Estadística

Variable aleatoria uniforme

Decimos que una X variable tiene distribución uniforme en un intervalo (a,b),

X ~ U(a,b)

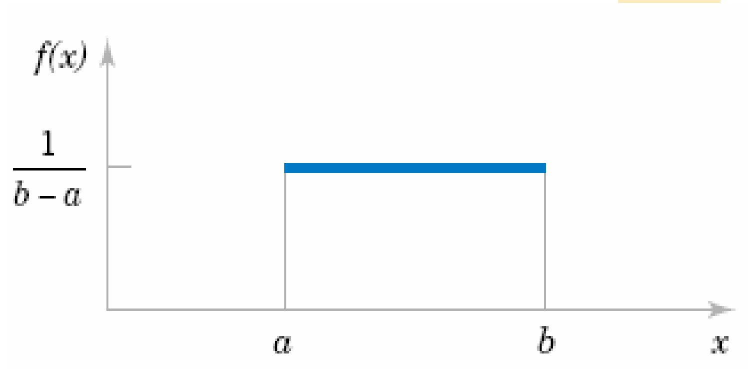
si X tiene función de densidad dada por

$$f(x) = \begin{cases} \dfrac{1}{b-a} & si\ a \leq x \leq b \\ 0 & en\ caso\ contrario \end{cases}$$

Las medidas características serán:

$$E[X] = \frac{a+b}{2} \qquad Var[X] = \frac{(b-a)^2}{12}$$

Variable aleatoria uniforme



Función de densidad de una uniforme U(a,b)

Variable aleatoria uniforme

Ejemplo:

La función de densidad del tiempo requerido para completar un ensamblaje es

$$f(x) = \begin{cases} 0.1 & si \;\; 30 < x < 40 \; segundos \\ 0 & en \; caso \; contrario \end{cases}$$

a)Calcular la proporción de ensamblajes que requerirán más de 35 segundos para ser completados.

b)¿Cuál es el tiempo excedido por el 90 % de los ensamblajes?

c)Determinar el tiempo medio necesario para completar un ensamblaje.

Variable aleatoria uniforme

Solución:

a)Calcular la proporción de ensamblajes que requerirán más de 35 segundos para ser completados.

$$P(X > 35) = \int_{35}^{40} 0.1dx = 4 - 3{,}5 = 0{,}5$$

b)¿Cuál es el tiempo excedido por el 90 % de los ensamblajes?

$$P(X > x) = 0{,}9 \Rightarrow \int_{x}^{40} 0.1dx = 4 - x \cdot 0{,}1 = 0{,}9 \Rightarrow x = 31\ segundos$$

c)Determinar el tiempo medio necesario para completar un ensamblaje.

E[X] = (40+30)/2 = 35 segundos

Variable aleatoria exponencial

La distribución exponencial está relacionada con el proceso de Poisson. Cuando el número de sucesos en un intervalo de medida sigue una distribución de Poisson, el tiempo entre sucesos sigue una distribución exponencial.

Se puede utilizar para modelizar

- Tiempo entre llamadas telefónicas
- Tiempo entre llegadas a un puesto de servicio
- Tiempo de vida de un componente eléctrico

Variable aleatoria exponencial

Sean

X = Número de sucesos en la unidad de tiempo

T = Tiempo hasta que ocurre el primer suceso

¿Cuál es la función de densidad de T?

Podemos calcular su función de distribución:

$$P(T > t_0) = P(\text{cero sucesos en } (0,t_0))$$

Supongamos que X ~ P(λ), entonces

Y = Número de sucesos en $(0,t_0)$ verifica $Y \sim P(\lambda t_0)$

$$P(T > t_0) = P(Y = 0) = e^{-\lambda t_0}$$

De donde tenemos que

$$F(t)=P(T \leq t)=1- e^{-\lambda t}$$

Variable aleatoria exponencial

¿Cuál es la función de densidad de T?

$$F(t) = P(T \leq t) = 1- e^{-\lambda t} \longrightarrow \quad T \sim Exp(\lambda)$$

implica que la función de densidad es

$$f(x) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t}, \;\; para \;\; t \geq 0$$

Las medidas características son:

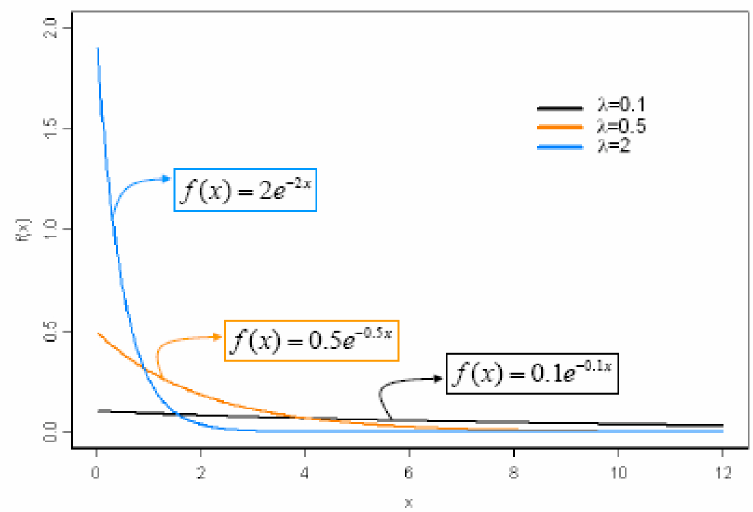
$$E[T] = 1/\lambda$$

$$Var[T] = 1/\lambda^2$$

Observación:

Si hay λ sucesos por término medio, en un intervalo de tiempo, entonces el tiempo medio entre sucesos es $1/\lambda$

Variable aleatoria exponencial



Variable aleatoria exponencial

Ejemplo:

El proceso de llegadas de clientes a un puesto de servicio se produce de manera estable e independiente. Por término medio llega un cliente cada minuto.

¿Cuál es la probabilidad de que pasen más de 3 minutos entre la llegada de dos clientes?

X = Número de clientes por minuto $\longrightarrow$ X ~ P(λ=1)

Y = Tiempo entre dos clientes $\longrightarrow$ Y ~ Exp(λ=1)

P(no hay clientes en 3 minutos) = $P(Y > 3) = 1 - P(Y \leq 3)$

$= 1 - F(3) = 1 - (1 - e^{-3}) = e^{-3}$

Variable aleatoria exponencial

Propiedad:

La distribución exponencial no tiene memoria

$$\underline{P(T > t_1 + t_2 | T > t_1)} = \underline{P(T > t_2)}$$

$$\frac{P(T > t_1 + t_2, T > t_1)}{P(T > t_1)} = \frac{P(T > t_1 + t_2)}{P(T > t_1)} = \frac{e^{-\lambda(t_1+t_2)}}{e^{-\lambda t_1}} = e^{-\lambda t_2}$$

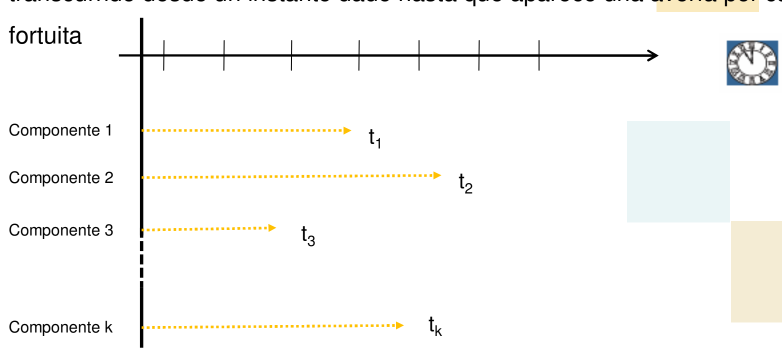
Si no ha habido clientes en 4 minutos, ¿cuál es la probabilidad de que no haya clientes en los próximos 3 minutos?

$P(T > 7 \mid T > 4) = P(T > 3) = 1 - P(T \leq 3) = 1 - F(3) = 1 - (1 - e^{-3}) = e^{-3}$

Universidad Europea de Madrid

Variable aleatoria exponencial

La distribución exponencial también es útil para modelizar el tiempo transcurrido desde un instante dado hasta que aparece una avería por causa fortuita



T = tiempo hasta que un equipo, componente, comunicación, etc, falla por causa fortuita $\longrightarrow$ $T \sim Exp(\lambda)$

¿Por qué fortuita? Porque NO HAY MEMORIA

No hay envejecimiento ni mejora, $E(T) = 1/\lambda$ es constante

Universidad Europea de Madrid

Variable aleatoria normal

La distribución Normal describe gran cantidad de procesos aleatorios

• Errores de medida

• Ruido en una señal digital

• Corriente eléctrica en un trozo de cable

…

En muchas situaciones otras distribuciones se pueden aproximar a una Normal

Es la base para la inferencia estadística

Universidad Europea de Madrid

Variable aleatoria normal

Está caracterizada por dos parámetros:

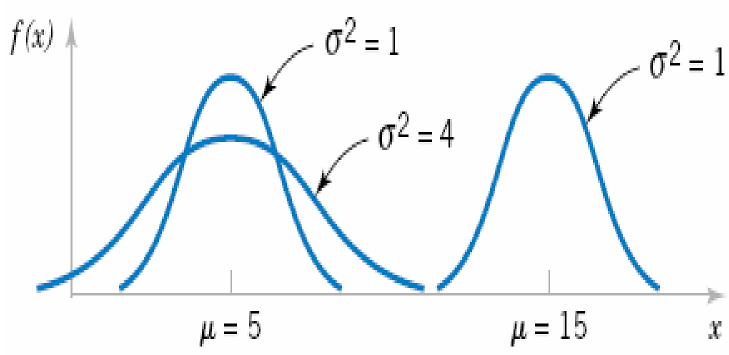
- La **media, μ, y la desviación típica, σ.**
- Toma valores en toda la recta real
- Su función de densidad y medidas características están dadas por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

$$E[X] = \mu \qquad Var[X] = \sigma^2$$

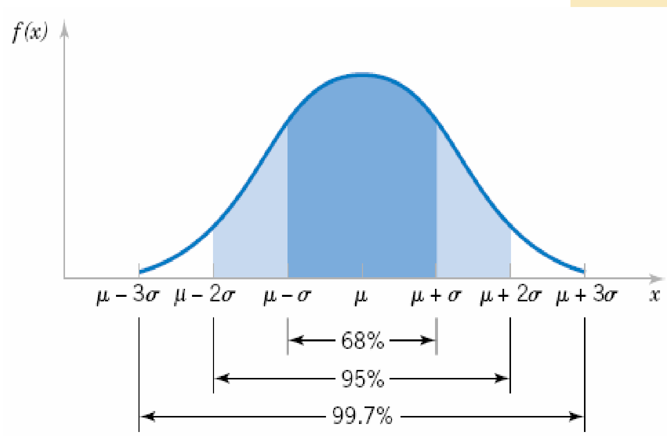
Universidad Europea de Madrid

Variable aleatoria normal



- Tiene forma de campana y es simétrica respecto de la media
- La media, mediana y moda coinciden.
- σ es un factor de escala y μ es un factor de traslación.

Variable aleatoria normal



Concentrada alrededor de la media

Variable aleatoria normal

Otras propiedades importantes:

• Si hacemos una transformación lineal de una normal, el resultado es un es variable aleatoria normal

• La combinación lineal de variables aleatorias normales, da como resultado una variable aleatoria normal

Variable aleatoria normal

Ejemplo:

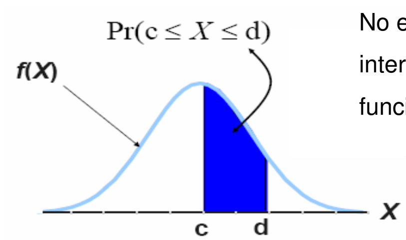
Sean $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$, dos variables aleatorias normales independientes la una de la otra. Entonces $Y = aX_1 + bX_2$ es también una variable normal. Además se tendrá que

$$E(Y) = E(aX_1 + bX_2) = a\mu_1 + b\mu_2$$

Por ser ambas variables aleatorias independientes se tiene también que

$$Var(Y) = Var(aX_1 + bX_2) = a^2\sigma_1^2 + b^2\sigma_2^2$$

Variable aleatoria normal



No es posible calcular la probabilidad de un intervalo simplemente usando la integral de la función de densidad: es difícil de integrar

¿Solución?

- Integramos usando ordenador
- Integramos usando Tablas

Sólo hay tablas de la N(0,1), que también se llama Normal estándar o Z

Transformamos un problema de $X \sim N(\mu,\sigma^2)$ a $Z \sim N(0,1)$.

Esta transformación se llama estandarización o tipificación

Variable aleatoria normal

Estandarización de una variable $X \sim N(\mu,\sigma^2)$

$$X \sim N(\mu,\sigma^2) \longrightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$E[Z] = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma} E[X-\mu] = 0$$

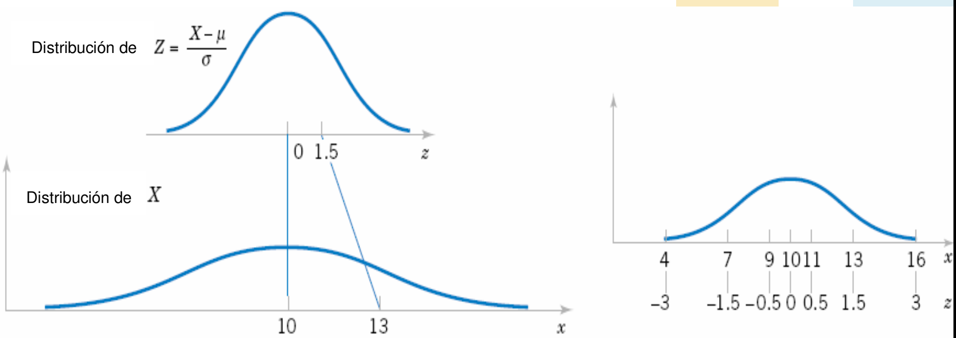
$$Var[Z] = Var\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2} Var[X-\mu] = 1$$

Notación:

$$P(Z \leq z) = \Phi(\mathrm{z})$$

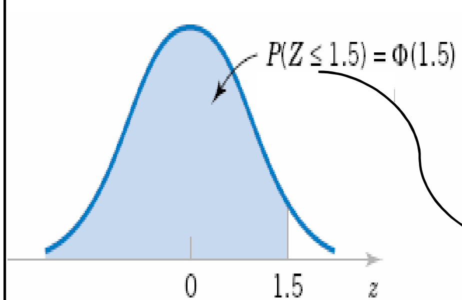
(estos son los valores que están tabulados)

Variable aleatoria normal



Estandarización de una variable normal

Variable aleatoria normal

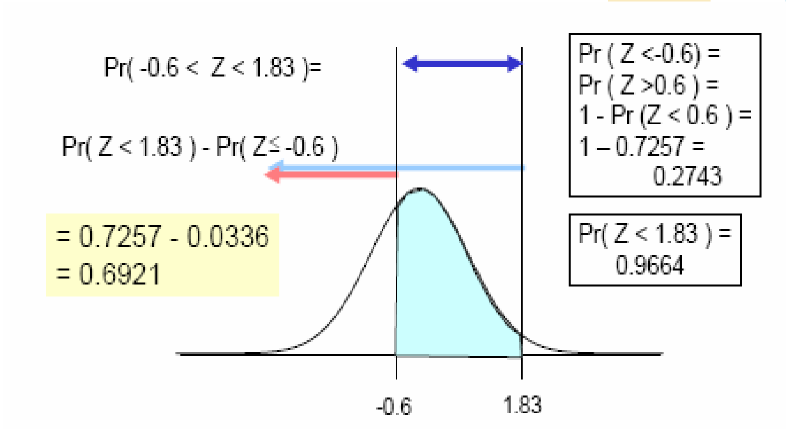


z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Distribución de la normal estándar mediante tablas

Variable aleatoria normal

Cálculo de probabilidades usando la normal:

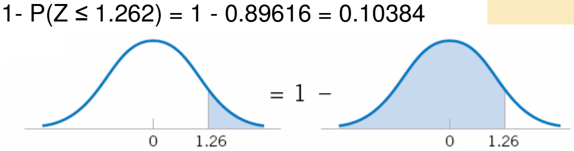


Universidad Europea de Madrid

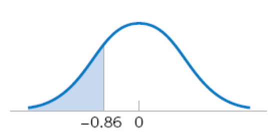
Variable aleatoria normal

Hallar las siguientes probabilidades:

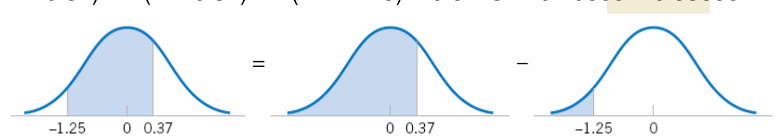
a) $P(Z > 1.26) = 1 - P(Z \leq 1.262) = 1 - 0.89616 = 0.10384$



b) $P(Z < -0.86) = 0.19490$



c) $P(-1.25 < Z < 0.37) = P(Z < 0.37) – P(Z < -1.25) = 0.64431 - 0.10565 = 0.53866$



Universidad Europea de Madrid

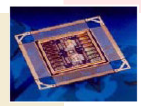
Variable aleatoria normal

Ejemplo 1:

El tiempo de vida de un semiconductor sigue una distribución normal con media 7000 horas y desviación típica 600 horas.

1.¿Cuál es la probabilidad de que el semiconductor falle antes de 6000 horas?

2.¿Qué tiempo de vida en horas es excedido por el 95,05% de los semiconductores?



Variable aleatoria normal

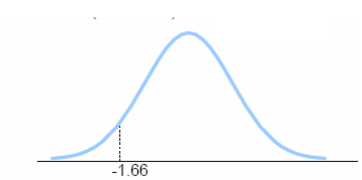
Solución:

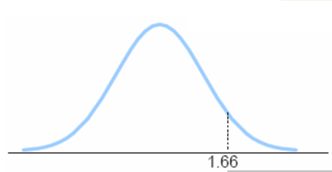
1.¿Cuál es la probabilidad de que el semiconductor falle antes de 6000 horas?

Sea X = tiempo de vida del semiconductor $\longrightarrow X \sim N(7000, 600^2)$

$$P(X < 6000) = P\left(Z < \frac{6000 - 7000}{600}\right)$$

$$= \quad P(Z < -1{,}66) = 1 - P(Z < 1{,}66) = 1 - 0{,}9515 = 0{,}0485$$





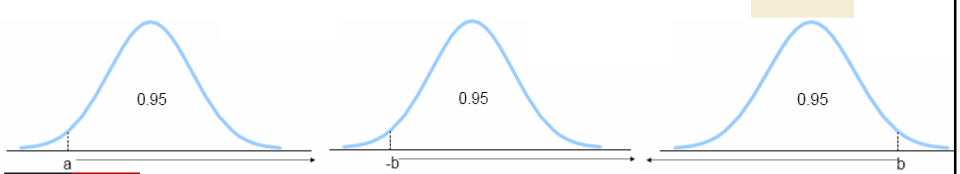
Variable aleatoria normal

Solución:

2.¿Qué tiempo de vida en horas es excedido por el 95,05% de los semiconductores?

$$P(X > a) = 0{,}9505 \longrightarrow P\left(Z > \underbrace{\frac{a-7000}{600}}_{-b}\right) = 0{,}9505$$

$$P\left(Z < -\frac{a-7000}{600}\right) = 0{,}9505$$



Variable aleatoria normal

Solución:

$$P\left(Z < -\frac{a-7000}{600}\right) = 0{,}9505 \Rightarrow -\frac{a-7000}{600} = 1{,}65$$

$$\Rightarrow a = 6010$$

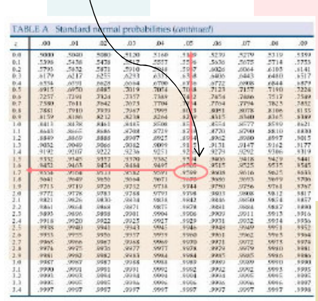


TABLE A Standard normal probabilities (continued)

El 95.05% de los semiconductores duran más de 6010 horas

Variable aleatoria normal

Ejemplo 2:

La longitud L en milímetros, de las piezas fabricadas en un proceso es una variable aleatoria que se distribuye según una N(32,0.3²), considerándose aceptables aquellas cuya medida se encuentra dentro del intervalo (31.1,32.6). Calcular la probabilidad de que una pieza elegida al azar sea aceptable.

$$P(31.1 < X < 32.6) = P\left(\frac{31.1-32}{0.3} < Z < \frac{32.6-32}{0.3}\right) = 0.976$$

Universidad Europea de Madrid

Variable aleatoria normal

Ejemplo 3:

Varios tests de inteligencia dieron una puntuación que sigue una ley normal con media 100 y desviación típica 15.

(a)Determinar el porcentaje de población que obtendría un coeficiente entre 95 y 100.

(b)¿Qué intervalo centrado en 100 contiene al 50% de la población?

Universidad Europea de Madrid

Variable aleatoria normal

Solución:

(a)Determinar el porcentaje de población que obtendría un coeficiente entre 95 y 100.

$$P(95 \leq X \leq 100) = P\left(\frac{95-100}{15} \leq Z \leq \frac{100-100}{15}\right)$$

$$= P\left(\frac{1}{3} \leq Z \leq 0\right) = 0.13$$

(b) ¿Qué intervalo centrado en 100 contiene al 50% de la población?

$$P(0 \leq Z \leq z) = 0.25 \Rightarrow z = 0.675 \Rightarrow \frac{X-100}{15} = 0.675$$

$$\Leftrightarrow X \simeq 110 \longrightarrow [90,110]$$

Teorema Central del límite

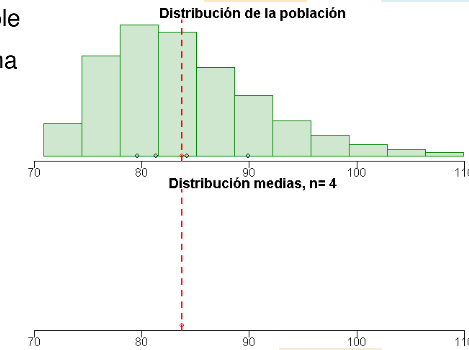
La Normal es importante, no sólo porque muchas variables comunes sigan esa distribución, sino porque aunque una v.a. no posea distribución normal, ciertos estadísticos/estimadores calculados sobre muestras elegidas al azar sí poseen una distribución Normal.

Es decir, tengan las distribución que tengan nuestros datos, los ‘objetos’ que resumen la información de una muestra, posiblemente tengan distribución normal.

Teorema Central del límite

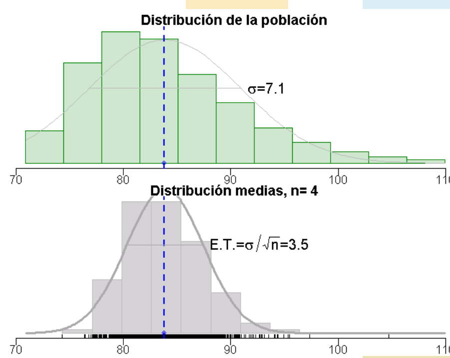
• Como ilustración mostramos una variable que presenta valores distribuidos de forma muy asimétrica. Claramente no normal.

• Saquemos muestras de diferentes tamaños, y usemos la media de cada muestra para estimar la media de la población.



Teorema Central del límite

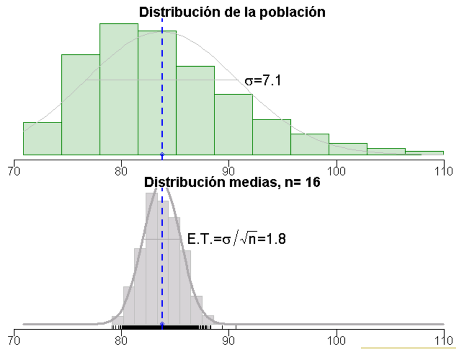
- Cada muestra ofrece un resultado diferente: La media muestral es una variable aleatoria.
- Su distribución es más parecida a la normal que la original.
- También está menos dispersa. A su dispersión ('desv. típica del estimador media muestral') se le suele denominar **<u>error típico</u>**.



Teorema Central del límite

Al aumentar el tamaño, n, de la muestra:

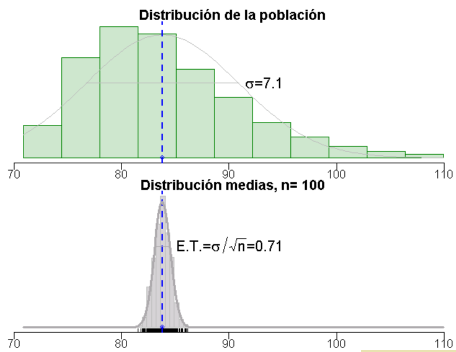
- ❑ La normalidad de las estimaciones mejora
- ❑ El error típico disminuye.



Universidad Europea de Madrid

Teorema Central del límite

- 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.



Universidad Europea de Madrid

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.

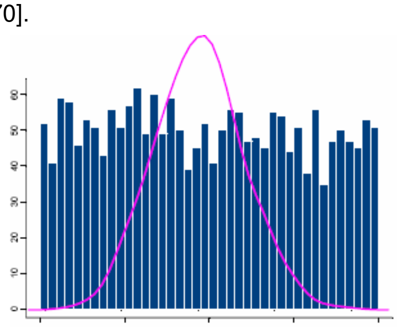
Teorema Central del límite

Sea X una variable Uniforme en el intervalo [50,70].

Tenemos una muestra de tamaño 2000.

La muestra tiene media 59.9 y desviación típica 4.57

El histograma no se parece a una distribución normal con la misma media y desviación típica



Teorema Central del límite

Elegimos aleatoriamente grupos de 10 observaciones.

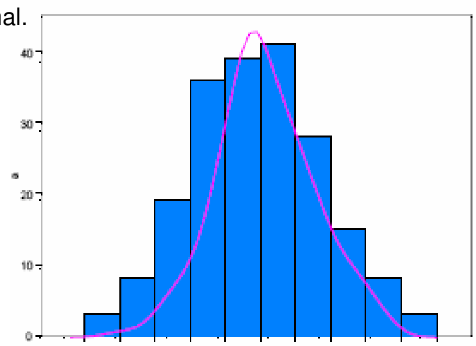
Para cada grupo de 10 obtenemos entonces una nueva medida: la media muestral.

Las medias de cada muestra están más o menos cerca de la media de la variable original.

Muestra		
1ª	2ª	3ª
59	63	59
60	60	69
66	58	60
54	53	65
51	51	69
54	59	54
51	53	59
63	62	65
66	57	56
70	69	55
59.4	58.5	61.1

Teorema Central del límite

La distribución de las medias muestrales tiene distribución aproximadamente normal.



La media de esta nueva variable es muy parecida a la de la variable original.

Las observaciones de la nueva variable están menos dispersas. La desviación típica es menor, en este caso 1.92

Teorema Central del límite

Supongamos que tenemos n variables aleatorias X_i independientes con medias (μ_i) y desviaciones típicas (σ_i) y distribución cualquiera

Teorema Central del Límite

Cuando n crece, $Y = X_1 + X_2 + \cdots + X_n$

$$Y \approx N\left(\sum_i \mu_i , \sqrt{\sum_i \sigma_i{}^2}\right)$$

o, en forma equivalente

$$\frac{Y - \sum_i \mu_i}{\sqrt{\sum_i \sigma_i{}^2}} \approx N(0,1)$$

Universidad Europea de Madrid

Ejemplo

El tiempo (en minutos) que una impresora en red tarda en imprimir un trabajo sigue una distribución exponencial de parámetro 4/3. Si a las 14.15 se encuentran 35 trabajos en cola para imprimir, ¿qué probabilidad hay de que a las 14.30 no quede nada por imprimir?.

Sea Y: "tiempo en imprimir 35 trabajos". $Y = \sum_{i=1}^{35} T_i$

Por el TCL sabemos que . $Y \approx N\left(\sum_i \mu_i , \sqrt{\sum_i \sigma_i{}^2}\right)$

$$E[Y] = \sum E[T] = \sum 3/4 = 35 \cdot \frac{3}{4} = 26.25 \qquad V[Y] = \sum V(t) = 35 \cdot \left(\frac{3}{4}\right)^2 = 19.69$$

Universidad Europea de Madrid

Ejemplo

- Ahora calculamos la probabilidad pedida:

$$P(X < 15) = P\left(Z < \frac{15 - 26.25}{\sqrt{19.69}}\right) = P(z < -2.54) = 0.0055$$

Universidad Europea de Madrid

Aproximación normal a la binomial

La variable binomial es suma de variables de Bernouilli, que toman el valor 0 ó 1.

$Y = X_1 + X_2 + \ldots X_n$ $\quad$ $E[X_i] = p$

$Var[X_i] = p(1 - p)$

$$\Big\downarrow \text{TCL}$$

$Y \approx N\,(np, np(1-p))$

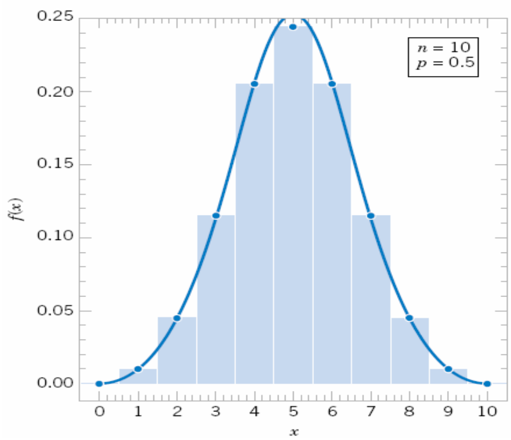
¿Cuándo?

$n > 30$

$np(1 - p) > 5$

Universidad Europea de Madrid

Variable aleatoria normal



Aproximación normal a la binomial

Aproximación normal a la binomial

La distribución Normal es continua pero la binomial es discreta.

Para mejorar la aproximación introducimos un factor de corrección que consiste en añadir o substraer 0.5 al valor al que le queremos calcular la probabilidad.

$$P(X \leq x) = P(X \leq x + 0.5) \cong P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(x \leq X) = P(x - 0.5 \leq X) \cong P\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}} \leq Z\right)$$

Aproximación normal a la binomial

Ejemplo:

Un fabricante de semiconductores admite que produce un 2% de chips defectuosos. Los chips se empaquetan en lotes de 2000 chips para su venta. Un comprador rechazará un lote si contiene 25 o más chips defectuosos ¿Cuál es la probabilidad de rechazar un lote?

Sea X = número de chips defectuosos $\longrightarrow$ X ~ B(2000, 0.02)

n > 30 np = 40 np(1 –p) = 39.2

$\Rightarrow$ X ≈ N(40,6.26)

$$P(X \geq 25) \approx P\left(Z \geq \frac{25 - 0.5 - 40}{6.26}\right) = P(Z \geq -2.47)$$

$$= P(Z < -2.47) = 0.9292$$

Universidad Europea de Madrid

Aproximación normal a la Poisson

La distribución de Poisson surge como límite de la binomial cuando el número de experimentos tiende a infinito.

Aproximamos a una Normal cuando λ grande (λ > 5)

$$X \sim P(\lambda)$$

$$X \approx N\left(\lambda, \sqrt{\lambda}\right)$$

Universidad Europea de Madrid

Aproximación normal a la Poisson

Ejemplo:

El número de defectos en la superficie de un material por metro cuadrado sigue una distribución de Poisson con media 100.

Si se analiza un metro cuadrado de dicho material, ¿Cuál es la probabilidad de encontrar 95 defectos o más?

Sea X = número de defectos en un metro cuadrado

$$P(X \geq 95) \approx P\left(Z \geq \frac{95 - 100 - 0.5}{10}\right)$$

$$= P(Z \leq 0.55) = 0.6915$$

Universidad Europea de Madrid

Ejemplo

Recuperemos el ejemplo de la impresora en red. Otra manera de resolver sería:

Como nos daban que el tiempo que tardaba en imprimir era una exponencial de parámetro 4/3, tenemos que el numero de documento impreso en un minuto es una Poisson de parámetro 4/3.

Como queremos calcular la probabilidad de imprimir 35 documentos en 15 minutos será: X->P(20). Ya que 15 ·4/3=20.

Aplicamos ahora la aproximación de la Poisson por la distribución normal y obtenemos:

$$X \approx N(20, \sqrt{20}) \qquad P(X > 35) = P\left(X > \frac{35 - 20}{\sqrt{20}}\right) = 0.0004$$

Universidad Europea de Madrid