Bioestadística

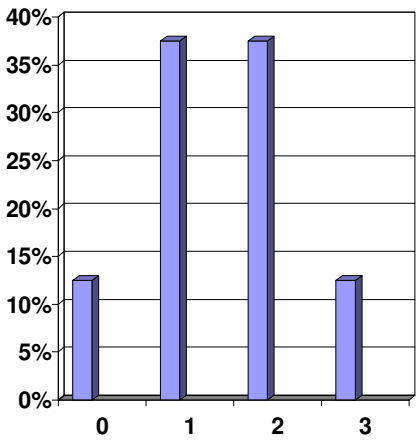
Tema 5: Modelos probabilísticos

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.

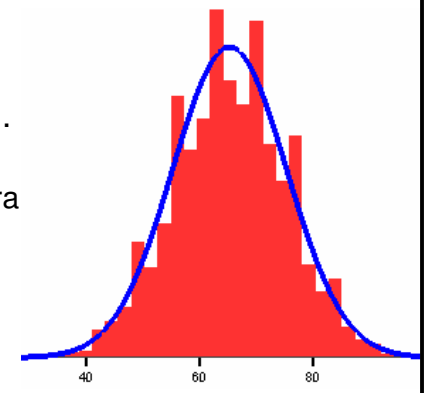
Función de probabilidad (V. Discretas)

- 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)

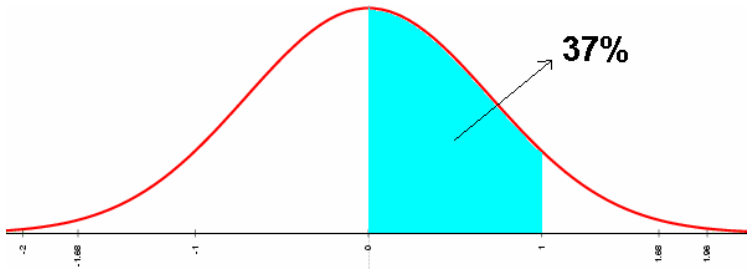
- Definición
  - Es una función no negativa de integral 1.
    - Piénsalo como la generalización del histograma con frecuencias relativas para variables continuas.



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

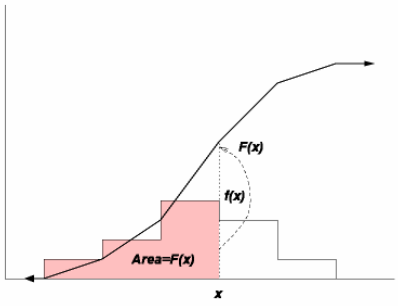
¿Para qué sirve la f. densidad?

- 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.

Algunos modelos de v.a.

- Hay v.a. que aparecen con frecuencia en las Ciencias de la Salud.
  - Experimentos dicotómicos.
    - Bernoulli
  - Contar éxitos en experimentos dicotómicos repetidos:
    - Binomial
    - Poisson (sucesos raros)
  - Y en otras muchas ocasiones…
    - Distribución normal (gaussiana, campana,…)
- El resto del tema está dedicado a estudiar estas distribuciones especiales.

Distribución de Bernoulli

- Tenemos un experimento de Bernoulli si al realizar un experimentos sólo son posibles dos resultados:
  - X=1 (**éxito**, con probabilidad p)
  - X=0 (**fracaso**, con probabilidad q=1-p)
    - Lanzar una moneda y que salga cara.
      - p=1/2
    - Elegir una persona de la población y que esté enfermo.
      - p=1/1000 = prevalencia de la enfermedad
    - Aplicar un tratamiento a un enfermo y que éste se cure.
      - p=95%, probabilidad de que el individuo se cure
- Como se aprecia, en experimentos donde el resultado es dicotómico, la variable queda perfectamente determinada conociendo el parámetro **p**.

Ejemplo de distribución de Bernoulli.

- Se ha observado estudiando 2000 accidentes de tráfico con impacto frontal y cuyos conductores no tenían cinturón de seguridad, que 300 individuos quedaron con secuelas. Describa el experimento usando conceptos de v.a.

- Solución.
  - La noc. frecuentista de prob. nos permite aproximar la probabilidad de tener secuelas mediante 300/2000=0,15=15%
  - X="tener secuelas tras accidente sin cinturón" es variable de Bernoulli
    - X=1 tiene probabilidad p ≈ 0,15
    - X=0 tiene probabilidad q ≈ 0,85

Ejemplo de distribución de Bernoulli.

- Se ha observado estudiando 2000 accidentes de tráfico con impacto frontal y cuyos conductores sí tenían cinturón de seguridad, que 10 individuos quedaron con secuelas. Describa el experimento usando conceptos de v.a.

- Solución.
  - La noc. frecuentista de prob. nos permite aproximar la probabilidad de quedar con secuelas por 10/2000=0,005=0,5%
  - X="tener secuelas tras accidente usando cinturón" es variable de Bernoulli
    - X=1 tiene probabilidad p ≈ 0,005
    - X=0 tiene probabilidad q ≈ 0,995

Observación

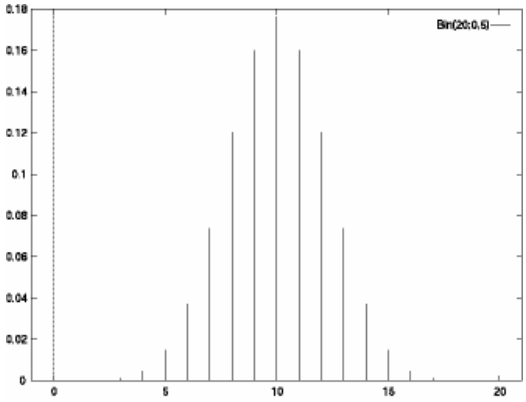
- En los dos ejemplos anteriores hemos visto cómo enunciar los resultados de un experimento en forma de estimación de parámetros en distribuciones de Bernoulli.
  - Sin cinturón: p ≈ 15%
  - Con cinturón: p ≈ 0,5%
- En realidad no sabemos en este punto si ambas cantidades son muy diferentes o aproximadamente iguales, pues en otros estudios sobre accidentes, las cantidades de individuos con secuelas hubieran sido con seguridad diferentes.
- Para decidir si entre ambas cantidades existen diferencias estadísticamente significativas necesitamos introducir conceptos de estadística inferencial (extrapolar resultados de una muestra a toda la población).
- Es muy pronto para resolver esta cuestión ahora. Esperemos a las pruebas de X^2.

Distribución binomial

- Función de probabilidad

$$P[X=k]=\binom{n}{k}p^k q^{n-k},\ 0\le k\le n$$

  - Problemas de cálculo si n es grande y/o p cercano a 0 o 1.



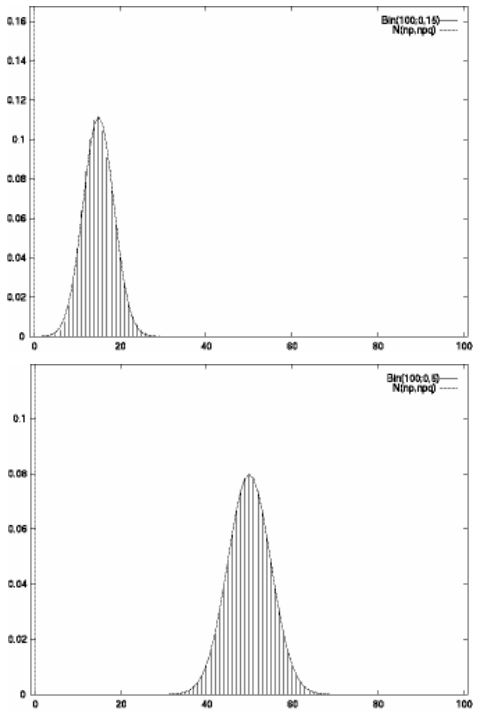
- Media: $\mu = n\,p$
- Varianza: $\sigma^2 = n\,p\,q$

Distribución Binomial

- Si se repite un número fijo de veces, **n**, un experimento de Bernoulli con parámetro **p**, el número de éxitos sigue una distribución **binomial** de parámetros (n,p).
  - Lanzar una moneda 10 veces y contar las caras.
    - Bin(n=10,p=1/2)
  - Lanzar una moneda 100 veces y contar las caras.
    - Bin(n=100,p=1/2)
    - Difícil hacer cálculos con esas cantidades. El modelo normal será más adecuado.
  - El número de personas que enfermará (en una población de 500.000 personas) de una enfermedad que desarrolla una de cada 2000 personas.
    - Bin(n=500.000, p=1/2000)
      - Difícil hacer cálculos con esas cantidades. El modelo de Poisson será más adecuado.

"Parecidos razonables"

- Aún no conocéis la distribución normal, ni de Poisson.
- De cualquier forma ahí tenéis la comparación entre valores de p no muy extremos y una normal de misma media y desviación típica, para tamaños de n grandes (n>30).
- Cuando p es muy pequeño es mejor usar la aproximación del modelo de Poisson.



Distribución de Poisson

- También se denomina de sucesos raros.
- Se obtiene como aproximación de una distribución binomial con la misma media, para 'n grande' (n>30) y 'p pequeño' (p<0,1).
- Queda caracterizada por un único parámetro μ (que es a su vez su media y varianza.)
- Función de probabilidad:

$$P[X=k]=e^{-\mu}\frac{\mu^k}{k!},\ k=0,1,2,...$$

Ejemplos de variables de Poisson

- El número de individuos que será atendido un día cualquiera en el servicio de urgencias del hospital clínico universitario.
  - En Málaga hay 500.000 habitantes (n grande)
  - La probabilidad de que cualquier persona tenga un accidente es pequeña, pero no nula. Supongamos que es 1/10.000
  - **Bin**(n=500.000,p=1/10.000) ≈ **Poisson**(μ=np=50)
- Sospechamos que diferentes hospitales pueden tener servicios de traumatología de diferente "calidad" (algunos presentan pocos, pero creemos que aún demasiados, enfermos con secuelas tras la intervención). Es dificil compararlos pues cada hospital atiende poblaciones de tamaños diferentes (ciudades, pueblos,…)
  - Tenemos en cada hospital n, nº de pacientes atendidos o nº individuos de la población que cubre el hospital.
  - Tenemos p pequeño calculado como frecuencia relativa de secuelas con respecto al total de pacientes que trata el hospital, o el tamaño de la población,…
  - Se puede modelar mediante Poisson(μ=np)

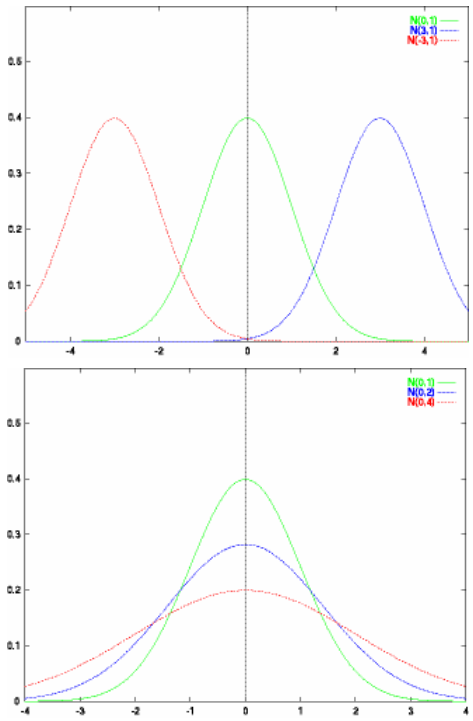
Distribución normal o de Gauss

- 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:

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

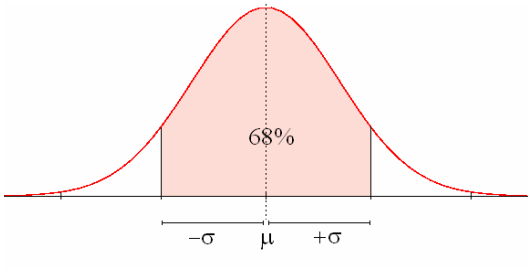
N(μ, σ): Interpretación geométrica

- Podéis interpretar la media como un factor de traslación.
- Y la desviación típica como un factor de escala, grado de dispersión,…

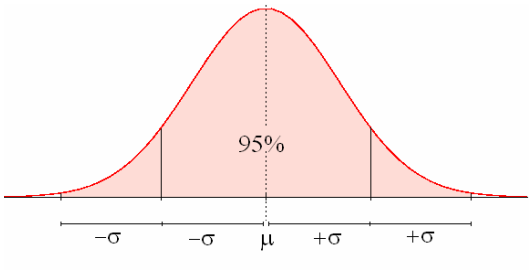


N(μ, σ): Interpretación probabilista

- 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.

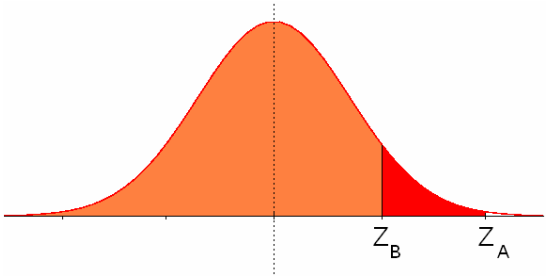
- Nos permite así comparar entre dos valores de dos distribuciones normales diferentes, para saber cuál de los dos es más extremo.

Ejemplo

- 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)

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

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



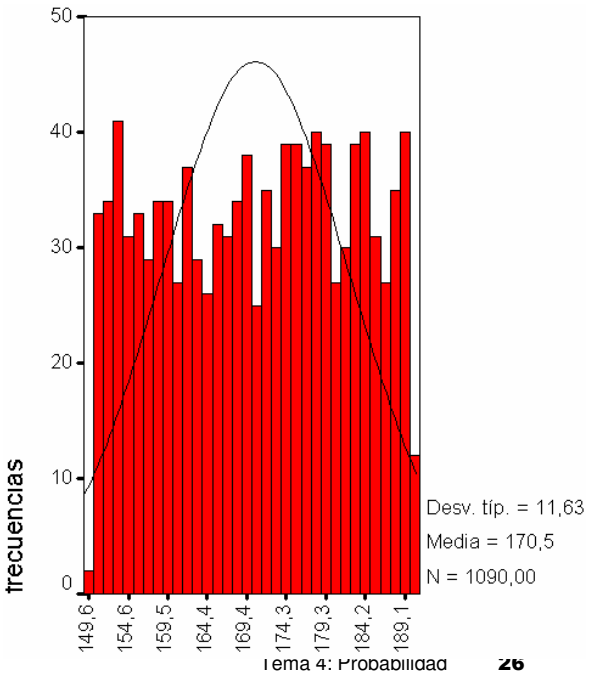
  - Como $Z_A > Z_B$, podemos decir que el porcentaje de compañeros del mismo sistema de estudios que ha superado en calificación el estudiante A es mayor que el que ha superado B. Podríamos pensar en principio que A es mejor candidato para la beca.

¿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).

Veamos aparecer la distribución normal

- Como ilustración mostramos una variable que presenta valores distribuidos más o menos uniformemente sobre el intervalo 150-190.
- Como es de esperar la media es cercana a 170. El histograma no se parece en nada a una distribución normal con la misma media y desviación típica.

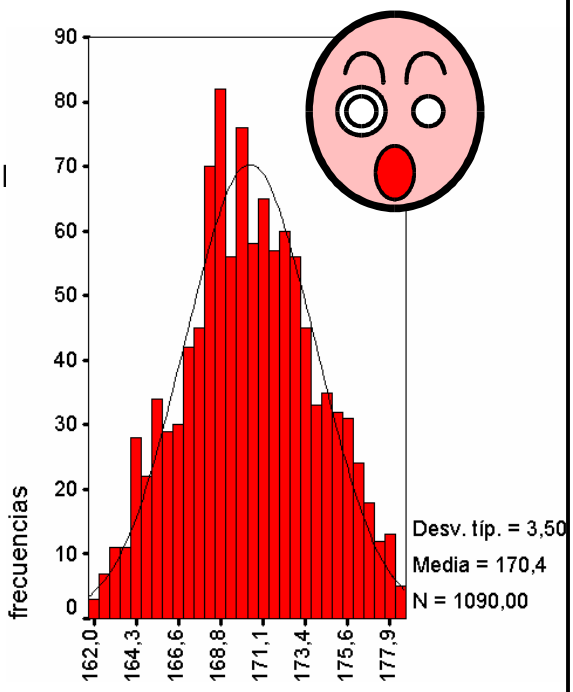


- A continuación elegimos aleatoriamente grupos de 10 observaciones de las anteriores y calculamos el promedio.
- Para cada grupo de 10 obtenemos entonces una nueva medición, que vamos a llamar promedio muestral.
- Observa que las nuevas cantidades están más o menos cerca de la media de la variable original.
- Repitamos el proceso un número elevado de veces. En la siguiente transparencia estudiamos la distribución de la nueva variable.

Muestra		
1ª	2ª	3ª
185	190	179
174	169	163
167	170	167
160	159	152
172	179	178
183	175	183
188	159	155
178	152	165
152	185	185
175	152	152

173 169 168 ⋯

- La distribución de los promedios muestrales sí que tiene distribución aproximadamente normal.
- La media de esta nueva variable (promedio muestral) es muy parecida a l de la variable original.
- Las observaciones de la nueva variable están menos dispersas. Observa el rango. Pero no sólo eso. La desviación típica es aproximadamente 'raiz de 10' veces más pequeña. Llamamos error estándar a la desviación típica de esta nueva variable.
- **Nada** de lo anterior **es casualidad**.



Teorema central del límite

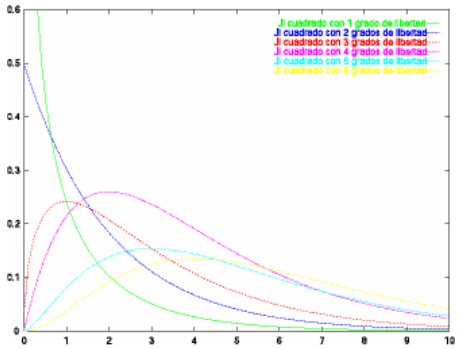
- Dada una v.a. **cualquiera**, si extraemos muestras de tamaño n, y calculamos los promedios muestrales, entonces:
- dichos promedios tienen distribución aproximadamente normal;
- La media de los promedios muestrales es la misma que la de la variable original.
- La desviación típica de los promedios 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.

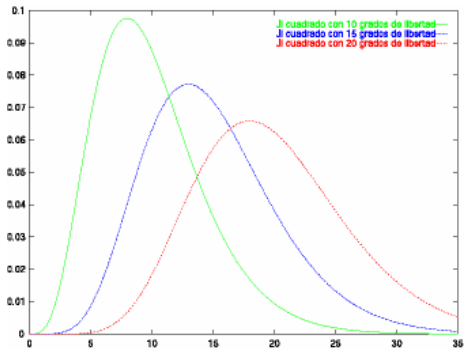
Distribuciones asociadas a la normal

- 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

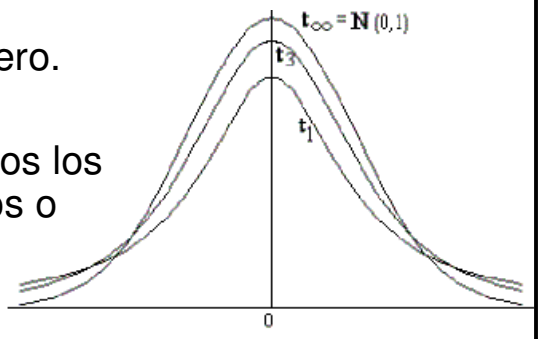
- Tiene un sólo parámetro denominado grados de libertad.
- La función de densidad es asimétrica positiva. Sólo tienen densidad los valores positivos.
- La función de densidad se hace más simétrica incluso casi gausiana cuando aumenta el número de grados de libertad.
- Normalmente consideraremos anómalos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





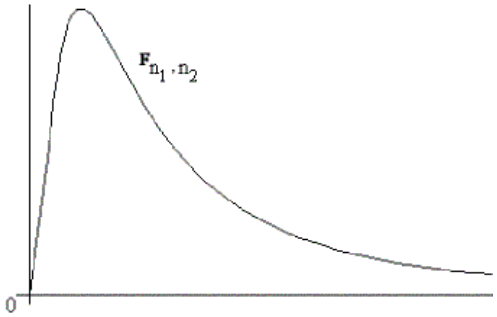
T de student

- 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, …
- Hay modelos de v.a. de especial importancia:
  - Bernoulli
  - Binomial
  - Poisson
  - Normal
    - Propiedades geométricas
    - Tipificación
    - Aparece tanto en problemas con variables cualitativas (dicotómicas, Bernoulli) como numéricas
    - Distribuciones asociadas
      - T-student
      - X2
      - F de Snedecor