Bioestadística

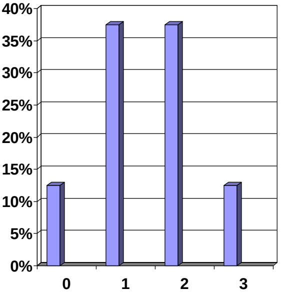
Tema 5: Modelos probabilísticos

Variable aleatoria

- El resultado de un experimento aleatorio puede ser descrito en ocasiones como una cantidad numérica.
- En estos casos aparece la noción de variable aleatoria
  - Función que asigna a cada suceso un número.
- Las variables aleatorias pueden ser discretas o continuas (como en el primer tema del curso).
- En las siguientes transparencias vamos a recordar conceptos de temas anteriores, junto con su nueva designación. Los nombres son nuevos. Los conceptos no.

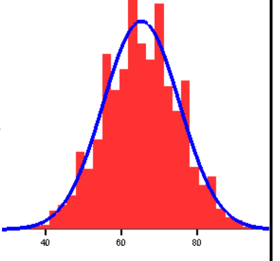
Función de probabilidad (V. Discretas)

- Asigna a cada posible valor de una variable discreta su probabilidad.
  - Recuerda los conceptos de frecuencia relativa y diagrama de barras.
- Ejemplo
  - Número de caras al lanzar 3 monedas.



Función de densidad (V. Continuas)

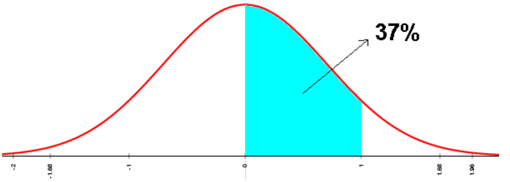
- Definición
  - Es una función no negativa de integral 1.
    - Piénsalo como la generalización del histograma con frecuencias relativas para variables continuas.



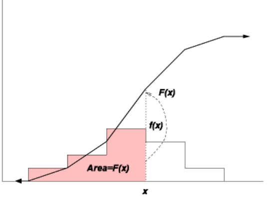
- ¿Para qué lo voy a usar?
  - Nunca lo vas a usar directamente.
  - Sus valores no representan probabilidades.

¿Para qué sirve la f. densidad?

- Muchos procesos aleatorios vienen descritos por variables de forma que son conocidas las probabilidades en intervalos.
- La integral definida de la función de densidad en dichos intervalos coincide con la probabilidad de los mismos.
- Es decir, identificamos la probabilidad de un intervalo con el **área** bajo la función de densidad.



Función de distribución



- Es la función que asocia a cada valor de una variable, la probabilidad acumulada de los valores inferiores o iguales.
  - Piénsalo como la generalización de las frecuencias acumuladas. Diagrama integral.
    - A los valores extremadamente bajos les corresponden valores de la función de distribución cercanos a cero.
    - A los valores extremadamente altos les corresponden valores de la función de distribución cercanos a uno.
- Lo encontraremos en los artículos y aplicaciones en forma de “p-valor”, significación,…
  - No le deis más importancia a este comentario ahora. Ya os irá sonando conforme avancemos.

¿Para qué sirve la f. distribución?

- Contrastar lo anómalo de una observación concreta.
    - Sé que una persona de altura 210cm es "anómala" porque la función de distribución en 210 es muy alta.
    - Sé que una persona adulta que mida menos de 140cm es "anómala" porque la función de distribución es muy baja para 140cm.
    - Sé que una persona que mida 170cm no posee una altura nada extraña pues su función de distribución es aproximadamente 0,5.
- Relaciónalo con la idea de cuantil.
- En otro contexto (contrastes de hipótesis) podremos observar unos resultados experimentales y contrastar lo "anómalos" que son en conjunto con respecto a una hipótesis de terminada.
    - Intenta comprender la explicación de clase si puedes. Si no, ignora esto de momento. Revisita este punto cuando hayamos visto el tema de contrastes de hipótesis.

Valor esperado y varianza de una v.a. X

- Valor esperado
    - Se representa mediante E[X] ó μ
    - Es el equivalente a la media
        - Más detalles: Ver libro.
- Varianza
    - Se representa mediante VAR[X] o σ^2
    - Es el equivalente a la varianza
    - Se llama desviación típica a σ
        - Más detalles: Ver libro.

Distribución normal o de Gauss

- Aparece de manera natural:
  - Errores de medida.
  - Distancia de frenado.
  - Altura, peso, propensión al crimen…
  - Distribuciones binomiales con n grande (n>30) y ‘p ni pequeño’ (np>5) ‘ni grande’ (nq>5).

- Está caracterizada por dos parámetros: La media, μ, y la desviación típica, σ.

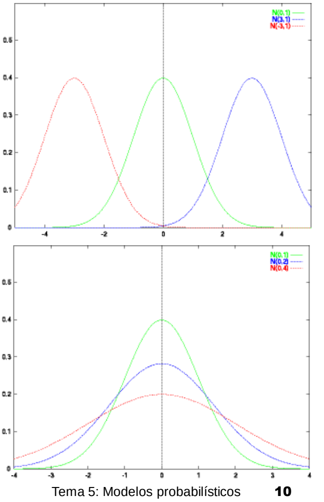
- Su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

N(μ, σ): Interpretación geométrica

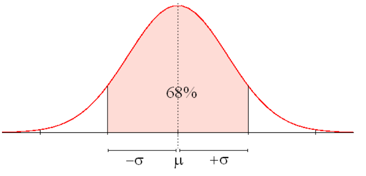
- Podéis interpretar la media como un factor de traslación.

- Y la desviación típica como un factor de escala, grado de dispersión,…

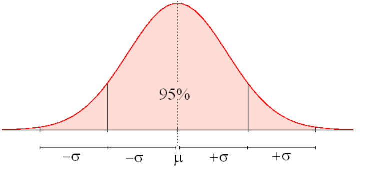


N(μ, σ): Interpretación probabilista

- Entre la media y una desviación típica tenemos siempre la misma probabilidad: aprox. 68%



- Entre la media y dos desviaciones típicas aprox. 95%



Algunas características

- La función de densidad es simétrica, mesocúrtica y unimodal.
  - Media, mediana y moda coinciden.
- Los puntos de inflexión de la fun. de densidad están a distancia **σ** de μ.
- Si tomamos intervalos centrados en μ, y cuyos extremos están…
  - a distancia **σ**, → tenemos probabilidad **68%**
  - a distancia **2 σ**, → tenemos probabilidad **95%**
  - a distancia **2'5 σ** → tenemos probabilidad **99%**
- No es posible calcular la probabilidad de un intervalo simplemente usando la primitiva de la función de densidad, ya que no tiene primitiva expresable en términos de funciones 'comunes'.
- Todas las distribuciones normales N(μ, σ), pueden ponerse mediante una traslación μ, y un cambio de escala σ, como **N(0,1)**. Esta distribución especial se llama **normal tipificada**.
  - Justifica la técnica de tipificación, cuando intentamos comparar individuos diferentes obtenidos de sendas poblaciones normales.

Tipificación

- Dada una variable de media μ y desviación típica σ, se denomina valor tipificado,z, de una observación x, a la distancia (con signo) con respecto a la media, medido en desviaciones típicas, es decir

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

- En el caso de variable X normal, la interpretación es clara: Asigna a todo valor de N(μ, σ), un valor de N(0,1) que deja **exáctamente** la misma probabilidad por debajo.

- Nos permite así comparar entre dos valores de dos distribuciones normales diferentes, para saber cuál de los dos es más extremo.

Tabla N(0,1)

normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

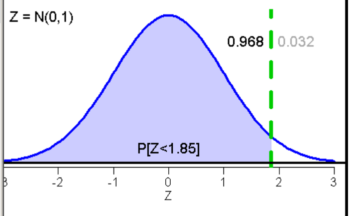
A1 Función de distribución (acumulativa) de la distribución normal tipificada.

	A	B	C	D	E	F	G	H	I	J	K
3		0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
4	0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
21	1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
22	1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
23	1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
24	2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
25	2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
26	2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
27	2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
28	2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
29	2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995

Normal_0_1

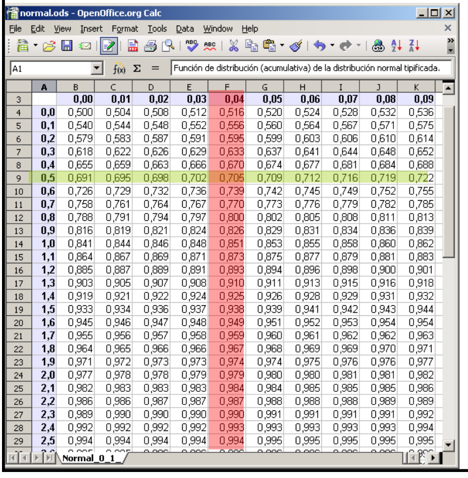
Z es normal tipificada.

Calcular P[Z<1,85]



Solución: 0,968 = 96,8%

Tabla N(0,1)



normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

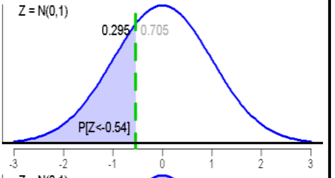
A1 | Función de distribución (acumulativa) de la distribución normal tipificada.

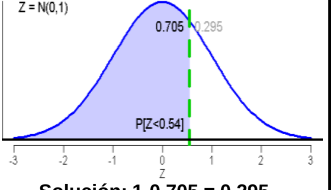
	A	B	C	D	E	F	G	H	I	J	K
3		0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
4	0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
21	1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
22	1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
23	1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
24	2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
25	2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
26	2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
27	2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
28	2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
29	2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995

Normal_0_1_

Z es normal tipificada.

Calcular P[Z<-0,54]



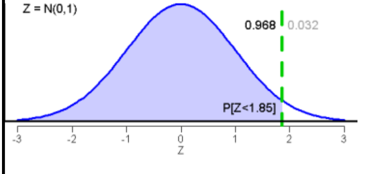


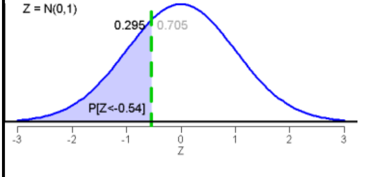
Solución: 1-0,705 = 0,295

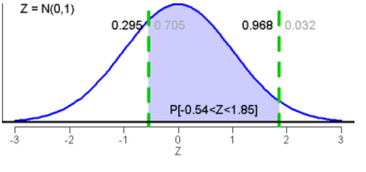
Tabla N(0,1)

Z es normal tipificada.

Calcular P[-0,54<Z<1,85]





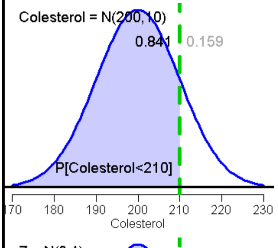


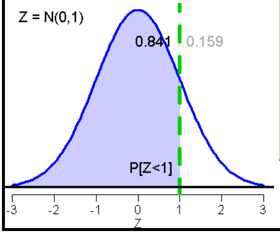
Solución: 0,968-0,295= 0,673

Ejemplo: Cálculo con probabilidades normales

- El colesterol en la población tiene distribución normal, con media 200 y desviación 10.
- ¿Qué porcentaje de indivíduos tiene colesterol inferior a 210?
- Qué valor del colesterol sólo es superado por el 10% de los individuos.

- Todas las distribuciones normales son similares salvo traslación y cambio de escala: Tipifiquemos.





$$z = \frac{x-\mu}{\sigma} = \frac{210-200}{10} = 1$$

normal.ods - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

A1 | Función de distribución (acumulativa) de la distribución normal tipificada.

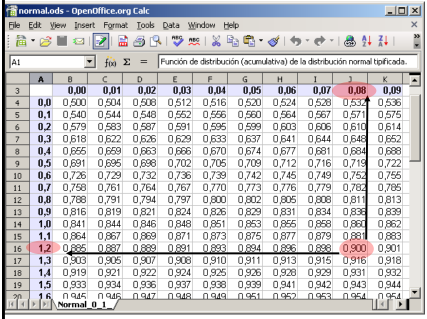
	A	B	C	D	E	F	G	H	I	J	K
3		**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
4	**0,0**	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	**0,1**	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	**0,2**	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	**0,3**	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	**0,4**	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	**0,5**	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	**0,6**	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	**0,7**	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	**0,8**	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	**0,9**	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	**1,0**	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	**1,1**	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883

Normal_0_1_

Sheet 1 / 1 | PageStyle_Normal_0_1_ | 100% | STD | Sum=0

$$P[Z < 1{,}00] = (\text{ver tabla}) = 0{,}841$$

- El valor del colesterol que sólo supera el 10% de los individuos es el percentil 90. Calculemos el percentil 90 de la N(0,1) y deshacemos la tipificación.

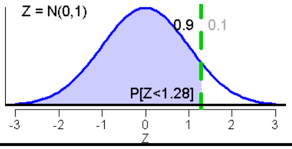


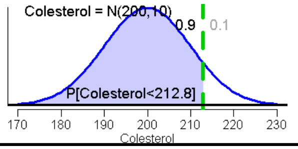
	A	B	C	D	E	F	G	H	I	J	K
3		**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
4	**0,0**	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
5	**0,1**	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
6	**0,2**	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
7	**0,3**	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
8	**0,4**	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
9	**0,5**	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
10	**0,6**	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
11	**0,7**	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
12	**0,8**	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
13	**0,9**	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
14	**1,0**	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
15	**1,1**	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
16	**1,2**	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
17	**1,3**	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
18	**1,4**	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
19	**1,5**	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
20	**1,6**	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954

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

$$1{,}28 = \frac{x-200}{10}$$

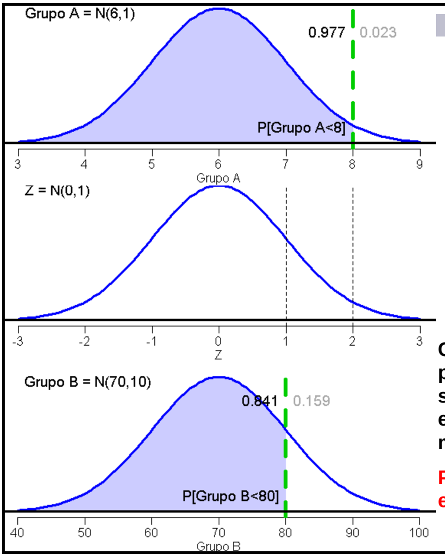
$$x = 200 + 10 \times 1{,}28 = 212{,}8$$





Ejemplo: Tipificación

- Se quiere dar una beca a uno de dos estudiantes de sistemas educativos diferentes. Se asignará al que tenga mejor expediente académico.
  - El estudiante A tiene una calificación de 8 en un sistema donde la calificación de los alumnos se comporta como N(6,1).
  - El estudiante B tiene una calificación de 80 en un sistema donde la calificación de los alumnos se comporta como N(70,10).
- **Solución**
  - No podemos comparar directamente 8 puntos de A frente a los 80 de B, pero como ambas poblaciones se comportan de modo normal, podemos tipificar y observar las puntuaciones sobre una distribución de referencia N(0,1)



$$z_A = \frac{x_A - \mu_A}{\sigma_A} = \frac{8-6}{1} = 2$$

$$z_B = \frac{x_B - \mu_B}{\sigma_B} = \frac{80-70}{10} = 1$$

Como $Z_A > Z_B$, podemos decir que el porcentaje de compañeros del mismo sistema de estudios que ha superado en calificación el estudiante A es mayor que el que ha superado B.

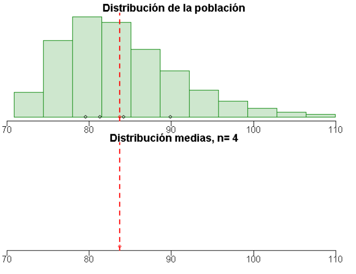
Podríamos pensar en principio que A es mejor candidato para la beca.

¿Por qué es importante la distribución normal?

- Las propiedades que tiene la distribución normal son interesantes, pero todavía no hemos hablado de por qué es una distribución especialmente importante.
- La razón es que aunque una v.a. no posea distribución normal, ciertos estadísticos/estimadores calculados sobre muestras elegidas al azar sí que poseen una distribución normal.
- Es decir, tengan las distribución que tengan nuestros datos, los 'objetos' que resumen la información de una muestra, posiblemente tengan distribución normal (o asociada).

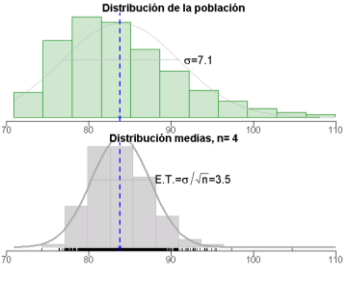
Aplic. de la normal: Estimación en muestras

- Como ilustración mostramos una variable que presenta valores distribuidos de forma muy asimétrica. Claramente no normal.
- Saquemos muestras de diferentes tamaños, y usemos la media de cada muestra para estimar la media de la población.



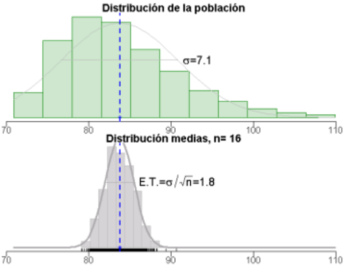
Aplic. de la normal: Estimación en muestras

- Cada muestra ofrece un resultado diferente: La media muestral es variable aleatoria.
- Su distribución es más parecida a la normal que la original.
- También está menos dispersa. A su dispersión ('desv. típica del estimador media muestral'… ¿os gusta el nombre largo?) se le suele denominar error típico.



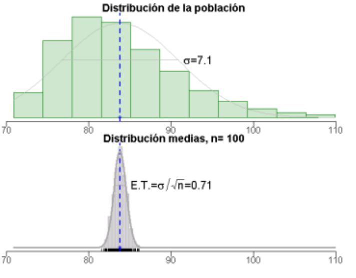
Aplic. de la normal: Estimación en muestras

- Al aumentar el tamaño, n, de la muestra:
  - La normalidad de las estimaciones mejora
  - El error típico disminuye.



Aplic. de la normal: Estimación en muestras

- Puedo 'garantizar' medias muestrales tan cercanas como quiera a la verdadera media, sin más que tomar 'n bastante grande'
- Se utiliza esta propiedad para dimensionar el tamaño de una muestra antes de empezar una investigación.



Resumen: Teorema del límite central

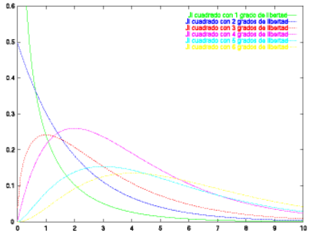
- Dada una v.a. **cualquiera**, si extraemos muestras de tamaño n, y calculamos los promedios muestrales, entonces:
  - dichos promedios tienen distribución aproximadamente normal;
  - La media de los promedios muestrales es la misma que la de la variable original.
  - La desviación típica de los promedios disminuye en un factor "raíz de n" (error estándar).
  - Las aproximaciones anteriores se hacen exactas cuando n tiende a infinito.
- Este teorema justifica la importancia de la distribución normal.
  - Sea lo que sea lo que midamos, cuando se promedie sobre una muestra grande (**n>30**) nos va a aparecer de manera natural la distribución normal.

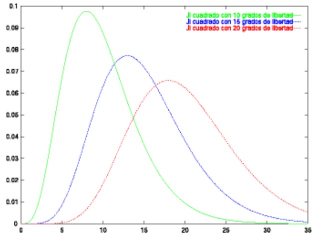
Distribuciones asociadas a la normal

- Cuando queramos hacer inferencia estadística hemos visto que la distribución normal aparece de forma casi inevitable.
- Dependiendo del problema, podemos encontrar otras (asociadas):
  - X^2 (chi cuadrado)
  - t- student
  - F-Snedecor
- Estas distribuciones resultan directamente de operar con distribuciones normales. Típicamente aparecen como distribuciones de ciertos estadísticos.
- Veamos algunas propiedades que tienen (superficialmente). Para más detalles consultad el manual.
- Sobre todo nos interesa saber qué valores de dichas distribuciones son "atípicos".
  - Significación, p-valores,…

Chi cuadrado

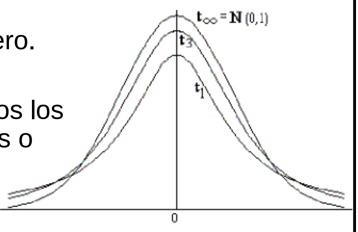
- Tiene un sólo parámetro denominado grados de libertad.
- La función de densidad es asimétrica positiva. Sólo tienen densidad los valores positivos.
- La función de densidad se hace más simétrica incluso casi gausiana cuando aumenta el número de grados de libertad.
- Normalmente consideraremos anómalos aquellos valores de la variable de la "cola de la derecha".





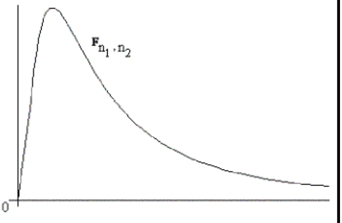
T de student

- Tiene un parámetro denominado grados de libertad.
- Cuando aumentan los grados de libertad, más se acerca a N(0,1).
- Es simétrica con respecto al cero.
- Se consideran valores anómalos los que se alejan de cero (positivos o negativos).



F de Snedecor

- Tiene dos parámetros denominados grados de libertad.
- Sólo toma valores positivos. Es asimétrica.
- Normalmente se consideran valores anómalos los de la cola de la derecha.



¿Qué hemos visto?

- En v.a. hay conceptos equivalentes a los de temas anteriores
  - Función de probabilidad ⇔ Frec. Relativa.
  - Función de densidad ⇔ histograma
  - Función de distribución ⇔ diagr. Integral.
  - Valor esperado ⇔ media, …
- Modelos de v.a. de especial importancia:
  - Normal
    - Propiedades geométricas
    - Tipificación
    - Aparece tanto en problemas con variables cualitativas (dicotómicas, Bernoulli) como numéricas
    - Distribuciones asociadas
      - T-student
      - X2
      - F de Snedecor