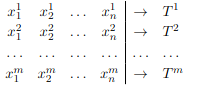
**Inferencia estadística**

**Conceptos**

* Una **muestra** es un subconjunto representativo de la población.
* Sea **X** la variable que estudiamos. Denotaremos por {X1, …, Xn} una **muestra aleatoria simple (m.a.s)**, donde cada variable Xi tomará un valor xi.
  + Al conjunto de valores que toma la muestra {x1, …, xn} se denomina realización muestral.
* Identificamos la población con la distribución de la variable que analizamos. Por ejemplo, ‘una población Normal’ hace referencia a la distribución de nuestra variable, no al conjunto de individuos.
* **Parámetro:** Característica de la población. Por ejemplo, en una población Binomial, el parámetro que caracteriza a nuestra variable es p[[1]](#footnote-0).
  + El parámetro puede ser unidimensional (caso previo) o bidimensional (por ejemplo, en una N(μ, σ2))
  + Denotamos por θ al parámetro de interés en la población. Se dirá que la variable tiene distribución F(θ).
* **Estadístico:** Cualquier función de la muestra (media, varianza, máximo…) Se denotan por T(X1, …, Xn)
  + **Estimadores:** Estadísticos independientes de los parámetros de la población. Por ejemplo, la media de la muestra es un estimador de la media poblacional, y la varianza muestral lo es de la poblacional
* **Método de muestreo:** Procedimiento por el cual se selecciona la muestra.

**Distribuciones en el muestreo**

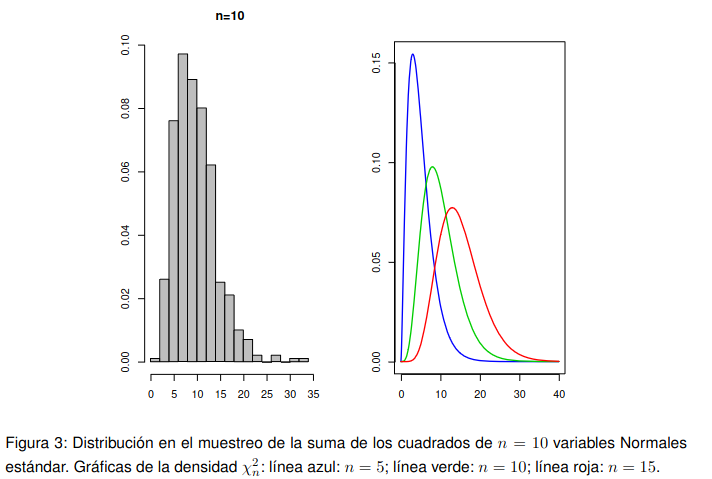
* Sea X1, …, Xn una muestra aleatoria simple de una v.a. X con distribución F(θ).
* Se considera que la muestra sigue la misma distribución que la variable aleatoria.
* Consideremos que F es una distribución normal. Entonces, debemos calcular su media y su varianza.
* Definimos T(X1, …, Xn) como un estadístico. Dado que es una función de la muestra aleatoria, tendrá también un comportamiento aleatorio y por tanto una distribución.
  + Se debe tener en cuenta que el valor del será distinto para distintas muestras de la misma población. 
  + Si realizamos m muestras de X1,..Xn, denotamos por T1, …, Tm los distintos valores del estadístico en cada una de las muestras.
  + Estos valores de T1, …, Tm seguirán una distribución.

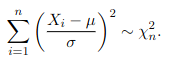
**Proporción muestral**

* La **proporción muestral** es la relación de casos de éxitos en una muestra respecto al tamaño de la muestra. Se denota con . (p^ nestes apuntes)
* Si me interesa conocer la proporción de elementos p de una población X ∼ Ber(p), seleccionamos una MAS X1,...,Xn de variables Ber(p).
* **Distribución:**  → 

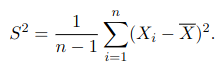
**Media muestral (varianza conocida)**

* Supongamos que disponemos de una m.a.s X1, …, Xn tal que X∼ N (μ, σ2). La **media muestral** Xbarra=  se puede escribir como la suma total de n términos .
  + **Nota:** no confundir con la **media real** de la población, denotada por **μ**
* Debido a esto, la media muestral sigue una distribución normal:  que se puede tipificar como 
  + Es decir, los posibles valores de Xbarra se distribuyen según una Normal, centrada en la media real **μ** y cuya varianza disminuye a medida que aumenta el tamaño n de la muestra.

**Distribución χ2 (ji-cuadrado)**

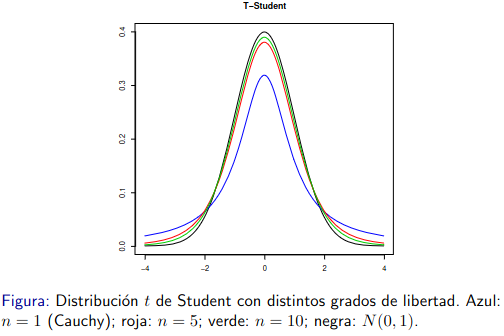
* Supongamos que disponemos de una m.a.s X1, …, Xn de variables Xi∼N(μ,σ2) Entonces, 
* Cuando la muestra es lo suficientemente grande, una distribución χ2n se puede aproximar por una N(n,2n)
* La distribución χ2n **no** es simétrica.

**Distribución de la varianza/cuasivarianza muestral**

* El **Teorema de Fisher** establece que si X1, ..., Xn es una m.a.s. de variables normales con varianza σ2, entonces X y s2 son independientes y además: (imagen) (s2 es la varianza muestral)
* Si en su lugar queremos utilizar la **cuasivarianza** S2,

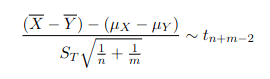
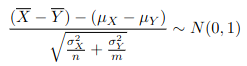
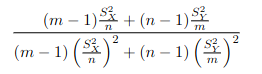
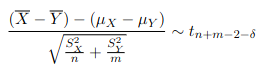
**Distribución t de Student**

* Consideremos una variable X ∼ N (0, 1) y otra v.a. Y ∼ χ2n independientes, el cociente: sigue una **distribución t** con **n grados de libertad**.
* Cuando n es suficientemente grande, se aproxima a una N(0,1).
* Permite describir la distribución de la **media muestral** si no se conoce la varianza poblacional.



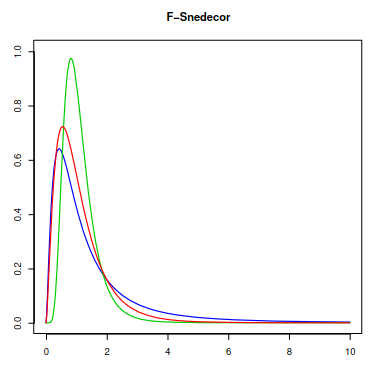


**Distribución de la diferencia de medias**[[2]](#footnote-1)

* Supongamos ahora que tenemos dos poblaciones y las correspondientes muestras X1, . . . , Xn m.a.s. de X ∼ N (μX , σ2X) e Y1, . . . , Ym una m.a.s. de Y ∼ N (μY , σ2Y ).
* Si σ2X y σ2y son conocidas: 
* Si σ2X y σ2y son desconocidas, pero iguales: ,
* Si σ2X y σ2y son desconocidas, y distintas

, sendo δ o enteiro máis proximo a

**Distribución F de Snedecor**

* Sexa X unha va con distribución χ2n e Y con distribución χ2m, ambas independientes. O cociente sigue unha distribución F con n e m graos de liberdade.
* Azul: n = 5, m = 5
* Roja: n = 5, m = 20
* Verde: n = 20, m = 20.

**Distribución do cociente de varianza**

* Polo teorema de Fisher, coñecemos que 
* Entón, o cociente de varianzas ten distribución F con n-1 e m-1 graos de liberdade:

**Intervalo de confianza**

* Dada unha m.a.s X1, …, Xn de X ∼ Fθ, un **intervalo de confianza de nivel (1-α)** con α € (0,1) é un intervalo aleatorio tal que[[3]](#footnote-2)
* 
  + Os extremos dos intervalos son aleatorios porque dependen da mostra.
* É dicir, hai unha probabilidade 1-α de que a m.a.s pertenza a ese intervalo.
  + Desta forma, aínda que non coñezamos o valor dun estadístico, podemos calcular un intervalo ao que existe unha determinada probabilidade ao que pertenzan, estimando o seu valor.

**Intervalo de confianza para a proporción p** (p descoñecido)[[4]](#footnote-3)

* Para unha proporción p que descoñecemos, grazas ao teorema central do límite coñecemos que p^∼N(p, p(1-p)/n).
* Tipificando a variable obtenemos que ∼ N(0,1). Denominamos esta nova m.a.s como p2.
  + p2 coñécese como o **estadístico pivote:** o elemento do que partimos para obter o intervalo de confianza.
* Non podemos calcular directamente o intervalo P(L1<=p<=L2)=1-α porque non coñecemos a distribución de p, pero si coñecemos a de p2.
* P(-z1-α/2 <= p2 <= z1-α/2) = 1-α. Ahora, despexamos p de dentro de p2. Finalmente, substituímos todas as instancias de p por p^ para que o intervalo non dependa de p (que é descoñecido)
  + **Nota:** zp denota o valor tal que P(N(0,1)<=zp)=p.
  + Por exemplo, se se pedise un intervalo de confianza 90%, teríamos 1-α=0.9→α=0.1 →Calculamos z0.95, é decir, consultamos a tabla para ver que valor de zp se corresponde coa p=0.95 (ver en que casilla se atopa este valor). Neste caso, o máis aproximado sería entre 1.75 e 1.76, poderiamos tomar zp=1.755.
* O intervalo final obtido é simétrico respecto de p^:[[5]](#footnote-4)

| (p^ - z1-α/2 , p^ + z1-α/2 ) |
| --- |

* Debido a que é simétrico, podemos expresar o intervalo como:

| **Ic =** |
| --- |

**Intervalo de confianza para a media** con varianza coñecida

* Sexa X1, …, Xn unha m.a.s de X∼N(μ,σ2) de varianza coñecida. Entón, ∼N(0,1). será o noso estadístico pivote.[[6]](#footnote-5)
* O intervalo do que partimos é P(-z1-a/2, , z1-a/2) = 1-a.

| **Ic =** |
| --- |

**Intervalo de confianza para a media** con varianza descoñecida

* Se non coñecemos a varianza, odemos utilizar ∼tn-1. Este será o estadístico pivote.
* O intervalo do que partimos é P(-tn-1, 1-a/2, , tn-1, 1-a/2) = 1-a.

| **Ic =** |
| --- |

**Exemplo de exercicio** (intervalo de confianza)

* Dada unha variable que estamos analizando, tomamos unha mostra de n=16 elementos. Nesta mostra, a varianza é de 25 e a media de 503. Calcular o intervalo tal que hai un 90% de probabilidade de que inclúa á media da poboación.
  + 1-a = 0.9 → a = 0.1.
  + Calculase z1-0.1/2= z0.95. mirase na tabla o valor de z que da p=0.95, aproximase a 1.645.
  + Ic = (503-1.645\*5/, 503+1.645\*5/) = (501.355, 504.645)
* Se non coñecemos a varianza, senon que só coñecemos que S=38.47, empregase a distribucion t de student (ten unha tabla distinta)
  + Trátase dunha t15, 0.95 . O valor da tabla correspondente é 1.753. (neste caso é ao reves, hai que consultar o valor da tabla na casilla indicada)
  + Ic =

**Intervalo de confianza para a varianza**

* Coñecemos que nS2/σ2 ∼ χ2n-1.. Este será o noso estadístico pivote.

| ) |
| --- |

**Lonxitude dun intervalo de confianza**

* Para calqueira intervalo excepto o da varianza[[7]](#footnote-6), o valor da lonxitude do Ic será 2\*(o extremo superior do intervalo).
* Despexando n na fórmula resultante, podemos calcular n para que un Ic teña unha lonxitude L. (ver fórmulas)

1. Se asume que n es conocido. [↑](#footnote-ref-0)
2. estas formulas danas no examen let’s fucking GOOOO [↑](#footnote-ref-1)
3. non confundir nivel de significación (α) con nivel de confianza (1-α). ‘Por ejemplo, si el nivel de confianza de un intervalo de confianza es del 95%, su nivel de significación es del 5%. Esto significa que si repetimos 100 veces el estudio estadístico, 95 veces obtendremos un resultado que coincide con el de la población real, mientras que 5 veces obtendremos un resultado erróneo.’ [↑](#footnote-ref-2)
4. Lembrar que p denota a proporción da poboación (descoñecida, queremos estimar o seu valor), e p^ denota a proporción muestral (coñecida) [↑](#footnote-ref-3)
5. en realidade, os p^ de dentro das raices serían p, pero como non o coñecemos asumimos que p^ é close enough [↑](#footnote-ref-4)
6. Xmedia é a x cunha barra arriba non sei escribila en gogoeld cdocs [↑](#footnote-ref-5)
7. Debido a que a distribución que emprega non é simétrica e as outras si. [↑](#footnote-ref-6)