

4. INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA Y DISTRIBUCIONES MUESTRALES

- 4.1. Introducción a la Inferencia.
 - 4.2. Muestreo. Tipos de muestreo.
 - 4.3. Distribuciones en el muestreo.
 - 4.4. Simulaciones informáticas .Remuestreo
-

Esquema de la inferencia estadística (mapa mental):

<https://www.uv.es/ceaces/intro/lexinfera.htm>

4.1. Introducción a la inferencia estadística.

Inferir es, en general, establecer un nuevo conocimiento partiendo de uno ya "dado".

Las inferencias suelen hacerse bien por **deducción** pasando de(l conocimiento de) lo general a (la obtención del conocimiento de) lo particular, conservando la certidumbre en cada paso de la deducción; o bien por **inducción** pasando de(l conocimiento de) lo particular a conclusiones generales que nunca podrán tener una certeza absoluta.

La inferencia estadística es un tipo particular de inferencia inductiva en la que las conclusiones generales aun no contando con certeza absoluta podrán expresarse con cierto grado de "incertidumbre" probabilística

La inferencia estadística va a ser, por lo tanto, una forma especial de realizar este proceso. Consiste, básicamente, en llegar a conocer algunas características desconocidas de una población partiendo de datos muestrales conocidos obtenidos aleatoriamente. Estas características poblacionales serán "inferidas" utilizando los recursos de la *TEORÍA MATEMÁTICA DE LA PROBABILIDAD* y finalmente, serán los principios del cálculo de probabilidades los que permitirán acotar la "incertidumbre" de las conclusiones

Fundamentalmente la Inferencia Estadística consiste en la resolución de dos grandes categorías de problemas:

□□□ **LA ESTIMACIÓN** : Aproximar, aventurar, "determinar" (entendiéndose como aproximación) el valor de una característica poblacional desconocida. Podrá ser:

- **Por punto**: Con la obtención de un valor estimado poblacional concreto
- **Por intervalo**: Con la obtención de un intervalo en el que quede incluido el valor de la característica con cierto grado de certeza (**Confianza**).

□□□ **EL CONTRASTE DE HIPOTESIS**: Determinar si es **aceptable (o no)**, partiendo de los datos muestrales, que la característica poblacional estudiada tome un valor determinado o bien que pertenezca a un intervalo de valores determinado.

(Es obvio que estos dos problemas de conocimiento pueden, muy bien, considerarse como dos tipos particulares de problemas de decisión estadística y así, de hecho, lo considera una de las escuelas metodológicas de la Estadística.)

4.2. Muestreo. Tipos de muestreo.

Entenderemos por muestreo el procedimiento o técnica empleada para obtener una muestra, que inicialmente consideramos como un cierto subconjunto de una población

POBLACION : Colectivo sujeto del estudio .

Cabe distinguir entre **Población** (colectivo en el que estamos considerando la magnitud sujeta a estudio) y **Universo** (colectivo de todos los elementos sujetos del estudio, en el que no consideramos la magnitud). El *universo* es, por tanto, el conjunto de individuos que poseen la característica o características sujetas a estudio, y éstas en su conjunto forman la *población*

Así ; analizando las estaturas de los españoles , la población sería el conjunto de todas las estaturas de todos los españoles , siendo el universo el conjunto de todos los españoles.

MUESTRA : Un subconjunto cualquiera de la población . Para que la muestra nos sirva para extraer conclusiones sobre la población deber ser **representativa** , lo que se consigue seleccionando sus elementos al azar , lo que da lugar a una muestra aleatoria

MUESTREO : Procedimiento para la obtención de una muestra

MUESTREO OPINATIVO : es aquel procedimiento de selección de los elementos muestrales que se realiza según el criterio subjetivo del investigador. Es, por tanto , subjetivo y la muestra obtenida puede no ser representativa de la población.

MUESTREO ALEATORIO :es aquel procedimiento de selección de la muestra en el que todos y cada uno de los elementos de la población tiene una cierta probabilidad de resultar elegidos . De esta forma, si tenemos una población de **N** elementos y estamos interesados en obtener una muestra de **n** elementos (muestra de **tamaño n**), cada subconjunto de **n** elementos de la población tendrá también una cierta probabilidad de resultar la muestra elegido.

Si designamos por M_i a cada uno de estos subconjuntos ,con $i= 1,2,3,...N$;

cada M_i tendrá una cierta probabilidad $P(M_i)$ de resultar elegido.

MUESTREO ALEATORIO SIMPLE: (M.A.S.): es aquel muestreo aleatorio en el que la probabilidad de que un elemento resulte seleccionado se mantiene constante a lo largo de todo el proceso de obtención de la misma. La técnica del muestreo puede asimilarse a un modelo de extracción de bolas de una urna con devolución (*reemplazamiento*) de la bola extraída. Un mismo dato, e incluso un mismo individuo, puede, en consecuencia, resultar muestreado más de una vez .Cada elección no dependerá de las anteriores y, por tanto, los datos muestrales serán **estocásticamente independientes**.

MUESTREO IRRESTRICTO (SIN REEMPLAZAMIENTO): en este tipo de muestreo la probabilidad de elegir/ extraer un individuo en cada selección viene influida por los resultados anteriores , en la medida en que en este muestreo no permitimos que un mismo individuo sea seleccionado más de una vez (lo que hace variar las probabilidades en cada extracción muestral) . Se corresponde con un modelo de extracción sin reemplazamiento .Teniendo en cuenta la convergencia de la distribución hipergeométrica a la binomial es fácil intuir que cuando la población sea muy grande ($N \gg n$) el muestreo irrestricto puede considerarse como muestreo aleatorio simple.

Por tanto , en el estudio de muestras para poblaciones grandes consideraremos sólo el muestreo simple .En el estudio de muestras de poblaciones finitas es, sin embargo , fundamental analizar las distribuciones muestrales que generará su adecuado muestreo irrestricto.

MUESTRA GENERICA DE TAMAÑO n : Es una **variable aleatoria**

n-dimensional ; $X=[x_1, x_2, x_3, ..., x_n]$ donde cada x_j (con $j=1,2,...,n$)

(cada dato muestral genérico) recorre todos los posibles valores que puede tomar el j-simo elemento de una muestra de **n** elementos.

Por tanto , una muestra concreta (ya obtenida) será un valor particular (una realización concreta) de la muestra genérica.

En la medida en que en el muestreo aleatorio cada elemento de la población tiene una probabilidad de ser elegido ,cada dato muestral genérico será una variable aleatoria que tendrá asociada una función de probabilidad $f(x)$ (de cuantía o de densidad) según una determinada distribución que llamaremos **distribución básica** , madre , o, simplemente, **distribución de la población** y recorrerá todos los posibles valores de la población.

Si trabajamos con un muestreo aleatorio simple (M.A.S.), cada dato muestral genérico será estocásticamente independiente de los demás y por tanto la función de probabilidad (cuantía o densidad) conjunta de la muestra genérica será :

$$f(x) = f(x_1, x_2, x_3, x_4, \dots, x_n) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot \dots \cdot f(x_n)$$

por ser las x_j variables aleatorias independientes.

Ejemplo 1:

Supongamos una *pequeña* población perfectamente conocida para poder ilustrar las ideas que vayamos introduciendo.

Supongamos que estudiamos un universo formada por 6 familias de las que consideramos la variable X =número de viajes que han realizado este año.

Supongamos que la distribución de esta población es:

X número de viajes realizados	Frecuencia (nº familias)	Frec.relativa/ probabilidad
0	1	1/6
1	1	1/6
2	2	2/6
3	1	1/6
4	1	1/6

Llamamos la atención para más adelante que : 1) la media de la población es 2 y 2) la varianza de la población es $10/6$.

Nos planteamos obtener una muestra aleatoria de tamaño 2 ($n=2$)

La variable Muestra genérica de tamaño 2, M , será una variable aleatoria bidimensional formada por datos muestrales genéricos X_1 y X_2

Tanto si se trata de MAS como de MI la distribución de probabilidad del primer dato muestral genérico, X_1 , será la propia distribución (de probabilidad) de la población

Sin embargo la distribución del segundo, X_2 ,variará en caso y otro:

En el caso de MAS: la distribución de X_2 seguirá siendo la distribución de la tabla de arriba.(X_1 y X_2 son estocásticamente independientes)

Sin embargo, en el caso de MI la distribución dependerá del resultado de X_1 .

Veámoslo:

MAS. Distribución de X_1 y X_2

X_1	$P(X_1)$		X_2	$P(X_2)$
0	1/6		0	1/6
1	1/6		1	1/6
2	2/6		2	2/6
3	1/6		3	1/6
4	1/6		4	1/6

MAS .Distribución de una muestra genérica de tamaño 2

$M=(X_1, X_2)$	$P(X_1, X_2)= P(X_1).P(X_2)$
(0,0)	$1/6.1/6= 1/36$
(0,1)	$1/6.1/6= 1/36$
(0,2)	$1/6.2/6= 2/36$
(0,3)	$1/6.1/6= 1/36$
(0,4)	$1/6.1/6= 1/36$
(1,0)	$1/6.1/6= 1/36$
(1,1)	$1/6.1/6= 1/36$
(1,2)	$1/6.2/6= 2/36$
(1,3)	$1/6.1/6= 1/36$
(1,4)	$1/6.1/6= 1/36$
(2,0)	$2/6.1/6= 2/36$
(2,1)	$2/6.1/6= 2/36$
(2,2)	$2/6.2/6= 4/36$
(2,3)	$2/6.1/6= 2/36$
(2,4)	$2/6.1/6= 2/36$
(3,0)	$1/6.1/6= 1/36$
(3,1)	$1/6.1/6= 1/36$
(3,2)	$1/6.2/6= 2/36$
(3,3)	$1/6.1/6= 1/36$
(3,4)	$1/6.1/6= 1/36$
(4,0)	$1/6.1/6= 1/36$

(4,1)	$1/6.1/6 = 1/36$
(4,2)	$1/6.2/6 = 2/36$
(4,3)	$1/6.1/6 = 1/36$
(4,4)	$1/6.1/6 = 1/36$
suma	$36/36 = 1$

M.I. Distribución de X_1 y X_2

X_1	$P(X_1)$	
0	1/6	
1	1/6	
2	2/6	
3	1/6	
4	1/6	

$X_2 X_1 = 0$	$P(X_2 X_1 = 0)$	$X_2 X_1 = 1$	$P(X_2 X_1 = 1)$	$X_2 X_1 = 2$	$P(X_2 X_1 = 2)$
0	0	0	1/5	0	1/5
1	1/5	1	0	1	1/5
2	2/5	2	2/5	2	1/5
3	1/5	3	1/5	3	1/5
4	1/5	4	1/5	4	1/5

$X_2 X_1 = 3$	$P(X_2 X_1 = 3)$	$X_2 X_1 = 4$	$P(X_2 X_1 = 4)$
0	1/5	0	1/5
1	1/5	1	1/5
2	2/5	2	2/5
3	0	3	1/5
4	1/5	4	0

MI .Distribución de una muestra genérica de tamaño 2

$M=(X_1, X_2)$	$P(X_1, X_2) = P(X_1) \cdot P(X_2 X_1)$
(0,0)	$1/6.0 = 0$
(0,1)	$1/6.1/5 = 1/30$
(0,2)	$1/6.2/5 = 2/30$
(0,3)	$1/6.1/5 = 1/30$
(0,4)	$1/6.1/5 = 1/30$
(1,0)	$1/6.1/5 = 1/30$
(1,1)	$1/6.0 = 0$
(1,2)	$1/6.2/5 = 2/30$
(1,3)	$1/6.1/5 = 1/30$
(1,4)	$1/6.1/5 = 1/30$
(2,0)	$2/6.1/5 = 2/30$
(2,1)	$2/6.1/5 = 2/30$
(2,2)	$2/6.1/5 = 1/30$
(2,3)	$2/6.1/5 = 2/30$
(2,4)	$2/6.1/5 = 2/30$
(3,0)	$1/6.1/5 = 1/30$
(3,1)	$1/6.1/5 = 1/30$
(3,2)	$1/6.2/5 = 2/30$
(3,3)	$1/6.0 = 0$
(3,4)	$1/6.1/5 = 1/30$
(4,0)	$1/6.1/5 = 1/30$
(4,1)	$1/6.1/5 = 1/30$
(4,2)	$1/6.2/5 = 2/30$
(4,3)	$1/6.1/5 = 1/30$
(4,4)	$1/6.0 = 0$
suma	$30/30 = 1$

ESTADÍSTICO : Es cualquier función de los valores muestrales que dependa exclusivamente de estos . En la medida en que los valores muestrales son variables aleatorias también lo serán las funciones de éstos: los estadísticos.

A modo de ejemplo podemos decir que son estadísticos la media muestral ,la varianza muestral , la cuasivarianza muestral , dado que son funciones de valores muestrales

exclusivamente y no sería estadístico la función $\frac{nS^2}{\sigma^2}$ que si bien contiene la varianza muestral, también depende de la poblacional, σ^2 , y por tanto no es función exclusiva de la muestra.

Como hemos visto, los estadísticos son variables aleatorias por lo que tendrán determinadas distribuciones de probabilidad y determinados parámetros (media, varianza, etc). Para el desarrollo de la inferencia es imprescindible conocer dichas distribuciones y parámetros, consiguiendo establecer entonces las relaciones entre éstas y las de la población, pudiendo entonces inferir las características desconocidas de ésta.

4.3. Distribuciones en el muestreo

4.3.1. Principales parámetros de la distribución de la media muestral (Población cualquiera)

Sea cualquier sea la distribución de la población la media muestral obtenida por M.A.S. tendrá una distribución cuya esperanza y varianza vendrán dadas por:

$$\begin{aligned} E(\bar{X}) &= \mu \\ \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \end{aligned}$$

En efecto:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n\right) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \dots + \frac{1}{n}E(X_n) = \\ &= \underbrace{\frac{1}{n}\mu + \frac{1}{n}\mu + \dots + \frac{1}{n}\mu}_{n \text{ veces}} = \mu \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \text{Var}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n\right) = \\ &= \underbrace{\frac{1}{n^2}\text{Var}(X_1) + \frac{1}{n^2}\text{Var}(X_2) + \dots + \frac{1}{n^2}\text{Var}(X_n)}_{\text{por ser independientes (MAS)}} = \underbrace{\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \dots + \frac{1}{n^2}\sigma^2}_{n \text{ veces}} = \frac{\sigma^2}{n} \end{aligned}$$

Si, el tamaño de la muestra **es muy grande** por el Teorema central del Límite podríamos, además afirmar la normalidad de la distribución:

$$\text{si } n \rightarrow \infty \quad \Rightarrow \quad \bar{X} \rightarrow N\left[\mu, \frac{\sigma}{\sqrt{n}}\right]$$

Si, se tratara de una muestreo sin remplazamiento la varianza se vería afectada por el coeficiente de exhaustividad o factor corrector para poblaciones finitas:

Si se trata de M.I.

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

y en caso de $n \rightarrow \infty \Rightarrow \bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right)$

Ejemplo 2: Desarrollamos un poco más el ejemplo 1 anterior:

En el caso de M.A.S. , a partir de la distribución de la muestra de tamaño 2 podemos construir la distribución de la media muestra (de una muestra con n=2)

M=(X ₁ ,X ₂)	P(X ₁ ,X ₂)= P(X ₁).P(X ₂)	$\bar{X}$	P($\bar{X}$)
(0,0)	1/6.1/6= 1/36	0	1/36
(0,1)	1/6.1/6= 1/36	0.5	1/36
(0,2)	1/6.2/6= 2/36	1	2/36
(0,3)	1/6.1/6= 1/36	1.5	1/36
(0,4)	1/6.1/6= 1/36	2	1/36
(1,0)	1/6.1/6= 1/36	0.5	1/36
(1,1)	1/6.1/6= 1/36	1	1/36
(1,2)	1/6.2/6= 2/36	1.5	2/36
(1,3)	1/6.1/6= 1/36	2	1/36
(1,4)	1/6.1/6= 1/36	2.5	1/36
(2,0)	2/6.1/6= 2/36	1	2/36
(2,1)	2/6.1/6= 2/36	1.5	2/36
(2,2)	2/6.2/6= 4/36	2	4/36
(2,3)	2/6.1/6= 2/36	2.5	2/36
(2,4)	2/6.1/6= 2/36	3	2/36
(3,0)	1/6.1/6= 1/36	1.5	1/36
(3,1)	1/6.1/6= 1/36	2	1/36
(3,2)	1/6.2/6= 2/36	2.5	2/36
(3,3)	1/6.1/6= 1/36	3	1/36
(3,4)	1/6.1/6= 1/36	3.5	1/36
(4,0)	1/6.1/6= 1/36	2	1/36
(4,1)	1/6.1/6= 1/36	2.5	1/36
(4,2)	1/6.2/6= 2/36	3	2/36
(4,3)	1/6.1/6= 1/36	3.5	1/36
(4,4)	1/6.1/6= 1/36	4	1/36
suma	36/36= 1		36/36= 1

Que convenientemente arreglada podría disponerse como:

$\bar{X}$	P($\bar{X}$)	$\bar{X} \cdot P(\bar{X})$
0	1/36	0
0.5	2/36	1/36
1	5/36	5/36
1.5	6/36	9/36
2	8/36	16/36
2.5	6/36	15/36
3	5/36	15/36
3.5	2/36	7/16
4	1/36	4/36
SUMA	1	E($\bar{X}$)=68/36=2

$$E[\bar{X}] = \sum \bar{X} \cdot P(\bar{X}) = 2 = \mu$$

Y la varianza:

$\bar{X}$	P($\bar{X}$)	$\bar{X}^2$	$\bar{X}^2 \cdot P(\bar{X})$
0	1/36	0	0

0.5	2/36	0.25	0.5/36
1	5/36	1	5/36
1.5	6/36	2.25	13.5/36
2	8/36	4	32/36
2.5	6/36	6.25	37.5/36
3	5/36	9	45/36
3.5	2/36	12.25	24.5/36
4	1/36	16	16/36
SUMA	1		$E(\bar{X}^2)=174/36$

$$Var[\bar{X}] = E(\bar{X}^2) - (E(\bar{X}))^2 = \frac{174}{36} - 4 = \frac{30}{36} = \frac{5}{6}$$

$$\frac{\sigma^2}{n} = \frac{10/6}{2} = \frac{10}{12} = \frac{5}{6}$$

En el caso del M.I.:

Partiendo de la distribución de la muestra de tamaño 2 (irrestricte) 2 podemos construir la distribución de la media muestra (de una muestra con n=2)

M=(X ₁ ,X ₂)	P(X ₁ ,X ₂)= P(X ₁).P(X ₂ X ₁)	$\bar{X}$	P($\bar{X}$)
(0,0)	1/6.0= 0	0	0
(0,1)	1/6.1/5= 1/30	0.5	1/30
(0,2)	1/6.2/5= 2/30	1	2/30
(0,3)	1/6.1/5= 1/30	1.5	1/30
(0,4)	1/6.1/5= 1/30	2	1/30
(1,0)	1/6.1/5= 1/30	0.5	1/30
(1,1)	1/6.0= 0	1	0
(1,2)	1/6.2/5= 2/30	1.5	2/30
(1,3)	1/6.1/5= 1/30	2	1/30
(1,4)	1/6.1/5= 1/30	2.5	1/30
(2,0)	2/6.1/5= 2/30	1	2/30
(2,1)	2/6.1/5= 2/30	1.5	2/30
(2,2)	2/6.1/5= 2/30	2	2/30
(2,3)	2/6.1/5= 2/30	2.5	2/30
(2,4)	2/6.1/5= 2/30	3	2/30
(3,0)	1/6.1/5= 1/30	1.5	1/30
(3,1)	1/6.1/5= 1/30	2	1/30
(3,2)	1/6.2/5= 2/30	2.5	2/30
(3,3)	1/6.0= 0	3	0
(3,4)	1/6.1/5= 1/30	3.5	1/30
(4,0)	1/6.1/5= 1/30	2	1/30
(4,1)	1/6.1/5= 1/30	2.5	1/30
(4,2)	1/6.2/5= 2/30	3	2/30
(4,3)	1/6.1/5= 1/30	3.5	1/30
(4,4)	1/6.0= 0	4	0
suma	30/30= 1		

Que convenientemente arreglada podría disponerse como:

$\bar{X}$	P($\bar{X}$)	$\bar{X} \cdot P(\bar{X})$
0	0	0
0.5	2/30	1/30
1	4/30	4/30
1.5	6/30	9/30
2	6/30	12/30
2.5	6/36	15/30
3	4/36	12/30
3.5	2/36	7/30
4	0	0
SUMA	1	$E(\bar{X})=60/30=2$

$$E[\bar{X}] = \sum \bar{X} \cdot P(\bar{X}) = 2 = \mu$$

Y la varianza

$\bar{X}$	$P(\bar{X})$	$\bar{X}^2$	$\bar{X}^2 P(\bar{X})$
0.5	2/30	0.25	0.5/30
1	4/30	1	4/30
1.5	6/30	2.25	13.5/30
2	6/30	4	24/30
2.5	6/30	6.25	37.5/30
3	4/30	9	36/30
3.5	2/30	12.25	24.5/30
SUMA	1		$E(\bar{X}^2)=140/30$

$$Var[\bar{X}] = E(\bar{X}^2) - (E(\bar{X}))^2 = \frac{140}{30} - 4 = \frac{20}{30} = \frac{2}{3}$$

$$\frac{\sigma^2}{n} \frac{N-n}{N-1} = \frac{10/6}{2} \left(\frac{6-2}{5} \right) = \frac{10}{12} \cdot \frac{4}{5} = \frac{2}{3}$$

4.3.2. Distribución de la varianza muestral (población cualquiera). [Esperanza]

La esperanza del estadístico varianza muestral procedente de una MAS sobre una población de cualquier distribución obedece a la expresión :

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

En efecto:

$$\begin{aligned} E(S^2) &= E \left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n} \right) = E \left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n} \right) = E \left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}^2 \right) = E \left(\sum_{i=1}^n \frac{X_i^2}{n} \right) - E(\bar{X}^2) = \\ &= \sum_{i=1}^n \frac{E(X_i^2)}{n} - E(\bar{X}^2) \quad (*)1 \end{aligned}$$

Ahora bien:

$$\text{Como } Var(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - E(\bar{X})^2 = E(\bar{X}^2) - \mu^2 \Rightarrow E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2 \quad (*)2$$

Y como

$$\forall i: Var(X_i) = \sigma^2 = E(X_i^2) - \mu^2 \Rightarrow E(X_i^2) = \sigma^2 + \mu^2 \text{ y por lo tanto:}$$

$$\sum_{i=1}^n \frac{E(X_i^2)}{n} = n \cdot \frac{(\sigma^2 + \mu^2)}{n} = \sigma^2 + \mu^2 \quad (*)3$$

Y sustituyendo (*)3 y (*)2 en (*)1:

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

Ejercicio 1. Comprobar el resultado en la situación del ejemplo 1 (con MAS) :

$M=(X_1, X_2)$	$P(X_1, X_2) = P(X_1) \cdot P(X_2)$	S^2	$P(S^2)$
----------------	-------------------------------------	-------	----------

(0,0)	1/6.1/6= 1/36	0	1/36
(0,1)	1/6.1/6= 1/36	0,25	1/36
(0,2)	1/6.2/6= 2/36	1	2/36
(0,3)	1/6.1/6= 1/36	2,25	1/36
(0,4)	1/6.1/6= 1/36	4	1/36
(1,0)	1/6.1/6= 1/36	0,25	1/36
(1,1)	1/6.1/6= 1/36	0	1/36
(1,2)	1/6.2/6= 2/36	0,25	2/36
(1,3)	1/6.1/6= 1/36	1	1/36
(1,4)	1/6.1/6= 1/36	2,25	1/36
(2,0)	2/6.1/6= 2/36	1	2/36
(2,1)	2/6.1/6= 2/36	0,25	2/36
(2,2)	2/6.2/6= 4/36	0	4/36
(2,3)	2/6.1/6= 2/36	0,25	2/36
(2,4)	2/6.1/6= 2/36	1	2/36
(3,0)	1/6.1/6= 1/36	2,25	1/36
(3,1)	1/6.1/6= 1/36	1	1/36
(3,2)	1/6.2/6= 2/36	0,25	2/36
(3,3)	1/6.1/6= 1/36	0	1/36
(3,4)	1/6.1/6= 1/36	0,25	1/36
(4,0)	1/6.1/6= 1/36	4	1/36
(4,1)	1/6.1/6= 1/36	2,25	1/36
(4,2)	1/6.2/6= 2/36	1	2/36
(4,3)	1/6.1/6= 1/36	0,25	1/36
(4,4)	1/6.1/6= 1/36	0	1/36
suma	36/36= 1		36/36= 1

Que convenientemente arreglada podría disponerse como:

S^2	$P(S^2)$	$S^2 P(S^2)$
0	8/36	0
0,25	12/36	3/36
1	10/36	10/36
2,25	4/36	9/36
4	2/36	8/36
suma	1	$E(S^2)=30/36=5/6$

$$E(S^2) = \left(\frac{n-1}{n} \right) \sigma^2 = \frac{2-1}{2} \cdot \frac{10}{6} = \frac{5}{6}$$

4.3.3. Distribución de la media muestral de una población normal

Si la muestra (supuesta M.A.S), $(X_1, X_2, \dots, X_n)$ proviene de una población normal de media μ y desviación típica σ , cada valor muestral X_i , tendrá esa distribución:

$X_i \rightarrow N(\mu, \sigma)$ y por lo tanto la media muestral, aplicando la linealidad de la Normal

tendrá una distribución normal con media μ y con desviación típica, $\frac{\sigma}{\sqrt{n}}$:

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Admitimos sin demostración que en el caso de tratarse de muestreo irrestricto también seguiría una distribución normal con la misma media y con la desviación típica afectada por el Factor corrector para poblaciones finitas:

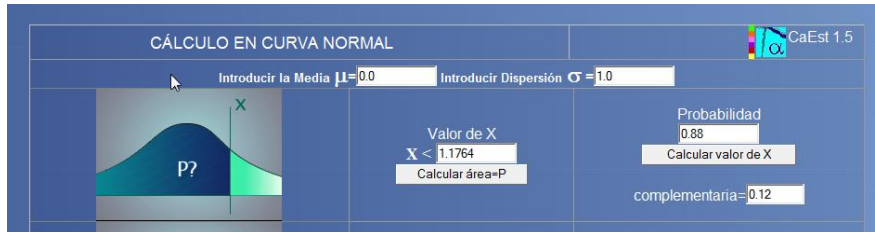
$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right)$$

Ejercicio 2. Si la estatura de una cierta población sigue una Normal de media 173 cm y desviación típica 17 cm y realizamos un muestreo de tamaño 100 ¿cuál es la probabilidad de que la media de las 100 personas muestreadas sea superior a 175?

$$X \rightarrow N(173, \sigma = 17)$$

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \bar{X} \rightarrow N\left(173, \frac{17}{\sqrt{100}}\right) \equiv N(173, 1.7)$$

$$P(\bar{X} > 175) = P\left(Z > \frac{175-173}{1.7}\right) = 1 - F_{N(0,1)}(1.1764) = 0.12$$



O en R:

```
pnorm(175,173,1.7,lower.tail = FALSE)
```

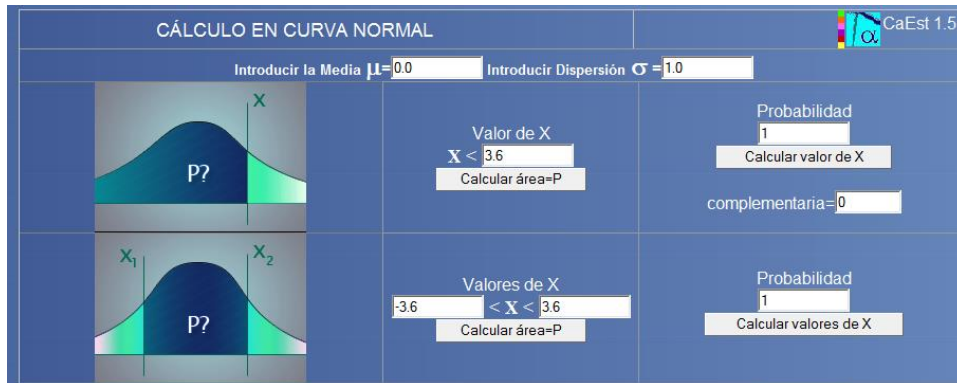
```
[1] 0.1197034
```

Ejercicio 3. Si llevamos a cabo un MAS de tamaño 144 en una normal de varianza poblacional 100 ¿cuál es la probabilidad de que la media muestral difiera de la media de la población (en valor absoluto) en más de 3 unidades?

$$X \rightarrow N(\mu, \sigma = 10)$$

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \bar{X} \rightarrow N\left(\mu, \frac{10}{\sqrt{144}}\right) \equiv N(\mu, 0.8333) \quad (\bar{X} - \mu) \rightarrow N(0, 0.8333)$$

$$\text{¿} P(|\bar{X} - \mu| > 3) \text{?} = P\left(|Z| > \frac{3-0}{0.8333}\right) = 2 \cdot (1 - F_{N(0,1)}(3.6)) = 0$$



O en R:

```
pnorm(-3,0,0.8333)+pnorm(3,0,0.8333,lower.tail = FALSE)
```

```
[1] 0.000318041
```

4.3.4. Distribución de la variable $\frac{nS^2}{\sigma^2}$ procedente de un M.A.S. sobre de una población normal.

En lugar de estudiar la distribución de la varianza muestral , va a resultar más conveniente de cara a las próximas aplicaciones inferenciales estudiar la distribución

de la variable $\frac{nS^2}{\sigma^2}$, que no es, propiamente un estadístico ya que depende de la varianza poblacional, σ^2 . Según el teorema de Cochran-Fischer, que no demostraremos aquí, la distribución de la variable $\frac{nS^2}{\sigma^2}$ procedente de un MAS sobre una población normal es estocásticamente independiente de la distribución de $\bar{X}$ y obedece a una distribución χ^2 con n-1 grados de libertad.

$$\frac{nS^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

Ejercicio 4. Sobre una población normal de varianza 150 se extrae una MAS de 30 individuos ¿Cual es la probabilidad de que la varianza muestral sobrepase el valor 160?

$$\frac{nS^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

$$\frac{30 \cdot S^2}{150} \rightarrow \chi^2_{29} \quad \frac{S^2}{5} \rightarrow \chi^2_{29} \quad P(S^2 > 160) = P\left(\frac{S^2}{5} > \frac{160}{5}\right) = P(\chi^2_{29} > 32) = 0.32$$

CÁLCULO PARA LA CHI-2			CaEst 1.5
Insertar si procede valor de chi-2	Insertar siempre grados de libertad	Insertar si procede valor de p	
32	29	0.320	
Calcular valor Chi-2		Calcular p	

O en R:

```
pchisq(32,29,lower.tail = FALSE)
[1] 0.3198733
```

4.3.5. Distribución de la variable $\frac{\bar{X} - \mu}{S} \sqrt{n-1}$ procedente de un M.A.S. sobre de una población normal.

Hemos visto más arriba cómo la media de una muestra procedente de una población normal tenía una distribución también normal centrada en la media poblacional

$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Aunque este resultado es muy conveniente de cara a las

aplicaciones inferenciales, ya que relaciona la distribución de la media muestral con el parámetro media de la población, presenta el problema de depender, también de la varianza o desviación típica poblacional. En un contexto en el que se pretenda realizar inferencias sobre la media poblacional, obviamente desconocida, cabe esperar que la varianza población también lo será.

Este inconveniente se resuelve mediante de la distribución de la variable (a veces

llamado impropriamente “estadístico t”) $\frac{\bar{X} - \mu}{S} \sqrt{n-1}$, cuya distribución será un de t

Student con n-1 grados de libertad.

En efecto:

Como por un lado $\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ y por otro : $\frac{nS^2}{\sigma^2} \rightarrow \chi^2_{n-1}$ y ambas variables son independientes ,resulta casi obvio que :

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{nS^2}{\sigma^2}}} \rightarrow t_{n-1} \text{ ya que } \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1) \text{ y el cociente entre una variable normal (0,1) y la raíz cuadrada de una variable Chi-dos dividida por sus g.l. es una t de Student.}$$

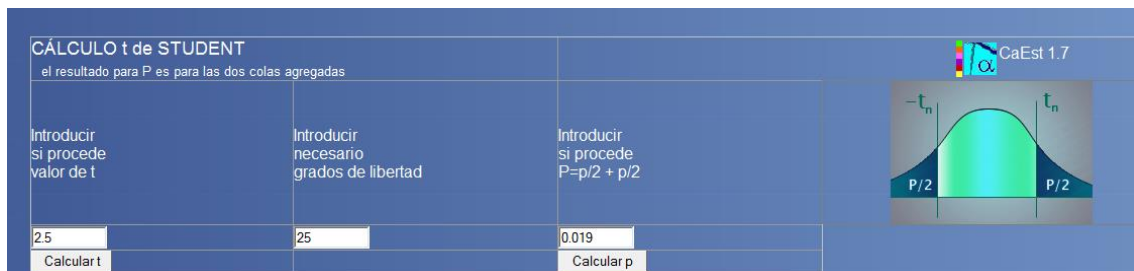
la raíz cuadrada de una variable Chi-dos dividida por sus g.l. es una t de Student.

Y finalmente la expresión se simplifica fácilmente

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{nS^2}{\sigma^2}}} = \frac{\bar{X} - \mu}{S} \sqrt{n-1}$$

Ejercicio 5. Sobre una población normal se ha realizado un MAS de tamaño 26 ¿Con qué probabilidad la media de esa muestra y la media de la población diferirán más de media desviación típica muestral?

$$\begin{aligned} &= P\left(\left|\bar{X} - \mu\right| > \frac{S}{2}\right) = P\left(\left|\frac{\bar{X} - \mu}{S}\right| > \frac{1}{2}\right) = P\left(\left|\frac{\bar{X} - \mu}{S} \cdot \sqrt{n-1}\right| > \frac{\sqrt{n-1}}{2}\right) = \\ &= P\left(\left|\frac{\bar{X} - \mu}{S} \cdot \sqrt{26-1}\right| > \frac{\sqrt{26-1}}{2}\right) = P(|t| > 2.5 | t \rightarrow t_{25}) = 0.019 \end{aligned}$$



O en R:

```
pt(2.5,25,lower.tail = FALSE)*2
# atención a las dos colas de la t de Student ( que recordemos que es simétrica)
[1] 0.01934313
```

4.3.4. Distribución de la diferencia de dos medias muestrales procedentes de dos poblaciones normales. (Suponemos 2 M.A.S. independientes)

Suponemos dos muestras aleatorias simples independientes de tamaños, en principio, diferentes n_x y n_y .

Es fácil ver que : $\bar{X} - \bar{Y} \rightarrow N\left(\mu_x - \mu_y; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$

Ya que $\bar{X} \rightarrow N\left(\mu_x, \frac{\sigma_x}{\sqrt{n_x}}\right)$ y $\bar{Y} \rightarrow N\left(\mu_y, \frac{\sigma_y}{\sqrt{n_y}}\right)$ y son independientes

De modo que aplicando la linealidad de la distribución normal el resultado es obvio.

(recordemos que trabajamos esta cuestión en la práctica 3)

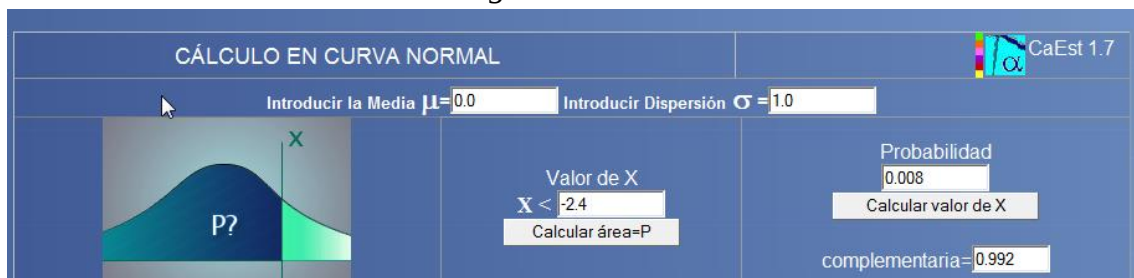
Ejemplo 6. Supongamos que la población masculina tiene una estatura $N(175,15)$ cm y la femenina $N(163,10)$ cm. Si extraemos al azar 15 hombres y 10 mujeres independientemente ¿Cuál será la probabilidad de que la media de las 10 mujeres supere al valor de la media de los 15 hombres?

$X \rightarrow N(175,15)$ se obtiene una M.A.S con $n_x = 15$

$Y \rightarrow N(163,10)$ se obtiene una M.A.S con $n_y = 10$

$$\bar{X} - \bar{Y} \rightarrow N\left(175 - 163; \sqrt{\frac{225}{15} + \frac{100}{10}}\right) \equiv N(12; 5)$$

$$P(\bar{Y} > \bar{X}) = P(\bar{X} - \bar{Y} < 0) = P\left(Z < \frac{0 - 12}{5}\right) = P(Z < -2.4) = 0.008$$



(Como -2.4 es un valor negativo en el gráfico superior en realidad habría que situarlo en su correspondiente lugar)

O en R:

```
pnorm(0,12,5)
[1] 0.008197536
```

4.3.5. Distribución de la variable aleatoria

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$$

En aquellas situaciones prácticas en las que se vayan a realizar inferencias sobre la diferencia de medias de dos poblaciones no suele ocurrir que las varianzas poblacionales resulten ser conocidas por lo que la distribución obtenida en el punto 4.3.4. no podrá ser utilizada más allá de algunos casos concretos o de ciertas aproximaciones.

En este contexto resulta útil, a veces, aplicar la distribución de la variable aleatoria de este epígrafe que sigue una distribución **t de Student** con n_x+n_y-2 grados de libertad, siempre que se cumpla que:

1.La varianza poblacional de ambas poblaciones sea la misma (o podamos suponerlo con cierta solvencia). Esta hipótesis llamada homoscedasticidad es bastante habitual en muchos modelos y aplicaciones estadísticas.

2.Los estadísticos muestrales (media y varianza muestrales de X y de Y) provenga de dos muestras aleatorias simples independientes.

Por lo tanto en estas condiciones:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{n_x + n_y} \cdot \sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{n_x \cdot n_y} \cdot \sqrt{n_x + n_y - 2} \rightarrow t_{n_x+n_y-2}$$

Más allá de la complicación de la expresión de esta variable aleatoria observemos, que (1) la diferencia de las dos medias muestrales seguirá una distribución Normal y que (2) cierto promedio ponderado de la sumas de cuadrados muestrales de X y de Y seguirá una distribución χ^2 . A partir de estos hechos se puede “construir” la variable de la que hablamos que parece razonable (no lo demostramos) que seguirá una **t de Student**.

Por último, comentar que la pérdida de dos grados de libertad se debe al hecho de que con los $n_x + n_y$ datos de las dos muestras se han tenido que calcular las dos medias muestrales y por lo tanto, no todos son libres (independientes a la hora que calcular el valor del promedio ponderado de la sumas de cuadrados muestrales).

Más allá, de nuevo, de la complejidad, de la expresión la importancia de esta distribución muestral es capital a la hora de estimar o contrastar hipótesis sobre las diferencias poblacionales de medias. Afortunadamente casi todas las herramientas de cálculo llevan incorporado un procedimiento “ t ” de prueba para cuando no es aplicable la distribución de 4.3.4. (varianzas conocidas – en la práctica tamaños muestrales grandes)

Cuando no se pueda asumir que las varianzas poblacionales son iguales existe una

aproximación t para la variable $\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n_x - 1} + \frac{S_y^2}{n_y - 1}}}$ que tendrá también una

distribución t con grados de libertad que vienen dados por una expresión aproximada que no reproducimos aquí pero que sí incorporan los procedimientos automáticos de cálculo.

Ejemplo 7. Consideremos la situación del ejemplo 6 pero ahora tengamos en cuenta que no son conocidas las varianzas poblacionales. Pero que suponemos que ambas son (o son aproximadamente) iguales . Por lo tanto:

Supongamos que la población masculina tiene una estatura $N(175, \sigma)$ cm y la femenina $N(163, \sigma)$ cm. (Con σ desconocida) Si extraemos al azar 15 hombres y 10 mujeres independientemente y ¿Cuál será la probabilidad de que la media de las 10 mujeres supere al valor de la media de los 15 hombres si las varianzas muestrales las suponemos iguales a 225 para los hombres y 100 para las mujeres ?

Nota: resulta muy poco realista desconocer los valores de la medias muestrales pero sí suponer los de las varianzas muestrales. Se trata ,obviamente, de un supuesto pedagógico.

$X \rightarrow N(175, \sigma)$ se obtiene una M.A.S con $n_x = 15$ y $S_x^2 = 225$

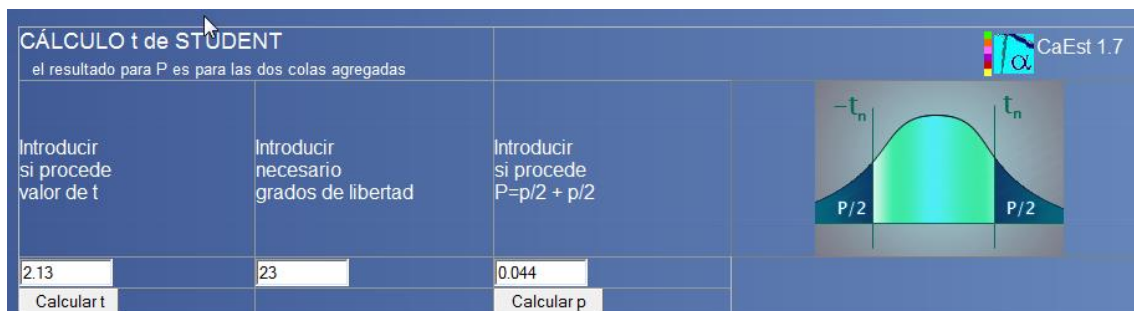
$Y \rightarrow N(163, \sigma)$ se obtiene una M.A.S con $n_y = 10$ y $S_y^2 = 100$

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{n_x + n_y} \cdot \sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{n_x \cdot n_y} \cdot \sqrt{n_x + n_y - 2} \rightarrow t_{n_x + n_y - 2} \text{ se traduce en :}$$

$$\frac{\bar{X} - \bar{Y} - (175 - 163)}{\sqrt{15 + 10} \cdot \sqrt{15 \cdot 225 + 10 \cdot 100}} \sqrt{15 \cdot 10} \cdot \sqrt{15 + 10 - 2} \rightarrow t_{15 + 10 - 2}$$

$$\frac{\bar{X} - \bar{Y} - (12)}{5.63} \rightarrow t_{23}$$

$$P(\bar{Y} > \bar{X}) = P(\bar{X} - \bar{Y} < 0) = P\left(\frac{\bar{X} - \bar{Y} - 12}{5.63} < \frac{0 - 12}{5.63}\right) = P(t_{23} < -2.13) = 0.022 \mid \dots [0.022 = 0.044 / 2]$$



O en R:

`pt(-2.13, 23)`

`[1] 0.02204537`

4.3.6. Distribución de una proporción muestral

Supongamos que queremos estudiar la proporción, p , con que se da una cierta característica, A , en una población. Para ello procedemos estadísticamente y realizamos una MAS de tamaño n . (Si la población fuera muy grande no sería necesario suponer la devolución)

Si llamamos X = número de individuos con la característica A en la muestra de tamaño n : $X \rightarrow B(n, p)$

Pero si la muestra es lo suficientemente grande podemos aplicar el Teorema de Moivre y , por lo tanto: $X \rightarrow N(np, \sqrt{npq})$

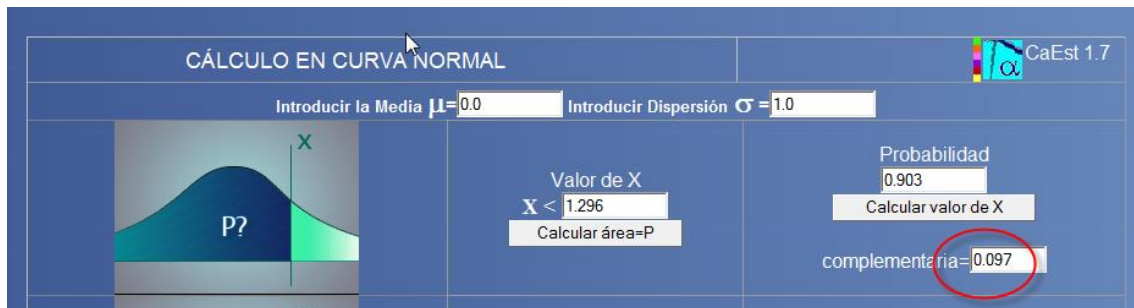
Si consideramos, ahora, el estadístico, $\hat{p} = \frac{X}{n} \equiv$ proporción muestral, por la linealidad de la D. Normal:

$$\hat{p} = \frac{X}{n} \rightarrow N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right) \text{ esto es : } \hat{p} = \frac{X}{n} \rightarrow N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Ejemplo 8: Supongamos que 43% de los estudiantes de la UV van a clase en metro. Si elegimos 100 estudiante al azar mediante MAS cual es la probabilidad de que vengan en metro más de la mitad (50%)

$$\hat{p} \rightarrow N\left(p, \sqrt{\frac{pq}{n}}\right) \equiv N\left(0.43, \sqrt{\frac{0.43 \cdot 0.67}{100}}\right) \equiv N(0.43; 0.054)$$

$$P(\hat{p} > 0.5) = P\left(Z > \frac{0.5 - 0.43}{0.054}\right) = P(Z > 1.296) = 0.097$$



O en R:

```
pnorm(0.5,0.43,0.054, lower.tail = FALSE)
[1] 0.09743671
```

4.3.7. Distribución de la diferencia de dos proporciones muestrales

Realizadas dos muestras **independientes** (MAS) de tamaños n_x y n_y la diferencia de las proporciones muestrales de las características X e Y se distribuirán según una normal. Resultado casi trivial aplicando las propiedades lineales de la D.normal:

$$\hat{p}_x - \hat{p}_y \rightarrow N\left(p_x - p_y, \sqrt{\frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}}\right)$$

4.3.8. Distribución para el cociente de la suma de cuadrados muestrales de dos MAS independientes de dos poblaciones normales con igual varianza (poblacional)

Partimos de dos poblaciones Normales $X \rightarrow N(\mu_x, \sigma_x)$ e $Y \rightarrow N(\mu_y, \sigma_y)$ Y consideramos dos MAS independientes de tamaños n_x e n_y

Nos es difícil probar que

$$\frac{n_x S_x^2}{n_y S_y^2} \frac{n_y - 1}{n_x - 1} \rightarrow F_{n_x - 1, n_y - 1}$$

Teniendo en cuenta la definición de la F de Snedecor y la distribución estudiada en 4.3.4.

ESQUEMA DE LAS DISTRIBUCIONES DE LOS PRINCIPALES ESTADÍSTICOS (<https://www.uv.es/ceaces/tex1t/3%20infemues/esquema.htm>)

