y el resto seleccionó la marca C. Si tomamos un sujeto de forma aleatoria ¿tiene las mismas probabilidades de seleccionar cualquiera de las tres variedades?
Vamos a realizar un contraste de hipótesis donde la hipótesis nula es las **tres categorías tienen la misma probabilidad de ser elegidas**. ♯

La idea básica de este contraste consiste en comparar las frecuencias observadas en la muestra, con las que se esperarían obtener si H_0 fuese cierta. El test chi-cuadrado se ha diseñado para su aplicación con distribuciones discretas, sin embargo, es válido también para distribuciones de tipo continuo.

Contraste de bondad de ajuste χ^2

En este contraste trabajaremos con una muestra aleatoria de tamaño *n* procedente de una v.a. (población) X dividida en *k* **clases exhaustivas y mutuamente excluyentes** (es decir, cada observación muestral puede pertenecer a una sola de las categorías). Denotaremos por O_i a la frecuencia de la categoría i **observada** en la muestra.
Imaginemos que las **probabilidades** de pertenencia a cada una de las categorías que especifica la **hipótesis nula** son p_1, p_2,...,p_k (por supuesto, la suma de estas probabilidades debe ser 1).
Entonces, si hay n observaciones muestrales, el **número esperado** en cada categoría, si se cumple la hipótesis nula, es

$$E_i = np_i, i = 1, 2, ...k$$

Toda esta información se resumiría en una tabla de este tipo:

C_i	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$
C_1	O_1	p_1	$E_1 = n \cdot p_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	O_k	p_k	$E_k = n \cdot p_k$
	n	1	n

Ejemplo 2

C_i	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$
$MarcaA$	75	$\frac{1}{3}$	100
$MarcaB$	110	$\frac{1}{3}$	100
$MarcaC$	115	$\frac{1}{3}$	100
	300	1	300

♯

A partir de los valores definidos definimos el estadístico de contraste como:

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

donde la distribución del estadístico de contraste será una χ^2_{k-1} de $k-1$ grados de libertad.

Para hallar χ^2_{exp} es aconsejable disponer los cálculos en la tabla (o utilizar R):

Ejemplo 3

C_i	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$	$\frac{(O_i - E_i)^2}{E_i}$
$MarcaA$	75	$\frac{1}{3}$	100	$\frac{(75-100)^2}{100} = 6.25$
$MarcaB$	110	$\frac{1}{3}$	100	$\frac{(110-100)^2}{100} = 1$
$MarcaC$	115	$\frac{1}{3}$	100	$\frac{(115-100)^2}{100} = 2.25$
	300	1	300	$\chi^2_{exp} = 9.5$

♯

La interpretación intuitiva del estadístico de contraste o experimental es *Valores pequeños del estadístico mostrarán una alta concordancia entre las frecuencias que se observan y las que se esperaban, por lo que no se podrá rechazar* H_0. Por otro lado, si el estadístico toma un valor muy grande mostrará la discrepancia entre estas frecuencias y habrá que rechazar H_0. Luego fijado un nivel de significación α, se rechaza H_0, si $\chi^2 > \chi^2_{k-1;1-\alpha}$.

Ejemplo 4 Para un nivel de significación del 1%, calculamos $\chi^2_{3-1,1-0.01} = \chi^2_{2,0.99} = 9.21$, como $\chi^2 = 9.5 > \chi^2_{2,0.99} = 9.21$ entonces existen razones para rechazar H_0 con una significación del 1%. Rechazamos que las preferencias por cada marca sean las mismas. ♯

CONDICIÓN DE VALIDEZ DEL TEST.

- Si $E_i < 5$ entonces χ^2 se hace grande y por tanto se rechaza H_0 sin razón. Este contraste es apropiado siempre que $E_i > 5$, para cualquier valor de i. Si esto no ocurre tendríamos que agrupar clases vecinas, pero por cada par de clases que se combinen hay que reducir en 1 los grados de libertad de la distribución chi del estadístico.
- Cuando se desee contrastar la hipótesis de que los datos están generados por alguna distribución (p.e. Binomial, Poisson o Normal), con parámetros de dicha distribución desconocidos se utilizarán los datos muestrales para obtener los estimadores puntuales correspondientes, pero los grados de libertad de la chi-cuadrado del contraste se reducirán en una unidad por cada parámetro estimado. En general, si es necesario estimar r parámetros, la distribución del estadístico de contraste será una χ^2_{k-r-1} de $k - r - 1$ grados de libertad.

Ejemplo 5 En un cajero automático situado en el extrarradio de una ciudad se ha detectado una baja utilización del mismo por los clientes de la sucursal bancaria. Con el fin de investigar esta afirmación, se ha controlado el número de llegadas al mismo durante las tardes en que la oficina permanece cerrada, contabilizándose los siguientes resultados:

Número de llegadas al cajero	Número de tardes
0	21
1	18
2	7
3	3
4 o más	1

Con base en esta información, ¿existe alguna razón para creer que el número de llegadas por tarde es una variable de Poisson con parámetro 0.9? ($\alpha = 0.05$) ♯

SOLUCIÓN

Definimos la variable aleatoria X como el número de llegadas al cajero por la tarde, cuya función de distribución es una Poisson de parámetro 0.9. Por tanto, planteamos las hipótesis:

$$\begin{cases} H_0 : X \to P(0.9) \\ H_1 : X \nrightarrow P(0.9) \end{cases}$$

Para hallar las p_i del cuadro resumen, utilizamos las tablas de la distribución de Poisson y calculamos $P[X=0], P[X=1], P[X=2], P[X=3], P[X \geq 4]$;

	O_i	$p_i = P[x \in S_i]$ bajo H_o	$E_i = n \cdot p_i$
$X = 0$	21	0.4066	20.33
$X = 1$	18	0.3659	18.3
$X = 2$	7	0.1647	8.24
$X = 3$	3	0.0494	**2.47** < 5
$X \geq 4$	1	0.0134	**0.67** < 5
	50	1	$\cong 50$

Como se puede observar, el número esperado de las observaciones de las clases $X = 3$ y $X \geq 4$ son inferiores a 5, por tanto para que se cumpla la condición de validez del test las agrupamos obteniendo:

	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$	$\frac{(O_i - E_i)^2}{E_i}$
$X = 0$	21	0.4066	20.33	0.02208
$X = 1$	18	0.3659	18.3	0.004918
$X \geq 2$	11	0.2215	11.38	0.012689
	50	1	$\cong 50$	$\chi^2 = 0.0393$

Calculamos $\chi^2_{3-1,1-0.05} = \chi^2_{2,0.95} = 5.99$, como $\chi^2 = 0.0393 \not> \chi^2_{2,0.95} = 5.99$ entonces no existen razones para rechazar H_0 con una significación del 5%. No podemos rechazar que los datos provengan de una distribución de Poisson de parámetro 0.9. Esta conclusión nos permite afirmar que el cajero es muy poco utilizado ya que el número medio de llegadas esperadas por tarde es menor de 1.

■

Ejemplo 6 Una muestra sobre el número de personas que diariamente requieren información de un producto financiero ofrece los resultados 3, 0, 1, 3, 2, 4, 4, 5, 5, 3, 3, 1, 2, 2, 3, 4, 3, 3, 2, 4, 5, 1, 0, 4, 2, 3, 1. ¿Se puede aceptar que el número de personas que requieren la mencionada información se distribuye según una ley de Poisson?

SOLUCIÓN

Definimos la variable aleatoria X como el número de personas que diariamente requieren información. Para estudiar si el número de personas sigue una distribución de Poisson, estimamos el valor del parámetro λ a partir de la media de la muestra, es decir, $\widehat{\lambda} = \overline{X} = \frac{73}{27} = 2.7$. Por tanto, planteamos las hipótesis

$$\begin{cases} H_0 : X \to P(2.7) \\ H_1 : X \nrightarrow P(2.7) \end{cases}$$

	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$
$X = 0$	2	0.0672	**1.8144** < 5
$X = 1$	4	0.1815	**4.9005** < 5
$X = 2$	5	0.2450	6.615
$X = 3$	8	0.2205	5.9536
$X = 4$	5	0.1488	**4.0176** < 5
$X \geq 5$	3	0.01370	**3.699** < 5
	27	1	27

Al igual que nos pasaba en el ejemplo anterior, agrupamos las dos primeras clases y las dos últimas clases, obteniendo la tabla:

	O_i	$p_i = P[x \in S_i]$ bajo H_0	$E_i = n \cdot p_i$	$\frac{(O_i - E_i)^2}{E_i}$
$X \leq 1$	6	0.2487	6.7138	0.0761
$X = 2$	5	0.2450	6.6140	0.3943
$X = 3$	8	0.2205	5.9526	0.7035
$X \geq 4$	8	0.2858	7.7195	0.0104
	27	1	27	$\chi^2 = 1.1841$

Hallamos $\chi^2_{4-\mathbf{1}-1,1-0.05} = \chi^2_{2,0.95} = 5.99$, como $\chi^2 = 1.1841 \not> \chi^2_{2,0.95} = 5.99$ entonces no rechazamos H_0 con una significación del 5%. Luego los datos provienen de una distribución de Poisson de parámetro 2.7.

■

Contraste de bondad ajuste de Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov (K-S) permite contrastar si los datos observados en una muestra proceden de una población con una distribución de probabilidad dada de antemano. La hipótesis que se establece es:

$$\begin{cases} H_0 : X \sim F(X) \\ H_1 : X \nsim F(X) \end{cases}$$

Para contrastar este supuesto disponemos de una muestra $x_1, \ldots, x_n$ extraída de esa población. Los pasos a seguir a la hora de realizar el contraste son:

1. Se ordenan los valores de la muestra de menor a mayor.
2. Para cada uno de los elementos muestrales, se calcula la función de distribución bajo H_0:
$$F_0(x_i) = P[X \leq x_i], i = 1, \ldots, n.$$
3. Se calcula la función de distribución empírica o muestral:
$$F_n(x_i) = \frac{\text{número de observaciones menores o iguales a } x_i}{\text{número total de datos}}$$
.

Contraste de bondad ajuste de Kolmogorov-Smirnov

Introducción
Contraste χ^2
Contraste de Kolmogorov-Smirnov
Contraste de Lilliefors
Contraste K-S para dos muestras
Contraste Kruskal-Wallis para k muestras

5. Se calcula el **Estadístico experimental**
$$D_{exp} = max \mid F_0(x_i) - F_n(x_i) \mid$$

6. Fijado el nivel de significación α y conocido el número de elementos en la muestra, n, se obtiene el valor crítico de la tabla correspondiente al Test K-S sobre Bondad de Ajuste, que denotaremos D_α.
Se rechaza H_0 si $D_{exp} > D_\alpha$.

Ejemplo 7 Con un nivel de significación del 5%, contraste la hipótesis de que los siguientes valores muestrales 4, 2, 5, 3, 4, 5, 4, 2 proceden de una distribución normal con media 3.5 y varianza 1.1^2.
♯

SOLUCIÓN

Planteamos la hipótesis a contrastar

$$\begin{cases} H_0 : X \sim N(3.5, 1.1^2) \\ H_1 : X \nsim N(3.5, 1.1^2) \end{cases}$$

Resumimos la información que tenemos en una tabla ordenando los valores muestrales en orden creciente:

Muestra	n_i	N_i	$F_0(x_i)$	$F_n(x_i)$	$\mid F_0(x) - F_n(x) \mid$
2	2	2	$P[X \leq 2] = 0.0863$	$\frac{2}{8} = 0.25$	0.1637
3	1	3	$P[X \leq 3] = 0.3247$	$\frac{3}{8} = 0.375$	0.0503
4	3	6	$P[X \leq 4] = 0.6753$	$\frac{6}{8} = 0.75$	0.0747
5	2	8	$P[X \leq 5] = 0.9137$	$\frac{8}{8} = 1$	0.0863
	8				

Se calcula el estadístico experimental

$$D_{exp} = max \mid F_0(x) - F_n(x) \mid = 0.1637$$

Fijado el nivel de significación al 5%, $D_{0.05} = 0.45427$. Por tanto, como $D_{exp} \not> D_{0.05}$ se mantiene la hipótesis nula. Luego, a un nivel de significación del 5% y teniendo en cuenta la información muestral, se concluye que los datos han sido extraídos de una población normal cuya media y varianza son 3.5 y 1.1^2 respectivamente.

■

Ejemplo 8 Supongamos una muestra aleatoria de tamaño $n = 6$ constituida por las observaciones 0.38, 0.55, 0.32, 0.48, 0.50, 0.20. A un nivel de significación del 5% ¿se puede afirmar que la muestra considerada procede de una población con función de distribución $F(X)$?

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

♯

SOLUCIÓN. Con el fin de saber si los datos considerados pertenecen a una población con función de distribución $F(X)$ recurriremos al test de K-S. En el cuadro siguiente se muestran los pasos necesarios con el fin de obtener el estadístico de contraste de K-S.

Muestra	$F_0(x_i)$	$F_n(x_i)$	$\mid F_0(x) - F_n(x) \mid$
0.20	0.20	0.167	0.033
0.32	0.32	0.333	0.013
0.38	0.38	0.500	0.120
0.48	0.48	0.667	0.187
0.50	0.50	0.833	0.333
0.55	0.55	1	0.450

Luego,

$$D_{exp} = max \mid F_0(x) - F_n(x) \mid = 0.450$$

Por otro lado, el valor crítico que se obtiene de la correspondiente tabla es $D_{0.05} = 0.51926$, y como consecuencia de que $D_{exp} \not> D_{0.05}$ no hay evidencias significativas, a un nivel de significación del 5%, para rechazar la hipótesis nula. Por tanto, se mantiene la hipótesis de que los datos muestrales han sido extraídos de una población cuya función de distribución es

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

■

Contraste de Normalidad de Lilliefors

El test de Kolmogorov-Smirnov no puede ser aplicado en aquellos casos en los que la distribución normal no tiene la información sobre los parámetros poblacionales. En 1967 Lilliefors presentó una modificación del test de K-S con el fin de contrastar la hipótesis de normalidad donde **no se especifica el valor de la media y varianza poblacional**.
La hipótesis que se desea contrastar sería pues:

$$\begin{cases} H_0 : X \sim N\left(\mu, \sigma^2\right) \\ H_1 : X \nsim N\left(\mu, \sigma^2\right) \end{cases}$$

donde μ y σ^2 son estimadas con $\overline{X}$ y S^2 respectivamente.
Los pasos a realizar coinciden con los desarrollados en el test de K-S, la única diferencia que presentan los dos test es la región de rechazo. Los valores críticos se buscan en la tabla "corrección de Lilliefors para normalidad" y se denotará por $D_{L,\alpha}$.

1. Se ordenan los valores de la muestra de menor a mayor.
2. Para cada uno de los elementos muestrales, se calcula la función de distribución real bajo H_0.
3. Se calcula la función de distribución empírica o muestral .
4. Se calcula el **Estadístico experimental** $D_{exp} = max \mid F_0(x_i) - F_n(x_i) \mid$.
5. Fijado el nivel de significación α, se rechaza H_0 si $D_{exp} > D_{L,\alpha}$.

Contraste de Normalidad de Lilliefors

Introducción
Contraste χ^2
Contraste de Kolmogorov-Smirnov
Contraste de Lilliefors
Contraste K-S para dos muestras
Contraste Kruskal-Wallis para k muestras

Ejemplo 9 Con el fin de estudiar el número de accidentes laborales en una determinada provincia se ha seleccionado, tras una campaña de información y prevención, una muestra en la cual se recogió el número de accidentes durante seis semanas consecutivas. Los datos muestrales son 8, 11, 9, 7, 9 y 10. ¿Puede decirse, al 10% de significación, que la muestra considerada proviene de una población normal? ♯

Solución. Considerando como estimador de μ y σ^2, la media muestral, $\overline{X} = 9$, y la cuasivarianza, $S^2 = 2$, respectivamente, la hipótesis a describir es

$$\begin{cases} H_0 : X \sim N(9, \sigma^2 = 2) \\ H_1 : X \nsim N(9, \sigma^2 = 2) \end{cases}$$

A continuación se resumen todos los resultados obtenidos en cada uno de los pasos:

Muestra	$F_0(x_i)$	n_i	$F_n(x_i)$	$\mid F_0(x) - F_n(x) \mid$
7	0.0786	1	0.167	0.0880
8	0.2398	1	0.333	0.0936
9	0.5000	2	0.667	**0.1667**
10	0.7602	1	0.833	0.0730
11	0.9214	1	1	0.0786

Como se verifica que $D_{exp} = 0.1667 \not> D_{L,0.10} = 0.294$ entonces se puede afirmar, a un nivel de significación del 10% que los datos han sido extraídos de una población normal.

Contraste de Kolmogorov-Smirnov para dos muestras

El contraste de Kolmogorov-Smirnov para dos muestras permite contrastar si dos muestras independientes (de tamaño n y m, respectivamente) han sido obtenidas de la misma población (se distribuyen igual).

$$\begin{cases} H_0 : F_n(x) = F_m(x) \\ H_1 : F_n(x) \neq F_m(x) \end{cases}$$

Este test detecta todo tipo de diferencias en las distribuciones: diferencias en la tendencia central (media, mediana), en la dispersión y en la asimetría.
Los pasos a realizar son:

1. Se entremezclan y se ordenan los valores de las dos muestras de menor a mayor.
2. Se calcula la función de distribución empírica de las dos muestras.
3. Se calcula el **Estadístico experimental**

$$D_{exp} = max \mid F_n(x_i) - F_m(x_i) \mid$$

.

4. Se busca el valor crítico (D_α) en las tablas de distribución del estadístico. Hay dos tablas distintas dependiendo de si $n = m$ o los tamaños muestrales son distintos.
5. Fijado el nivel de significación α, se rechaza H_0 si $D_{exp} > D_\alpha$.

Contraste de Kruskal-Wallis para k muestras

El contraste de Kruskal-Wallis permite contrastar si k muestras independientes han sido obtenidas de la misma población (se distribuyen igual).

$$\begin{cases} H_0 : F_1(x) = \cdots = F_k(x) \\ H_1 : \text{existe alguna distinta} \end{cases}$$

Consideremos k muestras aleatorias e independientes de tamaños n_1, n_2, ..., n_k.
Sea $n = n_1 + n_2 + \cdots + n_k$ el tamaño total de la muestra
Los pasos a realizar son:

1. Se asignan rangos desde 1 a n al conjunto de n observaciones como si se tratara de una sola muestra. Si existen empates se asigna el promedio de los rangos empatados.
2. Se calcula R_j como la suma de los rangos asignados a las n_j observaciones de la muestra j.
3. Se calcula el **Estadístico experimental** $H_{exp} = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$ que, bajo la hipótesis nula, tiene una distribución Chi-cuadrado con $k-1$ grados de libertad.
4. Fijado el nivel de significación α, se rechaza H_0 si $H_{exp} > \chi_{k-1;1-\alpha}$.

Contraste de Kruskal-Wallis para k muestras

Ejemplo 10 Contrastar la hipótesis de que los siguientes valores muestrales proceden de una misma población ($\alpha = 0.05$).
Muestra 1: 8, 15, 9, 13, 18, 11, 17
Muestra 2: 4, 13, 12, 13, 13, 10
Muestra 3: 13, 12, 12, 12, 14, 3, 6
Muestra 4: 14, 7, 9, 10, 6, 10 ♯

.