Universidad de Alcalá

Dpto. Ciencias de la Computación

Estadística

Tema 4. 2ª Parte: Modelos de distribución de probabilidad discretas y continuas

Distribuciones de Probabilidad

- Una vez expuesta la teoría general sobre variables aleatorias y sus distribuciones de probabilidad, vamos a describir algunas distribuciones particulares que han demostrado, empíricamente, ser modelos apropiados para situaciones que ocurren en la vida real.

- Estas distribuciones presentan un **carácter teórico** en el sentido de que sus funciones de probabilidad o de densidad se deducen matemáticamente en base a ciertas hipótesis que se suponen validas para los fenómenos aleatorios.

- La elección de una distribución de probabilidad para representar un fenómeno de interés practico debe estar motivada tanto por la comprensión de la naturaleza del fenómeno en si, como por la posible verificación de la distribución seleccionada a través de la evidencia empírica.

- Una distribución de probabilidad esta caracterizada, de forma general, por una o mas cantidades que reciben el nombre de **parámetros de la distribución**.

- Un parámetro puede tomar cualquier valor de un conjunto dado y, en ese sentido, se define una familia de distribuciones de probabilidad que tendrán la misma función genérica de probabilidad o función de densidad.

Modelos de Distribución de Probabilidad

En este tema estudiaremos algunas de las principales distribuciones de probabilidad, Discretas y Continuas

Discretas

- Distribución Uniforme Discreta
- **Distribución binomial**
- **Distribución de Poisson**
- Distribución Geométrica
- Distribución binomial Negativa
- Distribución Hipergeométrica

Continuas

- **Distribución Uniforme (continua)**
- **Distribución Exponencial**
- **Distribución Normal**

- **Distribución Chi-cuadrado**
- **Distribución t de Student**
- **Distribución F de Snedecor**

DISTRIBUCIONES DISCRETAS

Distribución de Bernoulli de parámetro p

- Supongamos un experimento aleatorio que sólo admite dos resultados excluyentes (éxito y fracaso).
- Por ejemplo sea la variable aleatoria discreta X asociada a este experimento, X toma el valor 1 cuando ocurre el suceso éxito con probabilidad p y el valor 0 cuando ocurre el suceso fracaso con probabilidad q = (1-p)

x	$P(X = x_i)$
0	q
1	p

$\mu_X = p$ $\sigma_X^2 = p \, . \, q$ $\sigma_X = \sqrt{p \, . \, q}$

Ejemplo:

El 10% de los trabajadores del país está desempleado,
¿Cuál es la probabilidad de seleccionar un individuo al azar y esté desempleado?

$X = 1 \Rightarrow$ Desempleado p = 0,1 p(x=1) = 0,1
$X = 0 \Rightarrow$ Empleado q = 1-p = 1-0,1 = 0,9 p(x=0) = 0,9

Distribución binomial introducción

- Supongamos un resultado elemental ***k***, de **éxitos** en ***n*** pruebas.
- Consideramos el **suceso *k:* elementos con éxito (*E*) en n pruebas seguidos de n-*k* de no éxito (*NE*)**

$$\underbrace{\mathrm{E, E, \ldots E}}_{k} \quad \underbrace{\mathrm{NE, NE, \ldots\ NE}}_{n-k}$$

- Por la hipótesis de la independencia, la probabilidad de este suceso será:

$$\underbrace{\mathrm{p\,p\ \cdots p}}_{K} \quad \underbrace{(1-\mathrm{p})(1-\mathrm{p})\cdots(1-\mathrm{p})}_{n-K} = \mathrm{p^k}(1-\mathrm{p})^{\mathrm{n-k}}$$

- La probabilidad de k resultados cualquier orden requiere contar las probabilidades de todos los sucesos excluyentes que verifican la condición E y NE en todas las ordenaciones posibles.

Distribución binomial

- Es una extensión de la distribución de Bernoulli. Supongamos que se repite un experimento Bernoulli "n" veces de forma idéntica e independiente. Los resultados de cada realización del experimento se clasifican en dos categorías, una será la probabilidad de éxito p, y otra q=1-p, la de fracaso.
- Sea X una variable aleatoria discreta, se dice que se distribuye como una **distribución binomial** de parámetros (n, p), con n>1 y 0≤p≤1, si su función de probabilidad viene dada por la expresión:

$$\mathrm{P}\left[X = x_i\right] = \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \quad x = 1,2,\ldots,n.$$

Distribución binomial

La función de distribución

$$F(x)= \begin{cases} 0 & x \le 0 \\ \sum_{I=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} & 0 \le x \le n \\ 1 & x > n \end{cases}$$

Se verifica que:

$E(x) = n\ p$;$Var\ (x) = n\ p\ (1-p) = n\ p\ q$

Ejemplos de variables que se distribuyen con una Binomial:
Número de caras al lanzar 20 veces una moneda, número de aprobados si se presentan 80 alumnos a un examen, número de familias con un solo hijo en una población de 120 familias, número de reacciones negativas ante un fármaco administrado a 40 pacientes, número de accidentes de tráfico si han circulado 1200 automóviles o número de semillas que germinan de las 20 semillas que se han plantado en suelos de idéntica composición.

Distribución binomial

Propiedades :

1. La distribución Binomial se puede obtener como suma de **n** variables aleatorias independientes Bernoulli con el mismo parámetro "p".
2. Si tenemos dos variables aleatorias que se distribuyen según una Binomial con el mismo parámetro "p", es decir, con la misma probabilidad de éxito,

   $X \rightarrow B(n, p)$ e, $Y \rightarrow B(m, p)$ entonces siempre se verifica: $X+Y \rightarrow B(n+m, p)$

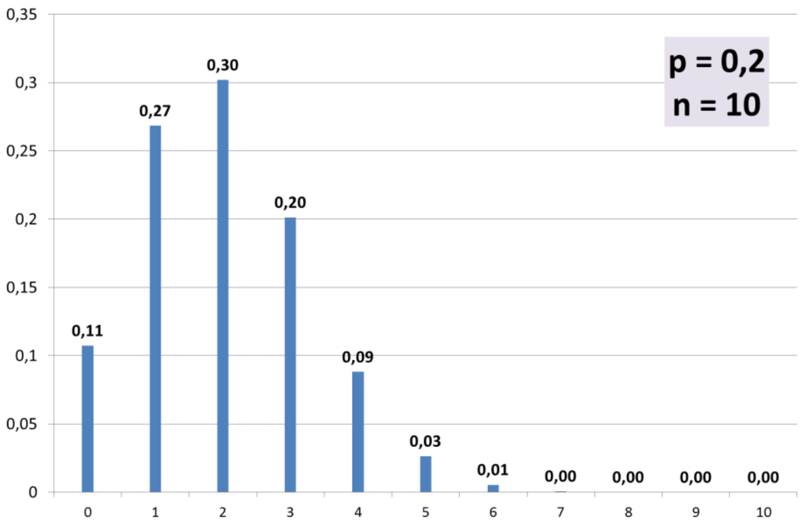
   OJO: si no tienen la misma probabilidad no se pueden sumar.

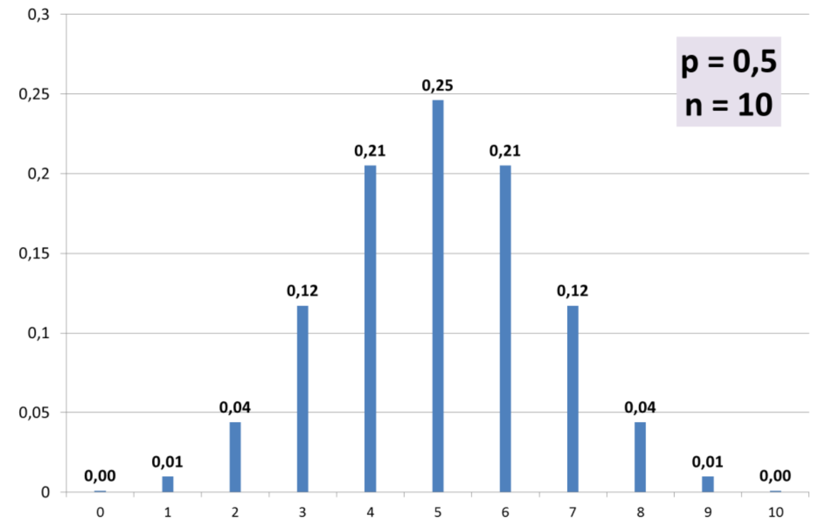
3. Sea X una variable aleatoria e Y otra variable aleatoria que verifican que

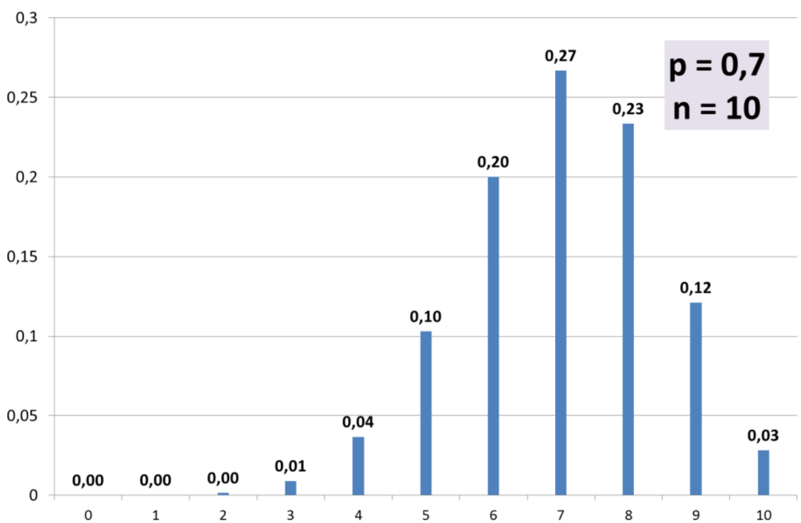
   $X \rightarrow B(n, p)$, $Y=X/n$, entonces se verifica $Y \rightarrow B(1, p/n)$

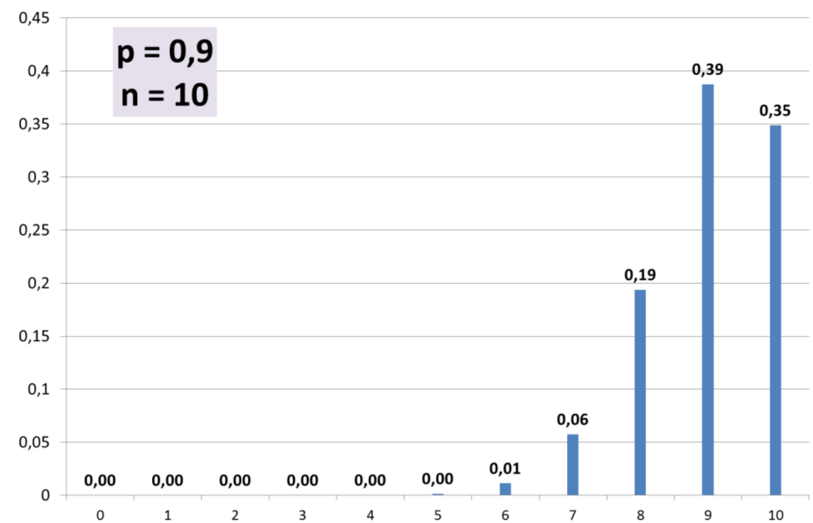
   La esperanza y la varianza son: $E[Y] = p$ y $V[Y] = \dfrac{pq}{n}$.

Distribución binomial









Distribución binomial

Ejercicio 1

La probabilidad de éxito de una determinada vacuna es 0,72. Calcular la probabilidad de que una vez administrada a 15 pacientes:

a) Ninguno sufra la enfermedad
b) Todos sufran la enfermedad
c) Dos de ellos contraigan la enfermedad

Solución :

Se trata de una distribución binomial de parámetros B(15;0,72)

$$\text{a)}\quad P(X=15)=\binom{15}{15}\cdot 0{,}72^{15}\cdot 0{,}26^{0}=0{,}00724=0{,}724\%$$

$$\text{b)}\quad P(X=0)=\binom{15}{0}\cdot 0{,}72^{0}\cdot 0{,}28^{15}=5{,}097\cdot 10^{-9}=5{,}097\cdot 10^{-7}\%$$

$$\text{c)}\quad P(X=13)=\binom{15}{13}\cdot 0{,}72^{13}\cdot 0{,}28^{2}=0{,}11503=11{,}503\%$$

Distribución binomial

- **Ejercicio 2**

  La probabilidad de que el carburador de un coche salga de fábrica defectuoso es del 4 por 100.

  Hallar :

  a) El nº de carburadores defectuosos esperados en un lote de 1000.
  b) La varianza y la desviación típica.

Solución :

$$a)\ E(X) = \mu = n \cdot p = 1000 \cdot 0{,}04 = 40\ carburantes\ defectuosos$$

$$b)\ Var[X] = n \cdot p \cdot q = n \cdot p \cdot (1-p) = 1000 \cdot 0{,}04 \cdot 0{,}96 = 38{,}4$$

$$DT[X] = \sqrt{Var[X]} = 6{,}19$$

Distribución binomial

- **Ejercicio 3:**

Un sistema con **9 componentes** requiere para su funcionamiento **que al menos 6 de los mismos estén disponibles** si la probabilidad de funcionamiento de un componente es de 0,95, calcular la fiabilidad del sistema (probabilidad de que funcione).

n =9 y p = 0,95 entonces (1-p) = 0,005

$$P(funcionen \ \geq 6 \ componentes) = 1 - (P(funcionen \leq 5) = 1 - F(5) = F(9) - F(5) = \sum_{k=6}^{9} \binom{9}{k} 0{,}95^{k} 0{,}05^{9-k} = \binom{9}{6} 0{,}95^{6} 0{,}05^{3} + \binom{9}{7} 0{,}95^{7} 0{,}05^{2} + \binom{9}{8} 0{,}95^{8} 0{,}05^{1} + \binom{9}{9} 0{,}95^{9} 0{,}05^{0} = 0{,}9993574 \approx 99{,}94\%$$

Distribución binomial

n	k	p	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,33	0,35	0,40	0,45	0,49	0,50
2	0		0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4449	0,4225	0,3600	0,3025	0,2601	0,2500
	1		0,0198	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4442	0,4550	0,4800	0,4950	0,4998	0,5000
	2		0,0001	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1109	0,1225	0,1600	0,2025	0,2401	0,2500
3	0		0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2967	0,2746	0,2160	0,1664	0,1327	0,1250
	1		0,0294	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4444	0,4436	0,4320	0,4084	0,3823	0,3750
	2		0,0003	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2219	0,2389	0,2880	0,3341	0,3674	0,3750
	3		0,0000	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0369	0,0429	0,0640	0,0911	0,1176	0,1250
4	0		0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1979	0,1785	0,1296	0,0915	0,0677	0,0625
	1		0,0388	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3953	0,3845	0,3456	0,2995	0,2600	0,2500
	2		0,0006	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,2960	0,3105	0,3456	0,3675	0,3747	0,3750
	3		0,0000	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,0985	0,1115	0,1536	0,2005	0,2400	0,2500
	4		0,0000	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0123	0,0150	0,0256	0,0410	0,0576	0,0625
5	0		0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1320	0,1160	0,0778	0,0503	0,0345	0,0313
	1		0,0480	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3295	0,3124	0,2592	0,2059	0,1657	0,1563
	2		0,0010	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3291	0,3364	0,3456	0,3369	0,3185	0,3125
	3		0,0000	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1643	0,1811	0,2304	0,2757	0,3060	0,3125
	4		0,0000	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0410	0,0488	0,0768	0,1128	0,1470	0,1563
	5		0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0041	0,0053	0,0102	0,0185	0,0282	0,0313
6	0		0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0881	0,0754	0,0467	0,0277	0,0176	0,0156
	1		0,0571	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2638	0,2437	0,1866	0,1359	0,1014	0,0938
	2		0,0014	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3292	0,3280	0,3110	0,2780	0,2436	0,2344
	3		0,0000	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2191	0,2355	0,2765	0,3032	0,3121	0,3125
	4		0,0000	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0821	0,0951	0,1382	0,1861	0,2249	0,2344
	5		0,0000	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0164	0,0205	0,0369	0,0609	0,0864	0,0938
	6		0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0014	0,0018	0,0041	0,0083	0,0138	0,0156

Distribución binomial

EJEMPLO

Supongamos que un alumno realiza un examen tipo test de 10 preguntas con cuatro opciones cada una de las que sólo una es correcta. Si responde de forma aleatoria a todas las preguntas. Calcula:

a) Probabilidad de contestar 5 preguntas bien.

b) Probabilidad de contestar bien al menos 3 preguntas.

El problema evidentemente se puede enmarcar en una binomial de parámetros n=10 y p=0,25

n	k	p	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,33	0,35	0,40	0,45	0,49	0,50
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0174	0,0135	0,0060	0,0025	0,0012	0,0010
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0870	0,0725	0,0403	0,0207	0,0114	0,0098
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1955	0,1757	0,1209	0,0763	0,0494	0,0439
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2603	0,2522	0,2150	0,1665	0,1267	0,1172
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2274	0,2377	0,2508	0,2384	0,2130	0,2051
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1362	0,1536	0,2007	0,2340	0,2456	0,2461
	6		0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0567	0,0689	0,1115	0,1596	0,1966	0,2051
	7		0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0162	0,0212	0,0425	0,0746	0,1080	0,1172
	8		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0389	0,0439
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0083	0,0098
	10		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0010

P(x=5) = 0,0584

Distribución binomial

b) Probabilidad de contestar bien al menos 3 preguntas.

n	k	p	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,33	0,35	0,40	0,45	0,49	0,50
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0174	0,0135	0,0060	0,0025	0,0012	0,0010
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0870	0,0725	0,0403	0,0207	0,0114	0,0098
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1955	0,1757	0,1209	0,0763	0,0494	0,0439
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2603	0,2522	0,2150	0,1665	0,1267	0,1172
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2274	0,2377	0,2508	0,2384	0,2130	0,2051
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1362	0,1536	0,2007	0,2340	0,2456	0,2461
	6		0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0567	0,0689	0,1115	0,1596	0,1966	0,2051
	7		0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0162	0,0212	0,0425	0,0746	0,1080	0,1172
	8		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0389	0,0439
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0083	0,0098
	10		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0010

$$b)\ p(x \geq 3) = p(x=3) + p(x=4) + \ldots + p(x=9) + p(x=10)$$

$$0{,}2503 + 0{,}1460 + 0{,}0584 + 0{,}0162 + 0{,}0031 + 0{,}0004 + 0{,}0000 + 0{,}0000 =$$

$$0{,}4744$$

o también:

$$p(x \geq 3) = 1 - p(x<3) = 1 - \big(p(x=0) + p(x=1) + p(x=2)\big) =$$

$$1 - (0{,}0563 + 0{,}1877 + 0{,}2816) = 0{,}4744$$

Distribución de Poisson

- Esta es una distribución discreta de gran utilidad sobre todo en procesos biológicos, donde X suele representar el **número de eventos independientes** que ocurren a velocidad constante **en un intervalo de tiempo o en un espacio**.
- Sea X una variable aleatoria discreta, se dice que se distribuye como una distribución de Poisson de parámetro λ , X→ P(λ) con λ>0, si su función o distribución de probabilidad viene dada por:

$$p(r) = P(x = r) = \frac{\lambda^r}{r!} e^{-\lambda}; con\ r = 0, 1, \ldots$$

El proceso es estable, esto es, el número medio de sucesos por unidad de tiempo, longitud, cantidad, etc., es constante.

$E(x) = \lambda$ y $VAR(x) = \lambda$

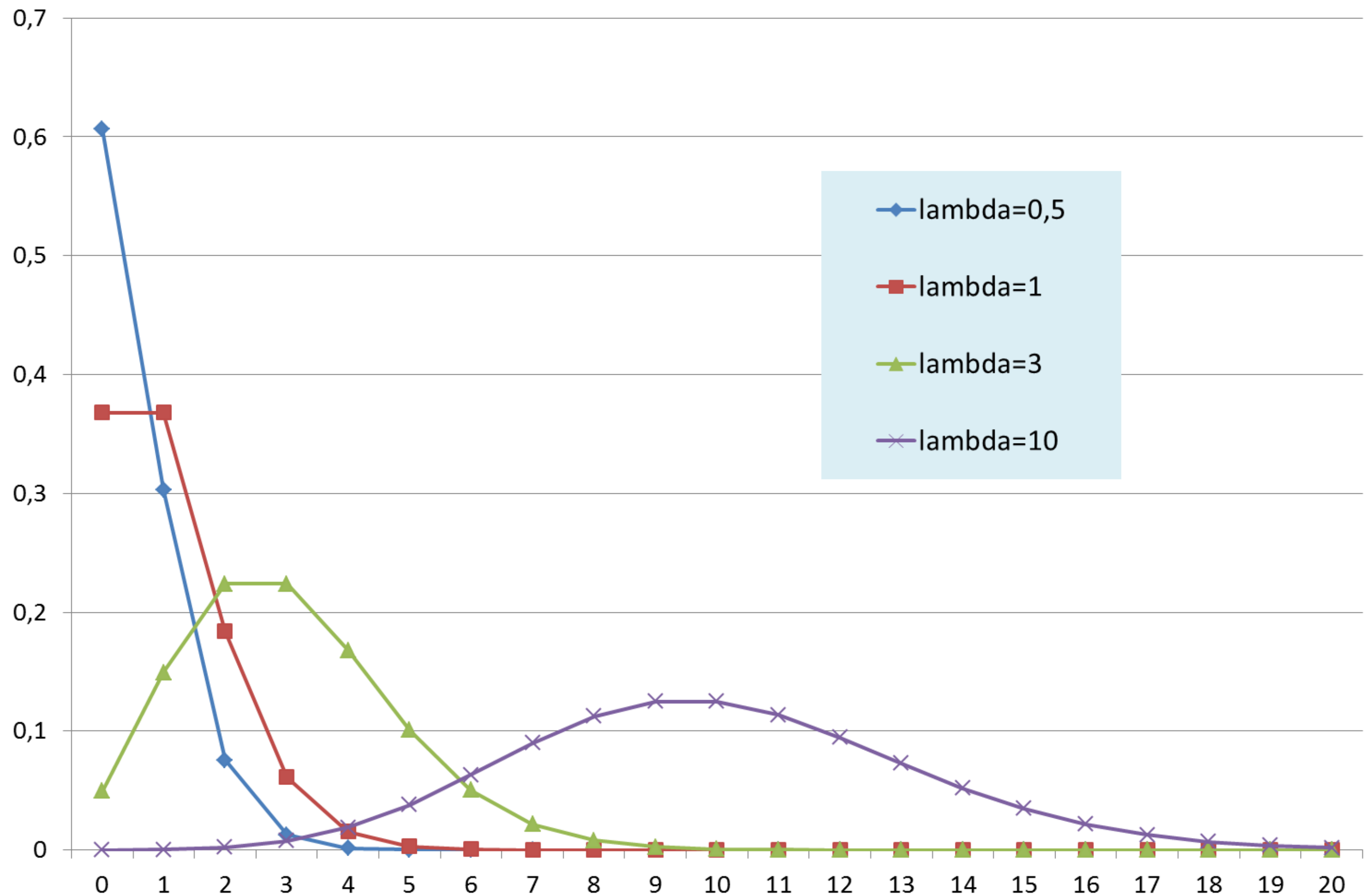
Simeón D. Poisson (1781-1840)

Distribución de Poisson

- La distribución de Poisson se utiliza en situaciones donde los sucesos son impredecibles o de ocurrencia aleatoria.
- También se utiliza cuando la probabilidad del evento que nos interesa se distribuye dentro de un segmento dado como una distancia, un área, un volumen o un tiempo definido.
- En la mayor parte de los sistemas y redes de telecomunicación el proceso de llegadas y de servicio sigue una distribución de Poisson.
- El proceso de llegadas de Poisson es el más utilizado en el diseño de colas, para modelar la generación de electrones en semiconductores y la emisión de partículas alfa en emisiones radiactivas

Distribución de Poisson

Representación de la Distribución de Poisson para diferentes valores de λ



Distribución de Poisson

- Si sumamos varias variables de Poisson independientes, el resultado es también una variable de Poisson.
  Es decir, sea Xi → P(λi), i = 1, ..., K, un conjunto de variables de Poisson independientes, entonces

  $$Y = \sum_1^k Xi$$

  $$Y \rightarrow P(\lambda *), \qquad \text{con } \lambda * = \sum_1^k \lambda i$$

- Si cada variable aleatoria Xi esta generando datos de forma independiente y con media constante, su suma sería también una sucesión de datos independiente y de media constante.

- Este resultado es útil para entender que **cuando hay independencia y estabilidad**, las propiedades estadísticas del **numero de sucesos observado por unidad de medida no depende de las unidades que utilicemos.**

Distribución de Poisson

Por ejemplo, si las llamadas telefónicas que recibe una centralita son sucesos independientes (el que una persona llame es independiente de las demás personas) y aparecen con una media de 5 llamadas al minuto.

La variable X =numero de llamadas en un minuto sería $X \rightarrow P(\lambda = 5)$

si sumamos las llamadas en 60 minutos, tendremos que la variable
Y =número de llamadas en una hora, seguiría siendo una Poisson, pues las llamadas seguirían siendo sucesos independientes y con media constante.

Ahora tendremos que $Y \rightarrow P(\lambda * = 5 \times 60)$. Por lo tanto, **la unidad de medida no afecta**.

Distribución de Poisson

Ejemplo:

Una central telefónica recibe una media de 480 llamadas por hora. Si el número de llamadas se distribuye según una Poisson y la central tiene una capacidad para atender a lo sumo 12 llamadas por minuto, ¿cuál es la probabilidad de que en un minuto determinado no sea posible dar línea a todos los clientes?

Si definimos X = "Nº de llamadas por minuto" (en una hora (480/60))

entonces X → P (8).

$$P\ (X > 12) = 1 - P\ (X \leq 12) = 1\text{-} \sum_{r=0}^{12} p(x = r) = 1\text{-} \sum_{r=0}^{12} \frac{8^r}{r!} e^{-8} =$$

$$1 - 0{,}9362 = 0{,}0638.$$

Distribución de Poisson

Ejemplo

Un servidor de una pequeña red recibe una media de 7 accesos al minuto. Suponiendo que los accesos a dicho servidor suceden de forma independiente y con un ritmo medio constante, se quiere calcular la probabilidad de que reciba más de 10 accesos en un minuto, que es el número de accesos a partir del cual el servidor tendría un rendimiento deficiente. Con las hipótesis de estabilidad e independencia se tiene que la variable

X =número de accesos en un minuto será una Poisson de media 7: X → P(7)

$$
\begin{aligned}
P(X > 10) &= 1 - P(X \leq 10) = 1 - \sum_{r=0}^{10} P(X = r) \\
&= 1 - \sum_{r=0}^{10} \frac{7^r}{r!} e^{-7} = 0{,}099,
\end{aligned}
$$

Entonces, más del 90% de los minutos recibe 10 accesos o menos.
Por tanto, el 10% del tiempo, el servidor está trabajando por encima de su capacidad.

Distribución de Poisson

Ejercicio

A una gasolinera llegan, de media, 3 coches por minuto. Determinar:

a. Cuál es la función de probabilidad.

b. Probabilidad de que en un minuto lleguen 2 coches.

c. Probabilidad de que en cinco minutos lleguen 12 coches.

Solución:

- Es un proceso que cuantifica el número *n* de elementos de la población que se observan en un intervalo de duración fija.
- Como observamos 1 minuto y en un minuto hay una media de 3 llegadas, estamos buscando números medios de llegadas en diferentes intervalos.

Distribución de Poisson

Solución:

a. En consecuencia la función de probabilidad es una de Poisson con $\lambda = 3$:

$$p(r) = P(x = r) = \frac{3^r}{r!} e^{-3}$$

b. 2 llegadas en 1 minuto:

$$p(2) = P(x = 2) = \frac{3^2}{2!} e^{-3} = 0{,}224 \rightarrow 22{,}4\%$$

b. 12 llegadas en 5 minutos: $\lambda' = \lambda * 5 = 3 * 5 = 15$

$$p(12) = \frac{15^{12}}{12!} e^{-15} = 0{,}0828 \rightarrow 8{,}28\%$$

Distribución de Poisson

- **EJERCICIO**

En promedio, cada uno de los 10 empleados de una empresa de servicios hace 3 llamadas telefónicas al día en una jornada laboral de 8 horas. Si se supervisa el número de llamadas cada hora.

A. ¿Cuál es el número medio de llamadas telefónicas que se registrará en cada hora?

B. ¿Con qué probabilidad encontraremos *x* llamadas para valores de *x*=0, 1, 2 y 3?

C. ¿Y la probabilidad de que $x \geq 4$?

Suponemos que la va X sigue una distribución de poisson de parámetro λ, es decir:

$$P_\lambda(x) = e^{-\lambda} \frac{(\lambda)^x}{x!}$$

Distribución de Poisson

SOLUCIÓN

A. Calculamos el promedio =μ = λ

λ = 10 empleados*3 llamadas/empleado/8 horas=3,75 llamadas/hora

B.

$$P_\lambda(x) = e^{-\lambda}\frac{(\mu)^x}{x!} = e^{-3,75}\frac{(3,75)^x}{x!}$$

$$P_\lambda(0)\ [\%] = e^{-3,75}\frac{(3,75)^0}{0!}\cdot 100 = 2,35$$

$$P_\lambda(1)\ [\%] = e^{-3,75}\frac{(3,75)^1}{1!}\cdot 100 = 8,82$$

$$P_\lambda(2)[\%] = e^{-3,75}\frac{(3,75)^2}{2!}\cdot 100 = 16,54$$

$$P_\lambda(3)[\%] = e^{-3,75}\frac{(3,75)^3}{3!}\cdot 100 = 20,67$$

$$P_\lambda(4)[\%] = e^{-3,75}\frac{(3,75)^4}{4!}\cdot 100 = 19,38$$

$$P_\lambda(5)[\%] = e^{-3,75}\frac{(3,75)^5}{5!}\cdot 100 = 14,53$$

$$P_\lambda(6)[\%] = e^{-3,75}\frac{(3,75)^6}{6!}\cdot 100 = 9,08$$

…

Distribución de Poisson

C. $\mathbf{P(x\geq 4) = 1- P(X<4)} = 1 - \sum_{0}^{3} e^{-3,75} \frac{(3,75)^x}{x!} =$

$$= 1 - \left[e^{-3,75} \frac{(3,75)^0}{0!} + e^{-3,75} \frac{(3,75)^1}{1!} + e^{-3,75} \frac{(3,75)^2}{2!} + e^{-3,75} \frac{(3,75)^3}{3!} \right]$$

$$\boxed{P_{3.75}(x \geq 4)[\%] = 51.62}$$

(También se dispone de tablas para el calculo de estos valores)

DISTRIBUCIONES CONTINUAS

Distribución Uniforme Continua

Es la más sencilla de las distribuciones continuas.

Surge cuando consideramos una variable aleatoria que toma valores en un intervalo finito de manera equiprobable.

Una variable aleatoria continua, X, se dice que se distribuye como una **distribución uniforme de parámetros a, b**, tales que $-\infty < a < b < +\infty$: X → U (a,b),

la función de densidad f(x) y la función de distribución F(x) son :

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

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

$$E(x) = \frac{a+b}{2}$$

$$\sigma_x^2 = Var(x) = \frac{(b-a)^2}{12}$$

Distribución Uniforme Continua

<u>Ejemplo</u>: Seleccionamos al azar un número real en el intervalo [2, 6] y definimos una variable aleatoria como X="número seleccionado". Calcula la probabilidad de que el número seleccionado sea menor de 5 y el número esperado.

En este caso $X \to U(2,6)$; Para calcular la probabilidad lo que hacemos es

$$P[X \leq 5] = \int_2^5 f(x)dx = \int_2^5 \frac{1}{6-2}dx = \int_2^5 \frac{1}{4}dx = \left.\frac{x}{4}\right]_2^5 = \frac{5}{4} - \frac{2}{4} = \frac{3}{4} = 0.75.$$

Esto se podía haber hecho más rápido con la función de distribución de la siguiente forma:

$$P[X \leq 5] = F(5) = \frac{x-2}{b-a} = \frac{5-2}{6-2} = \frac{3}{4} = 0.75.$$

Para calcular la esperanza, aplicamos la formula y nos queda,

$$E[X] = \frac{a+b}{2} = \frac{2+6}{2} = \frac{8}{2} = 4.$$

Distribución Exponencial

Esta distribución suele ser modelo de aquellos fenómenos aleatorios que miden el **tiempo que transcurre entre dos sucesos**. Por ejemplo, entre la puesta en marcha de un cierto componente y su fallo o el tiempo que transcurre entre dos llegadas consecutivas a un cajero automático.

Sea X una v.a. continua que puede tomar valores $x \geq 0$. *Se dice que X sigue una distribución exponencial de parámetro* λ *y se nota* $X \rightarrow \exp(\lambda)$ *si su función de densidad y de distribución son:*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & en\ otro\ caso \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$E(x) = \frac{1}{\lambda} \qquad Var(x) = \frac{1}{\lambda^2}$$

Distribución normal o gaussiana

La distribución normal fue descrita, por primera vez, en 1773 por De Moivre como límite de la distribución Binomial cuando el número de ensayos tiende a infinito.
Este hecho no llamó la atención y la distribución normal fue redescubierta" por **Laplace (1812) y Gauss (1809).** Ambos se ocupaban de problemas de Astronomía y Geodesia; cada uno obtuvo empíricamente la distribución normal (también llamada distribución gaussiana) como modelo que describe el comportamiento de los errores en las medidas astronómicas y geodésicas.

Carl Friedrich *Gauss*
(1777-1855)

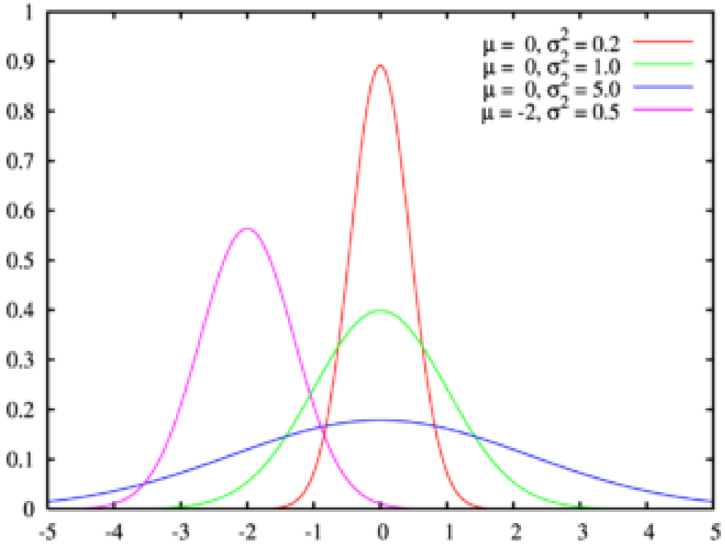
Es una de las distribuciones más importantes. Es el modelo de distribución más utilizado en la práctica, ya que multitud de fenómenos se comportan según una distribución normal.
Esta distribución de caracteriza porque los valores se distribuyen formando **una campana de Gauss**, en torno a un valor central que coincide con el **valor medio de la distribución**.
Las ventajas teóricas de este modelo hacen que su uso se generalice en las aplicaciones reales.

Distribución normal o gaussiana

- Se dice que una variable aleatoria X sigue una distribución normal de parámetros μ, σ , $X \rightarrow N(\mu, \sigma^2)$ o $X \rightarrow N(\mu, \sigma)$, $X \epsilon \mathbb{R}$, $\mu \ \epsilon \mathbb{R}$ y $\sigma > 0$

si su función de densidad tiene la siguiente forma:

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



La esperanza y la varianza:

$E(x) = \mu$

$Var(x) = \sigma^2$

Si una variable X_1 es $N(\mu_1, \sigma_1)$ y otra X_2 es $N(\mu_2, \sigma_2)$ independientes entre sí, entonces la nueva variable $X = X_1 \pm X_2$ sigue también una distribución normal $N\left(\mu_1 \pm \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$. Propiedad que se puede generalizar a n variables aleatorias independientes.

Teorema central del límite

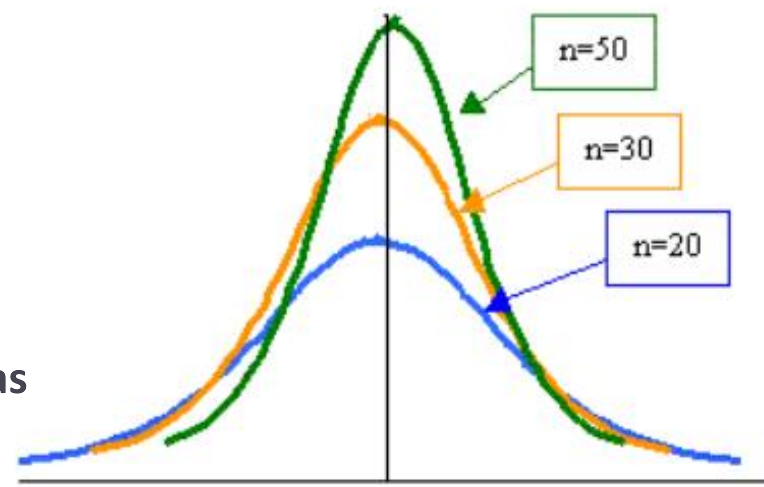
- Si tenemos $X_1, X_2, \cdots, X_n$ variables aleatorias independientes con media μ_i y varianza σ_i^2, y formamos la variable suma:

$$Y = X_1 + X_2 + \cdots + X_n$$

- Cuando *n* crece la distribución de Y se acerca a una distribución normal:

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

- Esto quiere decir que si tenemos una variable que es suma de otras:
  - La **variable suma** se distribuye **normalmente**
  - La **media de la suma** es la **suma de las medias**
  - La **varianza** es la **suma de las varianzas**



Distribución normal o gaussiana

Propiedades de la función de densidad

- Es contínua en todo $\mathbb{R}$.
- La moda (punto donde se alcanza el máximo) es $Mo = \mu$ y el valor de la función en este punto es $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$.
- Es estrictamente creciente para $x < \mu$ y estrictamente decreciente para $x > \mu$.
- Posee dos puntos de inflexión: $\mu + \sigma$ y $\mu - \sigma$. La situación de estos puntos determina la forma de la curva de modo que cuanto mayor es σ más lejos de la moda están los puntos de inflexión y más plana será la curva.
- Tiene como asíntota horizontal el eje de abcisas, ya que

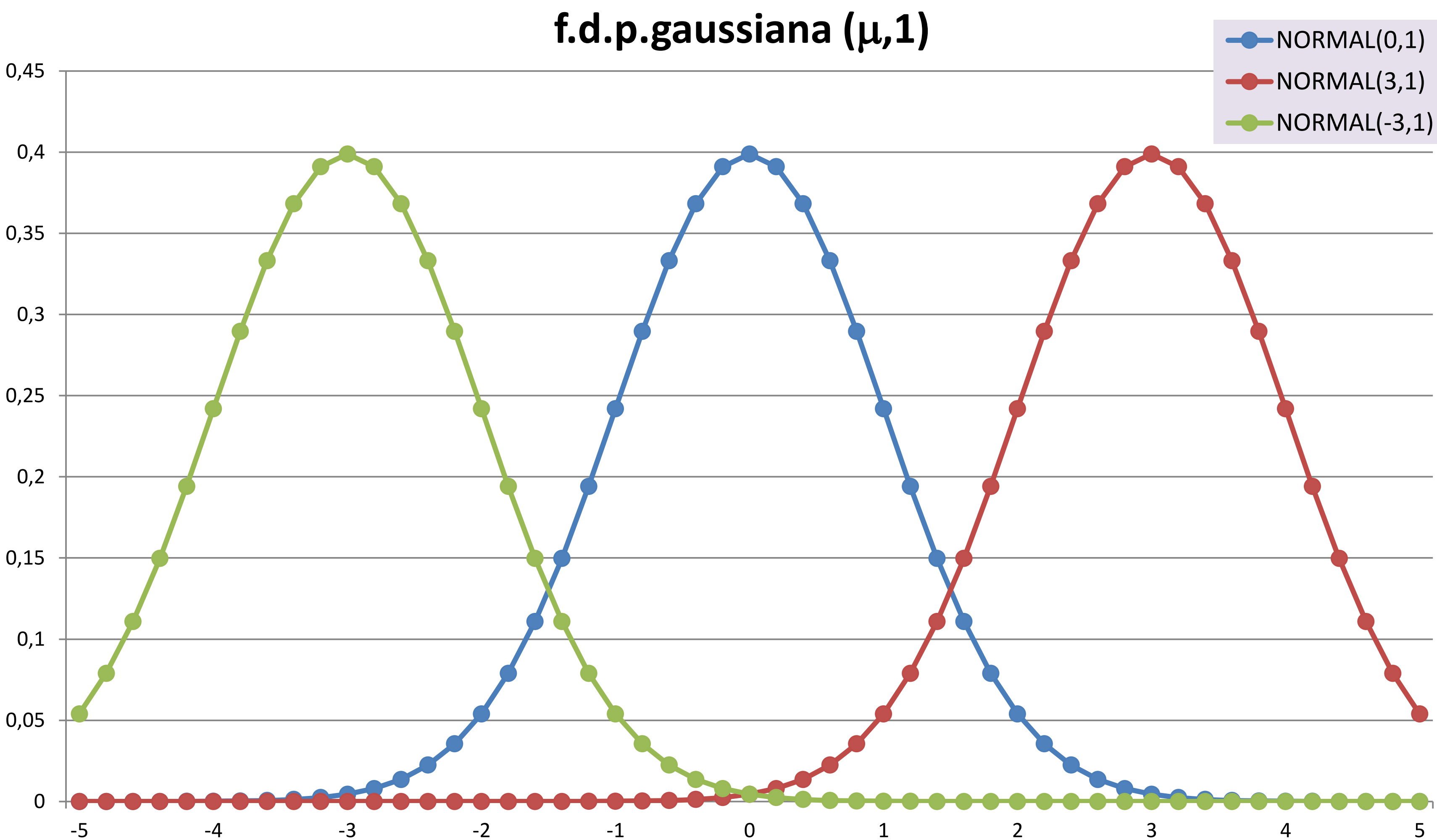
$$\lim_{x \to -\infty} f(x) = \lim_{x \to \infty} f(x) = 0.$$

- Es simétrica respecto a $x = \mu$, esto es

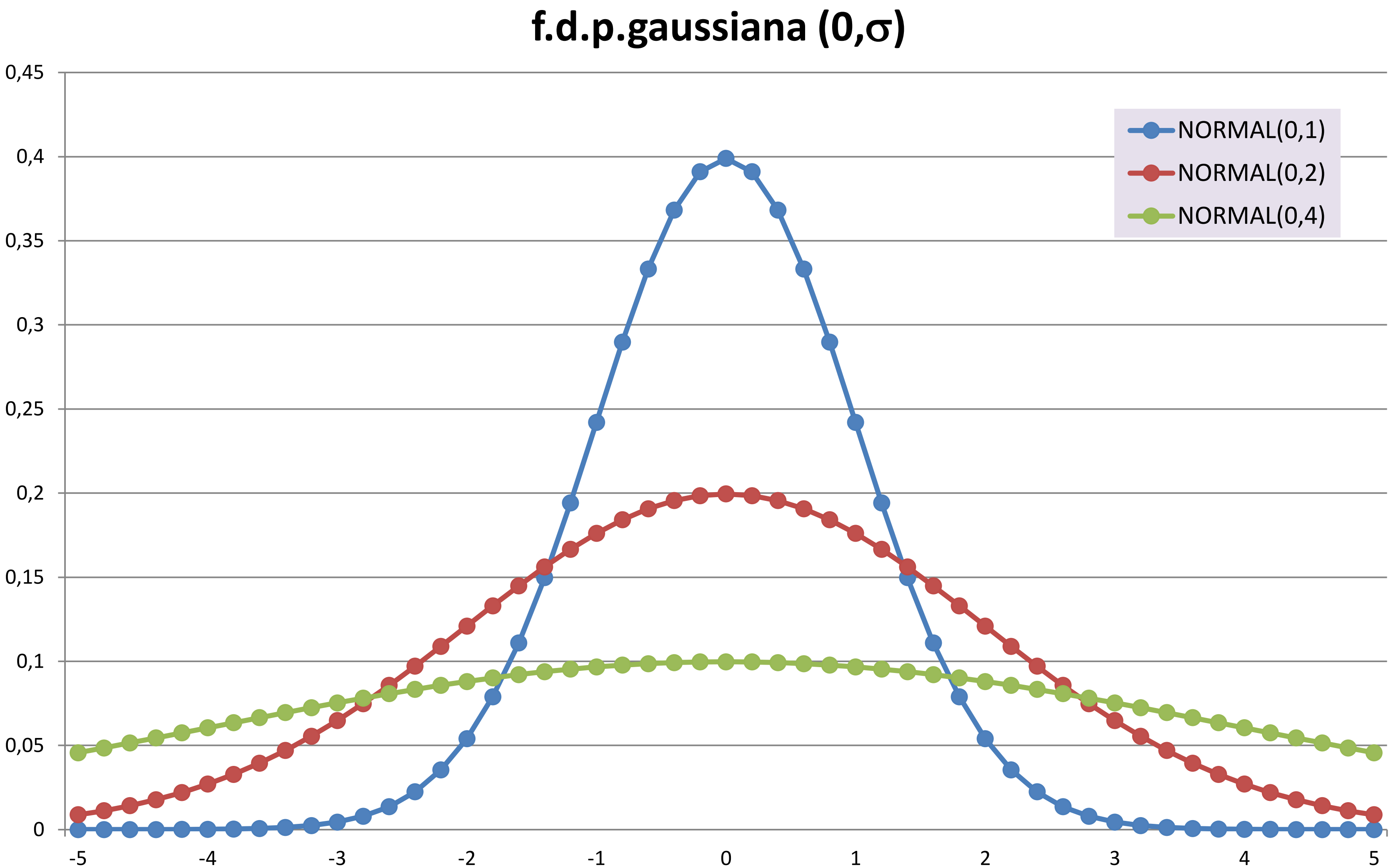
$$f(x - \mu) = f(x + \mu), \ \forall x \in \mathbb{R}$$

Por tanto, la mediana es también μ.

Distribución normal



Distribución normal

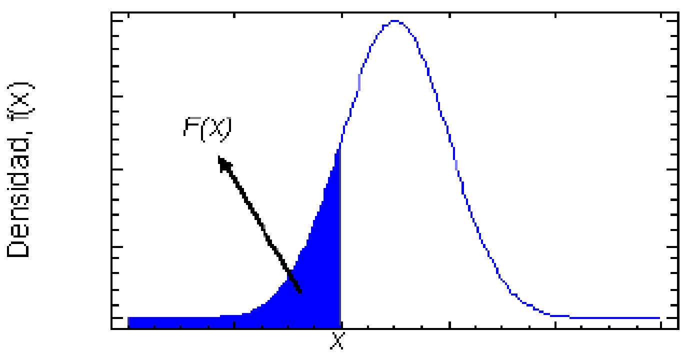


Distribución normal

Función de distribución normal:

- La función de distribución de la distribución normal está definida como sigue:

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt; \quad -\infty < x < \infty$$



La zona coloreada representa el valor, $F(x)$, de la función de distribución en el punto x

Distribución Normal estándar o tipificada

Se denomina **tipificación** a la transformación de una v.a. normal de parámetros, μ, σ , a la distribución normal estándar de parámetros 0,1.

la variable X → N (μ, σ) se transforma mediante la siguiente expresión:

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

Y la variable *z* sigue una Normal Estándar con $\mu \quad = 0$ y $\sigma \quad = 1$ con lo que la ecuación la función de densidad:

Z → N(0,1)

$$f(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right) \qquad z \epsilon \mathbb{R}$$

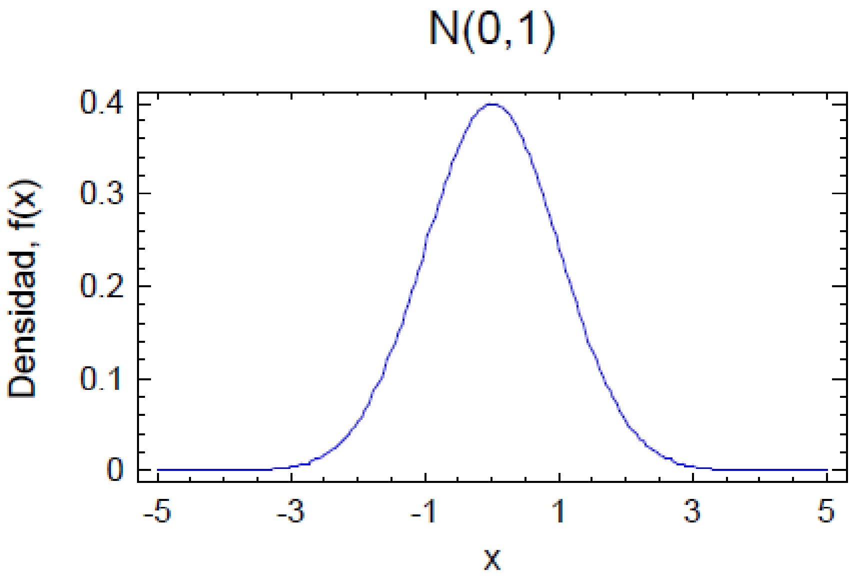
La Función de Distribución de la N(0,1) sería:
y está tabulada

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right)$$

Distribución Normal estándar o tipificada

La Figura muestra la representación grafica de la función de densidad de una N(0; 1). En ella podemos observar las siguientes características:

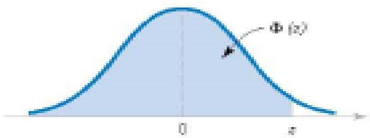
- Es simétrica respecto a $x = 0$.
- Alcanza una máximo en $x = 0$ que vale $f(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4$.
- Es creciente para $x < 0$ y decreciente para $x > 0$.
- $1, -1$ son puntos de inflexión.
- $y = 0$ es una asíntota horizontal.
- $F(-x) = 1 - F(x)$



Distribución Normal – Manejo de tablas

FUNCIÓN DE DISTRIBUCIÓN DE UNA N(0,1) PARA VALORES POSITIVOS

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201

Distribución Normal – Manejo de tablas

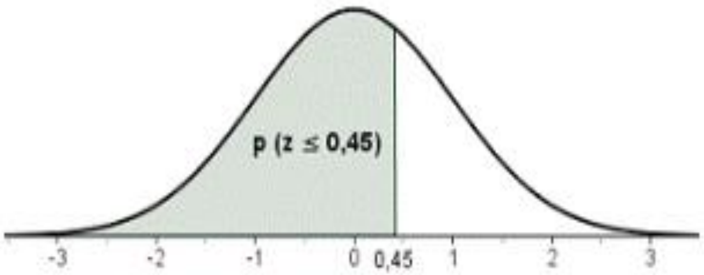
En nuestro caso, vamos a manejar una tabla de la distribución normal estándar que contiene los valores de su función de distribución para valores positivos, es decir, $P[Z \leq a]$ con $a \geq 0$. A partir de dicha tabla también se podrán calcular directamente aquellos valores $a \geq 0$ para los cuales aparezcan la $P[Z \leq a]$. Notemos que todas las probabilidades que aparecen en la tabla son mayores o iguales a 0.5, al no aparecer valores de la función de distribución para valores negativos.

A partir de dicha tabla también se puede obtener, para valores $a \geq 0$:

- $P[Z \geq a] = 1 - P[Z \leq a]$
- $P[Z \leq -a] = P[Z \geq a] = 1 - P[Z \leq a]$
- $P[Z \geq -a] = P[Z \leq a]$
- $P[|Z| \leq a] = P[-a \leq Z \leq a] = P[Z \leq a] - P[Z \leq -a] = 2P[Z \leq a] - 1$

Distribución Normal – Manejo de tablas

1 Cuando la probabilidad pedida se encuentra directamente en la tabla.
Hallar la probabilidad p (z ≤ 0,45).
En la tabla podemos leer directamente la probabilidad de valores menores o iguales que un número positivo.

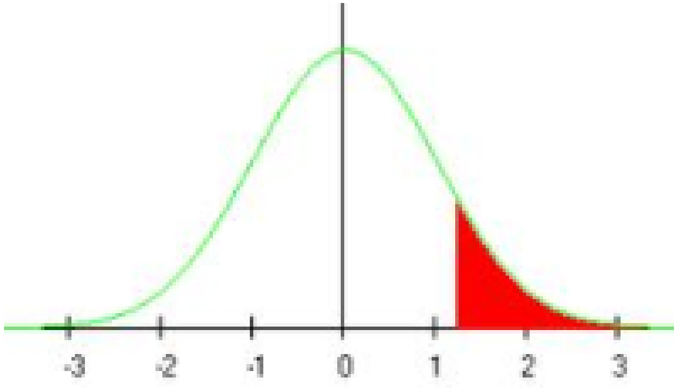


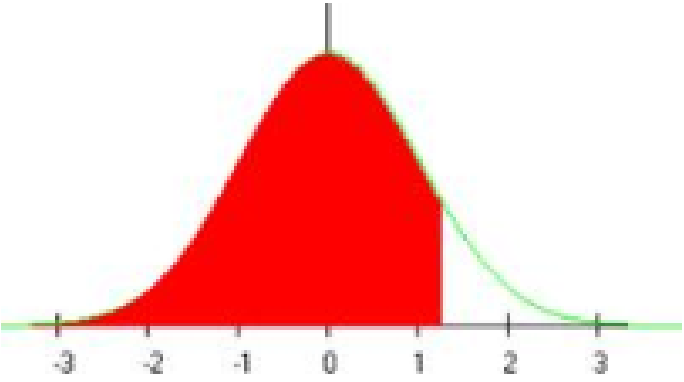
En la 1ª columna z buscamos el valor de las unidades y las decenas.
En la fila correspondiente al valor de la columna buscamos el valor de las centésimas.
Basta buscar 0,4 en la columna y 0,05 en la fila.
Su intersección nos da la probabilidad.
Leemos y nos da 0,6736.
La probabilidad p (z ≤ 0,45) = 0,6736

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123

Distribución Normal – Manejo de tablas

2 Probabilidad de un valor positivo p (z > 1,24)





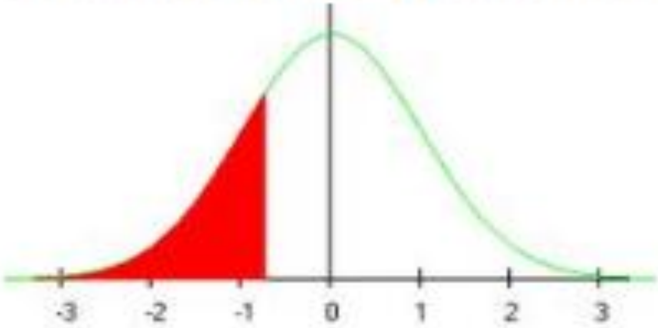
En este caso la probabilidad pedida no está en las tablas.
Sin embargo, si tenemos en cuenta que el área total bajo la gráfica ha de ser 1, deducimos de la figura que:

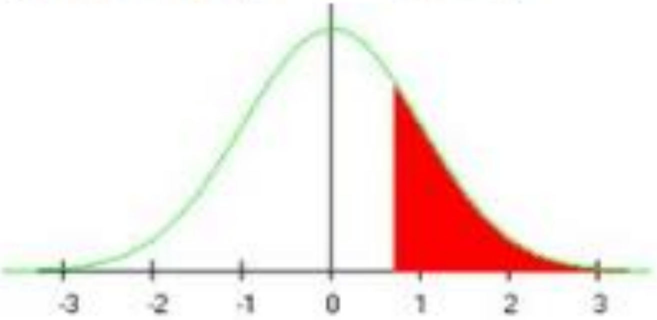
$p(z > 1,24) = 1 - p(z \leq 1,24) = 1 - 0,8925 = 0,1075$

En la tabla leemos $p(z \leq 1,24) = 0,8925$

Distribución Normal – Manejo de tablas

3 Probabilidad de un valor negativo p ($z \leq -0,72$)





Como la gráfica es simétrica respecto al eje de ordenadas,

$p(z \leq -0,72) = p(z \geq +0,72)$

Calculamos $p(z \geq +0,72)$ igual que en el caso 2.

$p(z \geq +0,72) = 1 - p(z < +0,72) = 1 - 0,7642 = 0,2358$

$p(z \leq -0,72) = p(z \geq +0,72) = 1 - p(z < +0,72) =$

$1 - 0,7642 = 0,2358$

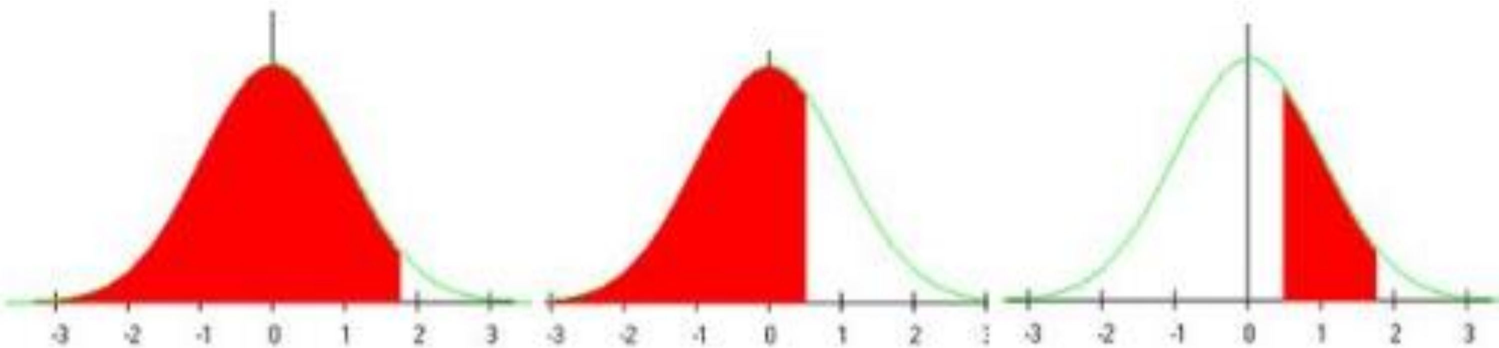
Distribución Normal – Manejo de tablas

4 Probabilidad entre dos valores positivos $p(0,5 \leq z \leq 1,76)$

Leemos directamente en la tabla la $p(z \leq 1,76)$ y la $p(z \leq 0,5)$.

La diferencia entre ellas es la probabilidad que nos piden.

$p(0,5 \leq z \leq 1,76) = p(z \leq 1,76) - p(z \leq 0,5) = 0,9608 - 0,6915 = 0,2693$



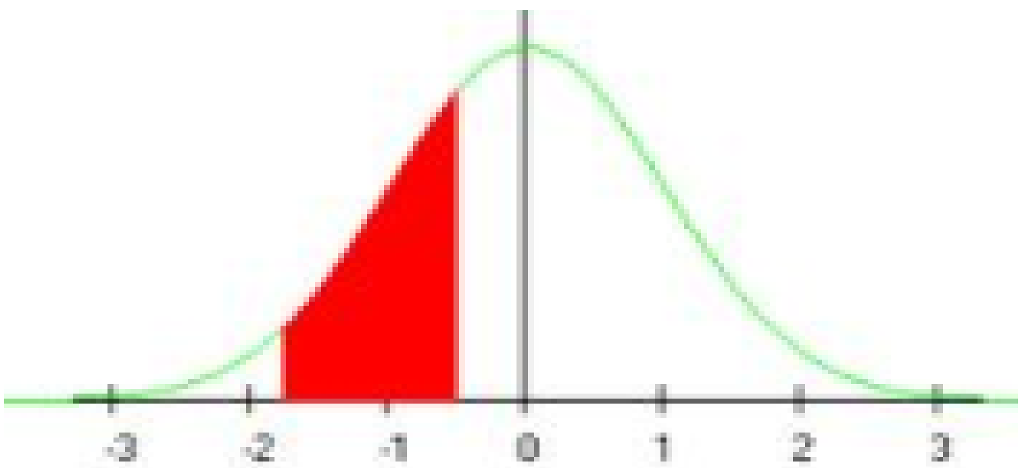
Distribución Normal – Manejo de tablas

5 Probabilidad entre dos valores negativos $p(-1{,}76 \le z \le -0{,}5)$

Por simetría cambiamos los dos valores negativos a positivos y calculamos sus probabilidades.

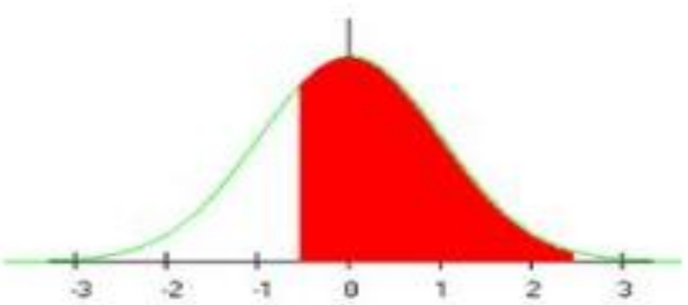
$P(-1{,}76 \le z \le -0{,}5) = p(0{,}5 \le z \le 1{,}76) = 0{,}9608 - 0{,}6915 = 0{,}2693$

Observa que el área sombreada es la misma que en el caso 4



Distribución Normal – Manejo de tablas

6 Probabilidad entre un valor positivo y uno negativo
$p(-0{,}53 \le z \le 2{,}46)$



$p(-0{,}53 \le z \le 2{,}46) = p(z \le 2{,}46) - p(z \le -0{,}53)$

Calculamos $p(z \le -0{,}53)$ igual que en el caso 3

$p(z \le -0{,}53) = p(z \ge 0{,}53) = 1 - p(z < 0{,}53) =$

$1 - 0{,}7019 = 0{,}2981$

La $p(z \le 2{,}46)$ la leemos directamente en la tabla.

$p(z \le 2{,}46) = 0{,}9931$

$p(-0{,}53 \le z \le 2{,}46) = p(z \le 2{,}46) - p(z \le -0{,}53) =$

$0{,}9931 - 0{,}2981 = 0{,}695$

Distribución Normal – Manejo de tablas

Uso de la tabla de forma inversa

Nos dan la probabilidad y calculamos el valor de la variable z que acumula dicha probabilidad.

Hallar z de la normal N(0,1) en cada caso:

- $p(z < k) = 0{,}7019$

Se busca en la parte central de la tabla normal N (0,1) la probabilidad 0,7019, observando que el valor de la 1ª columna es 0,5 y el valor de la primera fila es 0,03.

$p(k) = 0{,}7019 \Rightarrow k = 0{,}53$

- $p(z < k) = 0{,}8997$

Valor de la primera columna 1,2

valor de la primera fila 0,08.

$p(k) = 0{,}8997 \Rightarrow k = 1{,}28$

Distribución Normal – Manejo de tablas

Ejercicios resueltos

Sea z una variable normal N (0,1). Calcula:

a) $p(z \leq 2{,}17)$

$p(z \leq 2{,}17) = 0{,}9850$

b) $p(z \geq 1{,}32)$

$p(z \geq 1{,}32) = 1 - p(z < 1{,}32) = 1 - 0{,}9066 = 0{,}0934$

c) $p(1{,}52 < z \leq 2{,}03)$

$p(1{,}52 < z \leq 2{,}03) = p(z \leq 2{,}03) - p(z \leq 1{,}52) =$

$0{,}9788 - 0{,}9357 = 0{,}0431$

d) $p(z \geq -1{,}32)$

$p(z \geq -1{,}32) = p(z \leq 1{,}32) = 0{,}9066$

Distribución Normal - ejemplos

Ejemplo 1.

La vida de un semiconductor láser a una potencia constante se distribuye normalmente con media 7000 horas y desviación típica 600 horas. ¿Cuál es la probabilidad de que la vida del láser esté entre 6280 y 7120 horas?

$X \equiv$ vida del semiconductor (en horas) $\sim N(7000, 600)$.

Ejemplo 2.

Se quiere dar una beca a uno de dos estudiantes de sistemas educativos diferentes. Se asignará al que tenga mejor expediente académico.

a) El estudiante A tiene una calificación de 8 en un sistema donde la calificación de los alumnos se comporta como N(6,1).
b) El estudiante B tiene una calificación de 80 en un sistema donde la calificación de los alumnos se comporta como N(70,10).

Distribución normal

Ejemplo 1

La vida de un semiconductor láser a una potencia constante se distribuye normalmente con media 7000 horas y desviación típica 600 horas. ¿Cuál es la probabilidad de que la vida del láser esté entre 6280 y 7120 horas?

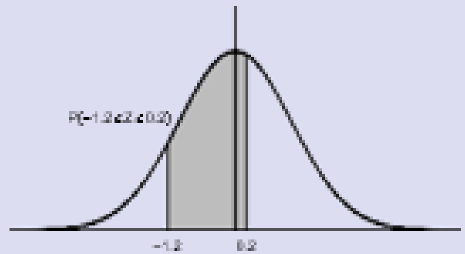
$X \equiv$ vida del semiconductor (en horas) $\sim N(7000, 600)$ Tipificación:

$$Z = \frac{X - 7000}{600} \sim N(0,1), \quad \frac{6280 - 7000}{600} = -1.2, \quad \frac{7120 - 7000}{600} = 0.2$$

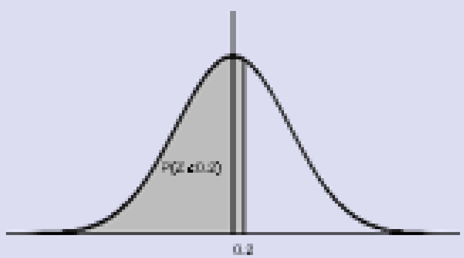
$$P(6280 \le X \le 7120) = P(-1.2 \le Z \le 0.2)$$
$$= P(Z \le 0.2) - P(Z \le -1.2) = 0.5793 - 0.1151 = 0.4642$$

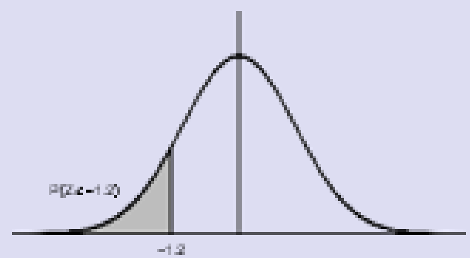
Distribución normal

Veamos una interpretación gráfica del cálculo anterior. La probabilidad que tenemos que calcular es igual al siguiente área:



Dicho recinto está incluido en el del gráfico que aparece a la izquierda. La parte que sobra es precisamente la que está sombreada en el gráfico de la derecha:





$$P(-1.2 \leq Z \leq 0.2) = P(Z \leq 0.2) - P(Z \leq -1.2)$$

Distribución normal

- **Ejemplo 2**
- 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)

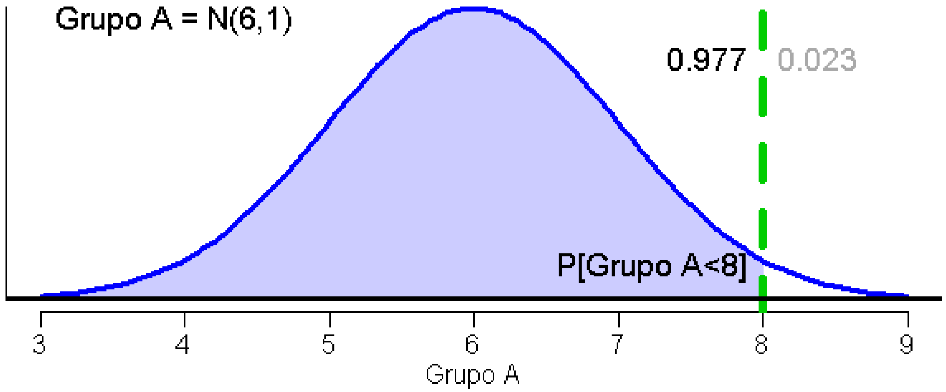
$$z_A = \frac{x_A - \mu_A}{\sigma_A} = \frac{8-6}{1} = 2$$

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

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 porcentaje de compañeros del mismo sistema de estudios que ha superado B.

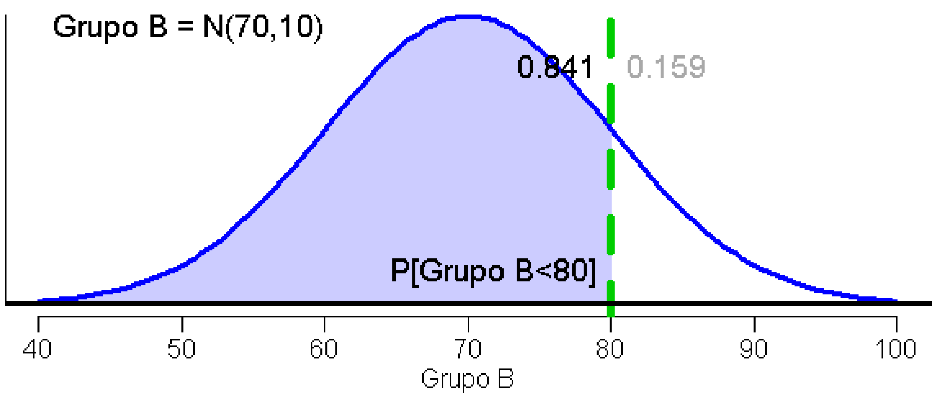
Podríamos pensar en principio que A es mejor candidato para la beca.

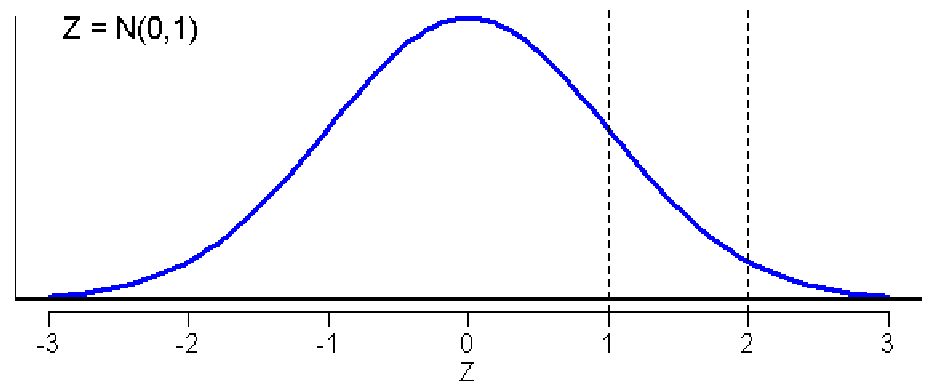
Distribución normal



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

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





Distribución normal

- **Ejercicio**
- Se sabe que, debido a los procesos de llenado, el contenido de una lata de bebida de 33 cl. No es exactamente de 33 cl. en todas las latas, sino que se distribuye normalmente con una media de 33cl y una desviación típica de 2cl. Determinar:
  a. La función de densidad.
  b. ¿Cuál es la probabilidad de que el contenido de una lata sea superior a 35 cl?
  c. Si un pack consta de 6 latas ¿cuál es la probabilidad de que el contenido sea inferior a 192 cl?

Distribución normal

Solución:

El enunciado indica que los valores de la variable contenido de la lata se distribuyen normalmente con una desviación de 2 cl. alrededor de una media de 33 cl. En consecuencia,

a. la función de densidad es una Normal (33, 2): $p(x_i) = f(x = x_i) = \frac{1}{2\sqrt{2\pi}} exp\left(-\frac{1}{8}(x_i - 33)^2\right)$

b. Contenido de una lata superior a 35 cl.:

$$p(x \geq 35) = p\left(z \geq \frac{35-33}{2}\right) = 1 - p(z \leq 1) = 1 - 0{,}8413 = 0{,}1587 \rightarrow 15{,}87\%$$

c. Contenido de 6 latas inferior a 192 cl. Por el Teorema Central del Límite: La suma de *n* variables que se distribuyen normalmente μ, σ, también tiene una distribución normal con media $n.\mu$ y desviación típica $\sigma\sqrt{n}$. Cada lata se distribuye como una N (33, 2), donde 2 es la desviación típica.

Sea Y = 6 latas, Y se distribuyen como N (6*33=198, $\sqrt{6}$*2) = N (198, 4,899)

$p(y < 192) = p\left(z < \frac{192-198}{\sqrt{6}.2}\right) = p(z < -1{,}22) = p(z > 1{,}22) = 1 - p(z < 1{,}22) = 1 - 0{,}8888 =$ $11{,}12\%$

Relaciones entre modelos de distribución

Aproximación de una Binomial por una Poisson

Sea X una variable aleatoria discreta que se distribuye como una Binomial con parámetros (n,p) donde **n** tiende a infinito y, **p** tiende a 0. Cuando esto ocurre podemos aproximar una distribución Binomial por medio de una distribución de Poisson, es decir,

$$X \to P(\lambda = np).$$

Por convenio se realizará esto cuando se verifiquen una de estas condiciones:

1. Cuando se verifique $n > 30$ y $p < 0'1$.
2. $n \cdot p < 5$.

Relaciones entre modelos de distribución

Aproximación de una Binomial por una Normal

Sea X una variable aleatoria discreta que se distribuye como una Binomial con parámetros (n,p), entonces De Moivre demostró que cuando $n \to \infty$ y, **p** es aproximadamente 0´5, esa variable aleatoria se puede aproximar como una distribución normal. El criterio que se toma es que $n > 50$ y $p \cong 0'5$.

Cuando esto ocurre se verifica que

$$X \to B(n,p) \text{ se dice que } X \to N(\mu = np; \sigma = \sqrt{npq}).$$

Aproximación de una distribución de Poisson por una Normal

Sea X una variable aleatoria discreta que se distribuye como una Poisson de parámetro (λ), se demuestra que cuando λ es muy grande, se puede aproximar por medio de una distribución normal, como ocurría anteriormente. Así, si

$$X \to P(\lambda) \text{ y } \lambda \to \infty \text{ entonces } X \to N\left(\mu = \lambda; \sigma = \sqrt{\lambda}\right).$$

La condición es que se verifique $\lambda > 16$.

Relaciones entre modelos de distribución

- **Ejemplo:**

Una prueba consta de 200 preguntas de verdadero o falso, para un sujeto que respondiese al azar.

¿Cuál sería la probabilidad de que acertase?:

a) 50 preguntas o menos
b) Más de 50 y menos de 100
c) Más de 120 preguntas

Relaciones entre modelos de distribución

- **Solución:** El número de preguntas sigue una distribución binomial con *n*=200 y *p*=0,5. Como n es elevado, esta distribución **se puede aproximar por una normal de media y varianza:**

$$\mu = n \cdot p = 200 \times 0.5 = 100 \quad \sigma = \sqrt{200 \times 0.5 \times 0.5} = 7{,}07$$

- a) $P(x \leq 50) \approx P(x \leq 50{,}5) = P\left(Z \leq \frac{50{,}5 - 100}{7{,}07}\right) = P(Z \leq -7) \approx 0$

- b) $P(50 < x < 100)) = P(x \leq 99{,}5) - P(x \leq 51) = P\left(Z \leq \frac{99{,}5 - 100}{7{,}07}\right) - P\left(Z \leq \frac{50{,}5 - 100}{7{,}07}\right) =$

  $P(Z \leq -0{,}07) - P(Z \leq -7) = 0{,}4721 = 47{,}21\%$

- c) $P(x > 120) = P\left(Z > \frac{120{,}5 - 100}{7{,}07}\right) = 1 - P(Z \leq 2{,}9) = 1 - 0.9981 = 0{,}0019 = 0{,}19\%$

Relaciones entre modelos de distribución

- **Ejemplo:**

Un banco recibe en promedio 6 cheques falsos al día, suponiendo que el número de cheques falsos sigue una distribución de Poisson ¿Cuál es la probabilidad de que se reciban más de 30 cheques falsos en una semana?

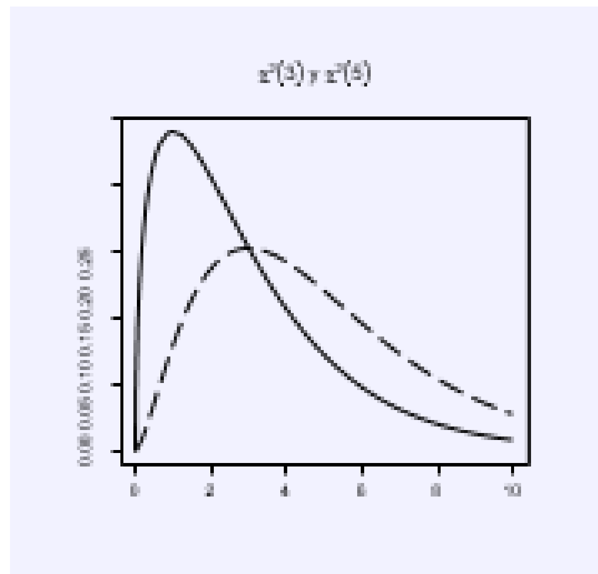
- Solución:
  - Para una semana: $\lambda = 6 \times 7 = 42 > 10$
  - La distribución de Poisson se puede aproximar por una normal por tanto

$$P(x > 30) = P\left(Z > \frac{30-42}{\sqrt{42}}\right) = P(Z > -1.85) = P(Z < 1,85) = 0,96 = 96\%$$

Distribución chi-cuadrado ($\chi^2(n)$)

La distribución de probabilidad **chi-cuadrado con n grados de libertad (χ2(n))** es la distribución asociada a una variable aleatoria que se obtiene como suma de los cuadrados de n variables independientes con distribución N(0, 1).
Por tanto, esta distribución sólo toma valores positivos y su función de densidad es muy compleja.

En este gráfico aparecen representadas las funciones de densidad de una χ2(3) (línea continua) y una χ2(5) (línea discontinua)



Intuitivamente, esta distribución es de utilidad para obtener información de la varianza poblacional a partir de un conjunto de datos extraídos de una variable normal.

Distribución chi-cuadrado ($\chi^2(n)$)

Si se tienen n variables aleatorias $x_1 \dots x_n$ independientes y distribuidas N(0, 1), y se define a partir de ellas la variable aleatoria $x = x_1^2 + \cdots + x_n^2$.

x que sigue una distribución de probabilidad χ^2 de Pearson con n grados de libertad.

La función de densidad de x es:

$$f(x) = \begin{cases} \dfrac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} si\ x \geq 0 \\ 0\ si\ x < 0 \end{cases}$$

La Función de Distribución x está tabulada

Distribución chi-cuadrado ($\chi^2(n)$)

Propiedades:

- Es una función asimétrica.
- E(x)= n.
- V(x)=2n.
- Sean dos variables aleatorias chi-cuadrado que se distribuyen $X_1 \rightarrow \chi^2_n$ y $X_2 \rightarrow \chi^2_m$, se define una nueva variable de la forma Y = X_1 + X_2, entonces esta nueva variable se distribuye como:

$$Y \rightarrow \chi^2_{n+m}$$

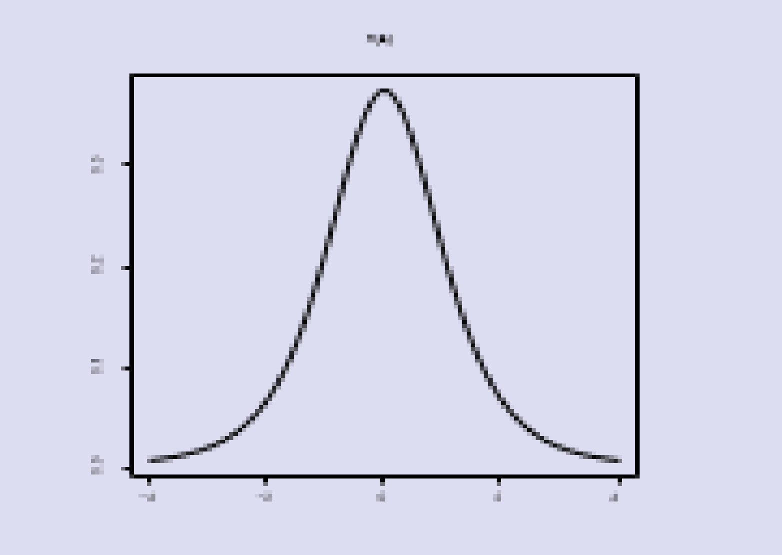
- Cuando el número de variables aleatorias es muy grande, es decir, cuando $n \rightarrow \infty$, la variable se puede aproximar por una normal.

Distribución t de Student

La distribución de probabilidad **t de Student** con *n* grados de libertad ($t(n)$), es la distribución asociada a una variable aleatoria que se obtiene a partir del cociente de una variable $N(0, 1)$ y la raíz cuadrada de una variable $\chi^2(n)$. Por tanto, esta distribución puede tomar cualquier valor real.

Su función de densidad es muy compleja y su gráfica es parecida a la de la distribución $N(0, 1)$.

En el gráfico aparece representada la función de densidad de una $t(4)$:



Intuitivamente, esta distribución es de utilidad para obtener **información o establecer comparaciones entre las medias poblacionales a partir de uno o dos conjuntos de datos extraídos de una variable normal**

Distribución t de Student

Si se tienen 2 variables aleatorias independientes, x distribuida N(0, 1) e y distribuida χ_n^2 de Pearson.

Definimos la variable $t = \frac{x}{\sqrt{\frac{y}{n}}}$.

Decimos que t sigue una distribución de probabilidad t de Student con n grados de libertad t_n . La función de densidad es:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} . \frac{(1+\frac{t^2}{n})^{\frac{n+1}{2}}}{\sqrt{n\pi}} \ \forall\ t \in \mathbb{R}$$

Y su Función de Distribución está tabulada

Distribución t de Student

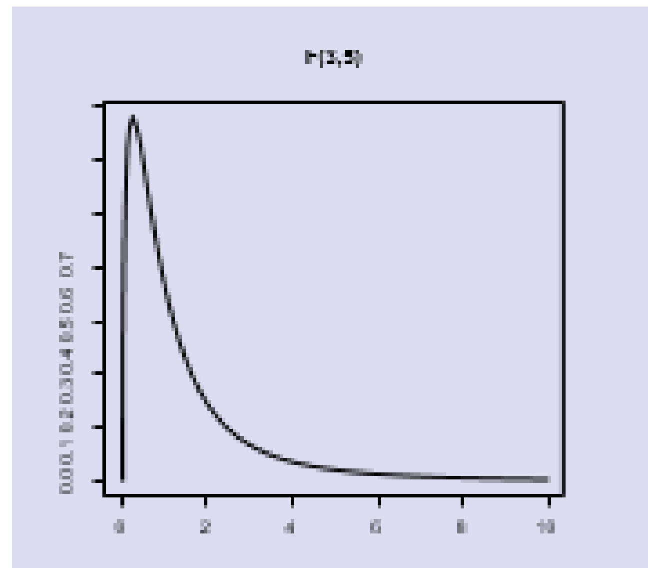
Propiedades:

- Es simétrica, está centrada en el punto (0,0)
- $M_o = M_e = 0$
- $E[T] = 0$ si $n>1$
- $V[T] = n/n-2$ si $n>2$.
- Cuando el número de variables aleatorias es muy grande, es decir, cuando $n \to \infty$, la variable se puede aproximar por una normal.

Distribución F de Snedecor

La distribución de probabilidad **F de Snedecor** con *n* y *m* grados de libertad (*F*(*n, m*)) es la distribución asociada a una variable aleatoria que se obtiene a partir del cociente de dos variables chi-cuadrado con *n* y *m* grados de libertad respectivamente.
Por tanto, esta distribución sólo tomar valores positivos.

Su función de densidad es muy compleja y su gráfica es parecida a la de la distribución chi-cuadrado. En el siguiente gráfico aparecen representada la función de densidad de una *F* de 3 *y* 6 grados de libertad



Intuitivamente, esta distribución es de utilidad para **establecer comparaciones entre las varianzas poblacionales a partir de dos conjuntos de datos extraídos de una variable normal**

Distribución F de Snedecor

Si se tienen 2 variables aleatorias independientes, x distribuida χ_m^2 e y distribuida χ_n^2 de Pearson.

Si definimos la variable $F = \frac{\frac{x}{m}}{\frac{y}{n}} = \frac{xn}{ym}$,

F sigue una distribución F de Snedecor (tb Snedecor –Fisher) con m y n grados de libertad (m grados de libertad en el numerador y n grados de libertad en el denominador) $Fm{,}n$

La función de densidad es:

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{F^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}F\right)^{\frac{m+n}{2}}} \; si\ F \geq 0 \\ 0\ si\ F < 0 \end{cases}$$

Y su Función de Distribución está tabulada

Distribución F de Snedecor

Propiedades:

- $E[F] = \dfrac{n}{m-2}$, si $m > 2$.
- $V[F] = \dfrac{m^2(2n+2m-4)}{n(m-2)^2(m-4)}$, si $m > 4$.
- Si $m \to \infty$ entonces la distribución $X \to F_{n,m} \equiv \chi^2_n$.
- Si $X \to F_{n,m}$ entonces la distribución $\dfrac{1}{X} \to F_{m,n}$.

Estadística

Tema 5: Estimación Puntual y por Intervalos

Contenido

Tema 5. Estimación puntual y por intervalos

5.1. Métodos de muestreo.

5.2. Estimación puntual.

5.3. Estimación por intervalos.

Métodos de Muestreo

- Con la **Estadística Descriptiva**, se estudiaron una serie de procedimientos y técnicas, que permitían un conocimiento descriptivo de las características básicas de una población.
- En general, no podremos casi nunca tratar con poblaciones al completo, por ejemplo no podemos estudiar todos los coches que salen de una cadena de producción para determinar su calidad, ni es posible ensayar un medicamento en todas las personas, ni podemos costearnos preguntar a todos los españoles sobre una cuestión cualquiera (salvo en un referéndum, votaciones, o en el censo, siendo estos los pocos casos en que un estudio comprende a toda la población).
- Para una mayor rapidez en la recogida y presentación de los datos, lo que se suele hacer es obtener los datos de tan sólo una parte de la población que denominamos **Muestra** .
- Por tanto utilizaremos muestras, que sean capaces de revelarnos algo acerca de la población de las que han sido extraídas.

Métodos de Muestreo

- La **Estadística Inferencial** se ocupa de extender o extrapolar a toda una población, informaciones obtenidas de una muestra, así como de la toma de decisiones.

- Al trabajar con muestras, hay que diferenciar los valores observados en la muestra, que llamaremos **estadísticos muestrales**, de los valores reales correspondientes a la población, que llamaremos **parámetros poblacionales.**

- En la primera parte de este tema trataremos de la forma de elegir las muestras y las condiciones que han de verificar.

Métodos de Muestreo

- **Problema**: extraer conclusiones válidas respecto de un grupo grande *(***población***)*
  - *En la práctica tarea difícil o imposible controlar todos los elementos de la población*
- **Solución***: Examinar a una parte pequeña de la población (***muestra***) y deducir ciertos hechos para toda la población*

  → ***Inferencia estadística***

Ejemplos:

- Extraer conclusiones respecto a la estatura de 12.000 estudiantes examinando solamente 100 estudiantes seleccionados
- Extraer conclusiones respecto al porcentaje de tornillos defectuosos producidos en una fábrica durante una semana examinando 20 tornillos diariamente producidos en momentos diferentes del día

Métodos de Muestreo

- La población ideal que se pretende estudiar se denomina **población objetivo.**
  - Idealmente deberíamos realizar una aproximación mediante muestras que den a cada individuo la **misma probabilidad de ser elegido.**
  - Tampoco es fácil elegir muestras de la población objetivo:
    - Si llamamos por teléfono excluimos a los que no tienen.
    - Si elegimos personas en la calle, olvidamos los que están trabajando, etc.
- El grupo que en realidad se puede estudiar se denomina **población de estudio**. (ej. Los usuarios de Internet por Fibra Óptica)

Métodos de Muestreo

Nuestro objetivo va a ser a partir de ahora, el tratamiento estadístico de muestras.

¿Pero bajo que condiciones, resulta apropiada una muestra?

Para que los resultados obtenidos a partir de una muestra sean fiables, esta tiene que cumplir dos condiciones fundamentales:

- Tener un tamaño adecuado.
- Que sus elementos hayan sido seleccionados de manera aleatoria.

- Si cumple estas dos condiciones diremos que la muestra es **representativa**.
- En el caso en que la selección **no** sea aleatoria se dirá que la muestra es **sesgada**.

Métodos de Muestreo

- **Muestreos probabilistas**
  - Conocemos la probabilidad de que un individuo sea elegido para la muestra.
  - Se utiliza algún sistema de selección **aleatoria** para garantizar que cada individuo de la población tenga una probabilidad específica y conocida de ser seleccionado.
- **Muestreos no probabilistas**
  - No se conoce la probabilidad.
  - Son muestreos que seguramente esconden sesgos.
  - En principio no se pueden extrapolar los resultados a la población.

En adelante vamos a tratar ***exclusivamente*** con muestreos representativos o con menor posibilidad de sesgo es decir **muestreos probabilistas.**

Métodos de Muestreo

Fuentes de sesgo

- Las poblaciones objetivo y de estudio **pueden diferir** en cuanto a las variables que estudiamos.
  - El nivel económico en la población de estudio es mayor que en la población objetivo, los individuos que se eligen en la calle pueden ser de mayor edad (ej. mayor frecuencia de jubilados)
- En este caso, diremos que las muestras que se elijan estarán **sesgadas**.
  - Al tipo de sesgo debido a diferencias sistemáticas entre población objetivo y población de estudio se denomina **sesgo de selección**.
  - Hay otras fuentes de error/sesgo (no responder a encuestas embarazosas o mentir en las preguntas “delicadas”: consumo de drogas, violencia doméstica, prácticas poco éticas, etc.) Para evitar este tipo de sesgo se utilizan diversas técnicas, por ejemplo, la técnica de respuesta **aleatorizada.**

Métodos de Muestreo

- **Muestreo aleatorio simple**: Cada elemento de la poblacion tiene la misma probabilidad de ser elegido para la muestra.
- **Muestreo aleatorio sistemático**: Se eligen los elementos de la población con una periodicidad determinada
- **Muestreo aleatorio estratificado**: Los elementos de la poblacion se dividen en estratos o clases
  - Se utiliza cuando la población presenta diferentes característica (ejp. sondeo de opinión)
  - Se toman elementos de cada estrato proporcionalmente al tamaño del estrato dentro de la población
- **Muestreo aleatorio por conglomerados**: Los elementos de la población se encuentran en conglomerados que se suponen representativos de la población
  - Ejemplo agrupación de la población por ciudades, provincias, etc.

Métodos de Muestreo

En todos los casos se puede dar:

- **Muestreo con remplazamiento** es aquel en el que cada miembro de una población puede ser seleccionado más de una vez, es decir todos los elementos de la población **tienen la misma probabilidad siempre**.
- **Muestreo sin remplazamiento** es aquel en el que cada miembro sólo puede ser seleccionado una sola vez, es decir una vez elegido un elemento de la población para incluirlo en la muestra, se excluye de la población para que no pueda volver a ser elegido.

- Las técnicas de Estadística Inferencial son adecuadas para el muestreo con remplazamiento
- En poblaciones suficientemente grandes se utilizan de manera similar tanto para el muestreo con remplazamiento como para el muestreo sin remplazamiento, ya que **en la practica se utilizan generalmente muestras sin remplazamiento** en casos concretos pudiera ser necesario aplicar algún ajuste o factor de corrección.
- En poblaciones muy pequeñas, pueden existir variaciones en la estimación de la probabilidad de los diferentes elementos.

Muestreo aleatorio simple (m.a.s.)

- Partimos de una lista de los individuos de la población de estudio.
- Se eligen individuos de la población de estudio, de manera que todos **tienen la misma probabilidad de ser elegidos**, hasta alcanzar el tamaño muestral deseado.
- La selección de los elementos de la muestra se realiza en una sola etapa (en la practica se realiza sin remplazamiento) .
  - Ejemplo: se parte de un listado de individuos de la población, y se eligen individuos aleatoriamente mediante tablas de números aleatorios o un procedimiento similar que garantice la aleatoriedad.
- Normalmente tiene un coste bastante alto su aplicación.
- En general, las técnicas de inferencia estadística suponen que la muestra ha sido elegida usando m.a.s., aunque en realidad se use alguna de las técnicas que veremos a continuación.

Muestreo aleatorio sistemático

- Es análogo al anterior, ya que también se tiene una lista de los individuos de la población de estudio, aunque resulta más cómoda la elección de los elementos.
- Si queremos una muestra de un tamaño dado, elegimos individuos igualmente espaciados de la lista, donde el primero ha sido elegido al azar.
- Para obtener una muestra de tamaño n, en una población de tamaño N, se ordenan y numeran los elementos de la población. El primer elemento de la muestra, llamado **origen**, se obtiene al azar. Posteriormente, hallamos el entero k más próximo a N/n. Los demás valores se obtienen sumando al primer elemento el número k, teniendo en cuenta que, al sobrepasar N, debemos empezar de nuevo.

- **¡CUIDADO!**: Si en la lista existen periodicidades, obtendremos una muestra sesgada.

  - **Un caso real:** Se eligió una de cada 5 casas para un estudio de salud pública en una ciudad donde las casas se distribuyen en manzanas de 5 casas. Salieron con mucha frecuencia las de las esquinas, que reciben más sol, están mejor ventiladas, etc.

Muestreo aleatorio estratificado

- Se aplica cuando sabemos que hay ciertos factores (variables, subpoblaciones o estratos) que pueden influir en el estudio y queremos asegurarnos de tener **cierta cantidad mínima** de individuos de cada tipo:
  - Hombres y mujeres,
  - Jóvenes, adultos y ancianos…
- Se realiza entonces una m.a.s. de los individuos de cada uno de los estratos.
- Al extrapolar los resultados a la población hay que tener en cuenta el tamaño relativo del estrato con respecto al total de la población.

Muestreo aleatorio por conglomerados

- Se aplica cuando no se dispone o es difícil tener una lista de todos los individuos que forman parte de la población de estudio, pero sabemos que se encuentran **agrupados naturalmente en grupos**.
- Se realiza eligiendo varios de esos grupos al azar, y ya elegidos algunos podemos estudiar a todos los individuos de los grupos elegidos o bien seguir aplicando dentro de ellos más muestreos por grupos, por estratos, aleatorios simples,…
  - Ejemplo: para conocer la opinión de los médicos del sistema nacional de salud, podemos elegir a varias regiones de España, dentro de ellas varias comarcas, y dentro de ellas varios centros de salud, y…
- Al igual que en el muestreo estratificado, al extrapolar los resultados a la población hay que tener en cuenta **el tamaño relativo de unos grupos con respecto a otros.**
  - Regiones con diferente población pueden tener probabilidades diferentes de ser elegidas, comarcas, hospitales grandes frente a pequeños,…

Parámetros Poblacionales

- Se considera que se conoce una población cuando conocemos la distribución de probabilidad *f*(x) de la variable aleatoria X

- **Parámetros poblaciones:** son los parámetros que describen la distribución de probabilidad *f*(x), por ejemplo μ y σ, en la distribución normal, *p* en la binomial; otras como la mediana, los momentos, etc.

- En ocasiones *f*(x) no se conoce, aunque podemos formular alguna hipótesis relativa a su comportamiento. En tal caso se deben **estimar los parámetros poblacionales.**

- Cualquier cantidad obtenida de una muestra con el propósito de obtener un parámetro poblacional se llama **estadístico muestral**

Estimación puntual

- Una **estimación es puntual** cuando se usa un solo valor extraído de la muestra para estimar el parámetro desconocido de la población. Al valor usado se le llama estimador o estadístico muestral.
- La media de la población se puede estimar puntualmente mediante la media de la muestra, la desviación típica de la población se puede estimar puntualmente mediante la desviación típica de la muestra, aunque hay mejores estimadores, como la cuasidesviacion típica, la proporción de la población se puede estimar puntualmente mediante la proporción de la muestra.
- Sea una población donde se observa la variable aleatoria X con los siguientes parámetros $\mathrm{E}[X]=\mu;\qquad \mathrm{Var}[X]=\sigma^2$
- Consideramos una m.a.s. de tamaño n formada por las variables $X_1, X_2, \ldots, X_n$ independientes entre si. Definimos los estimadores o estadísticos muestrales:

✓ Media muestral:

$$\overline{X}=\frac{X_1+X_2+\ldots+X_n}{n}$$

✓ Varianza muestral:

$$S^2=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}$$

✓ Cuasi-Varianza muestral:

$$\widehat{S^2}=Sn_{-1}{}^2=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}$$

Estimación puntual

Propiedades de los estimadores

Insesgadez : Un estimador es **insesgado** cuando la esperanza matemática del este es igual al parámetro que se desea estimar. Por tanto, la diferencia entre el parámetro a estimar y la esperanza de nuestro estimador tendría que ser 0.

Eficiencia: Un estimador es más **eficiente** o tiene la capacidad de estimar de forma precisa cuando su varianza es reducida. Por lo tanto ante 2 estimadores, siempre elegiremos el que tenga una varianza menor.

Consistencia: Un estimador **consistente** es aquel que a medida que la muestra crece se aproxima cada vez más al valor real del parámetro. Por lo tanto, cuantos más valores entran en la muestra, el parámetro estimado será más preciso

Estimación puntual

Los estadísticos muestrales, media, varianza y cuasivarianza verifican las siguientes propiedades:

Para la Media Muestral

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

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Para la Varianza Muestral

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

Para la Cuasivarianza muestral

$$\mathrm{E}(S_{n\,\_\,1}{}^2) = \sigma^2$$

- ***Estas propiedades se verifican siempre, cualquiera que sea la distribución de la variable X***
- Siempre que el **valor esperado de un estadístico** sea igual al parámetro poblacional correspondiente el estadístico se denomina, **estimador insesgado**.
- La **media muestral** y la **cuasivarianza muestral** son estimadores insesgados de la media y varianza poblacional respectivamente

Distribución muestral

- Las **muestras aleatorias** obtenidas de una población son **impredecibles**
- Dos muestras aleatorias del mismo tamaño y tomadas de la misma población no tendrán la misma media o no tienen por qué ser muy parecidas
- Cabe esperar que cualquier **estadístico**, como la media muestral, calculado a partir de las medias en una muestra aleatoria, **cambie su valor de una muestra a otra → *estudiar la distribución de todos los valores posibles de un estadístico***.
- Las **distribuciones de los estadísticos** son de gran importancia en el estudio de la **estadística inferencial**, porque las inferencias sobre las poblaciones se harán usando **estadísticas muestrales**.
- **Analizando las distribuciones de los estadísticos muestrales** puede conocerse la **fiabilidad o confiabilidad de un estadístico muestral** como instrumento para **realizar inferencias sobre un parámetro poblacional desconocido.**

Distribución muestral

- Como los valores de un estadístico varían de una muestra aleatoria a otra, un estadístico se puede considerar como una **variable aleatoria** con su correspondiente **distribución de frecuencias**.

- La distribución de frecuencia de un estadístico muestral se denomina ***distribución muestral.***

- En general, podemos considerar todas las muestras posibles de tamaño ***n*** que pueden extraerse de la población y para cada muestra computar el estadístico que queramos estudiar . De esta forma obtenemos la **distribución del estadístico**.

- Para una distribución muestral podemos computar la media, proporción, varianza, desviación típica, etc.

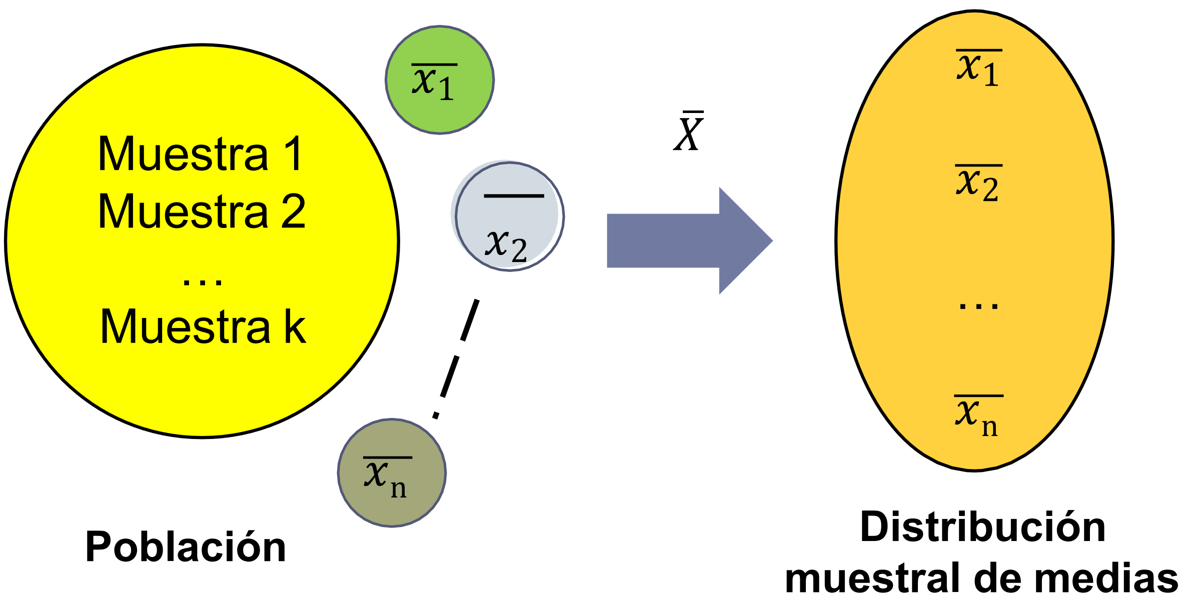
Distribución muestral de medias

Supongamos que se quiere estudiar la **media μ de una población**.

Para ello consideramos todas las muestras de tamaño *n* de la población objeto de estudio, M_1, M_2, …, M_n y calculamos sus medias $\overline{x}_1$, $\overline{x}_2$,…, $\overline{x}_n$

Sea $\bar{X}$ la variable aleatoria que asigna a cada muestra su media: $M_i \xrightarrow{\bar{X}} \overline{x}_i$

Esta variable se denomina **media muestral** y la distribución que sigue se llama **distribución muestral de las medias.**



Para la Media Muestral

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

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

Distribución muestral de medias

Según el **teorema central del limite (TCL),** podemos decir que si una población tiene **media μ y desviación típica σ conocida** , y tomamos muestras de tamaño n (n ≥ 30, o cualquier tamaño si la población de partida es "normal"), **las medias de estas muestras** siguen aproximadamente la distribución:

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

Por lo tanto

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0;1)}$$

Distribución muestral de medias

- Por tanto, la distribución de las medias muestrales, es una distribución de tipo “normal”, siempre que la población de procedencia lo sea, o incluso si no lo es, siempre que el **tamaño de las muestras sea 30 o mayor**.
- La desviación típica de la distribución de las medias **es el grado de variabilidad** de las medias muestrales. Cuanto menor sea, más ajustadas a la media de la población serán las medias que obtengamos de una muestra.
- De su propia definición, es fácil darse cuenta de que **cuanto mayor es el tamaño de la muestra, menor es este grado de variabilidad**, y por tanto más similar a la media de la población será la media obtenida de la muestra.
- Es decir, cuanto mayor es el valor de *n*, mejor es la aproximación “normal”.
- En términos más coloquiales, lo que en definitiva establece el TCL, es que la distribución de la media (o de las sumas), de diferentes valores da como resultado una distribución normal.

Distribución Muestral de Medias

Factor de Corrección para Población Finita

El teorema central del límite y en general las técnicas de inferencia estadística están basadas en que las muestras son seleccionadas con reemplazamiento, lo que significa que trabajamos con poblaciones infinitas, pero en la practica, en todos los casos reales, el muestreo se hace **sin reemplazamiento** por lo que trabajamos con poblaciones con tamaño finito N.

Para estos casos puede ser necesario emplear un **Factor de Corrección** para la desviación estándar, denominado **coeficiente corrector** de poblaciones finitas" o "**coeficiente de exhaustividad**“ cuya expresión es:

$$\sqrt{\frac{N-n}{N-1}}$$

Trabajaremos generalmente suponiendo que la población es infinita y aplicaremos el factor de corrección si se indica.

Distribución Muestral de Medias

Esta corrección modifica la ecuación del estimador o estadístico de la siguiente forma.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Siendo n el tamaño de la muestra y N el de la población.

Por tanto **las medias de estas muestras**, siguen aproximadamente la distribución:

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

Distribución Muestral de Medias

- **Ejemplo:** Sea X una población con distribución N (μ = 90, σ = 20).

A) Si se obtiene una m.a.s. de tamaño 16,¿cuál es la probabilidad de que la media muestral X sea mayor o igual que 92?
B) Determinar el tamaño muestral para que la probabilidad de que la media muestral sea menor o igual que 98 sea 0,99.

$$\left.\begin{array}{c} \mathbb{E}\left(\overline{X}\right) = \mu = 90 \\ Var\left(\overline{X}\right) = \dfrac{\sigma^2}{n} = \dfrac{20^2}{16} \Rightarrow \sigma_{\overline{X}} = \sqrt{\dfrac{400}{16}} = \dfrac{20}{4} = 5 \end{array}\right\} \Rightarrow \overline{X} \text{ - } N(90, 5)$$

$$a)\ \mathbb{P}\left(\overline{X} \geq 92\right) = \mathbb{P}\left(N(90,5) \geq 92\right) = \mathbb{P}\left(N(0,1) \geq \frac{92-90}{5}\right) =$$

$$= \mathbb{P}\left(Z \geq \tfrac{2}{5}\right) = \mathbb{P}\left(Z \geq 0.4\right) = \boxed{0,3446}$$

$$b)\ 0.99 = \mathbb{P}\left(\overline{X} \leq 98\right) = \mathbb{P}\left(N\left(90, \frac{20}{\sqrt{n}}\right) \leq 98\right) = \mathbb{P}\left(N(0,1) \leq \frac{98-90}{\frac{20}{\sqrt{n}}}\right) =$$

$$= \mathbb{P}\left(Z \leq \sqrt{n}\tfrac{2}{5}\right) \Longrightarrow \sqrt{n}\frac{2}{5} = 2,33 \Longrightarrow n = \left(\frac{2,33 \cdot 5}{2}\right)^2 = 33,9 \Longrightarrow \boxed{n \geq 34}$$

Distribución Muestral de Medias

Funcion de Distribucion Normal Estandar F.D. $P[Z<z]$; $z \rightarrow N(0,1)$										
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,10	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,20	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,30	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,40	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,50	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,60	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,70	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,80	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,90	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,00	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,10	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,20	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,30	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,40	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,50	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,60	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,70	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,80	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,90	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,00	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,10	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,20	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,30	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,40	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,50	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,60	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,70	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,80	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,90	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,00	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900

Tabla:

Función de distribución normal estándar o tipificada **N(0,1)**

a) $P[Z < 0{,}4]=0{,}65542$

$P[Z \geq 0{,}4]=1-[Z<0{,}4]=$
$1-0{,}65542=0{,}34458=$

$\rightarrow$ 34,458%

b) $0{,}99 \rightarrow P[Z < z_x]=2{,}33$

Distribución Muestral de Medias

- **Ejemplo:**

Las estaturas de 1.000 estudiantes están distribuidas aproximadamente en forma normal con una media de 174,5 centímetros y una desviación estándar de 6,9. centímetros.
Si se extraen 200 muestras aleatorias de tamaño 25 **sin reemplazo** de esta población, determinar:
a) El número de las medias muestrales que caen entre 172,5 y 175,8 centímetros.
b) El número de medias muestrales que caen por debajo de 172 centímetros.

N = 1.000 y n= 25

$$\mu = 174{,}5$$
$$\sigma = 6{,}9$$

Distribución Muestral de Medias

Solución:

En este ejercicio se realiza un **muestreo sin reemplazo,** por lo que podemos considerar que hablamos de población finita, por lo que aplicaremos el **factor de corrección**.

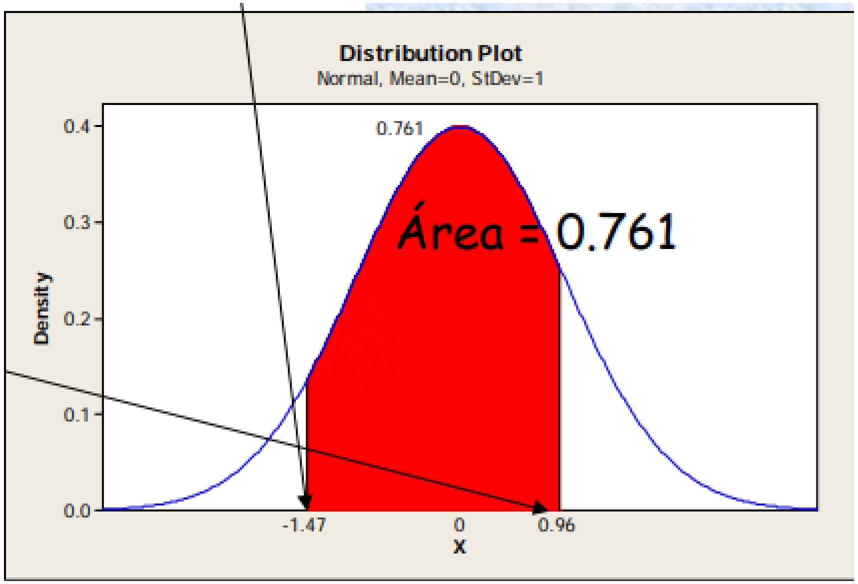
Calculamos el denominador de Z

$$\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}} = 6.9/\sqrt{25}\sqrt{\frac{1000-25}{1000-1}} = 1.36$$

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}} = \frac{172.5-174.5}{1.36} = -1.47$$

Distribución Muestral de Medias

-1.47



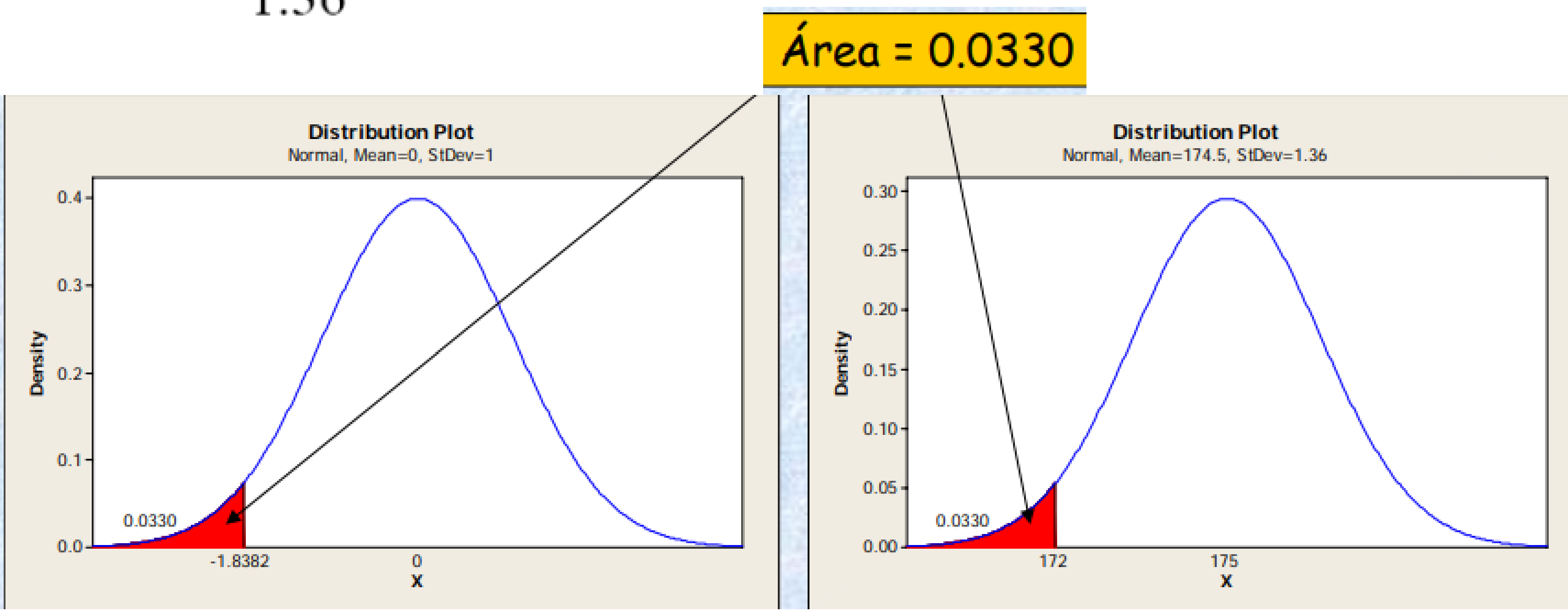
$$z = \frac{175.8 - 174.5}{1.36} = 0.96$$

Luego el área comprendida entre 0,96 y -1,47 es 0,761
Por tanto sobre 200 muestras tendremos

(0.761)(200)=152 medias muestrales caerán dentro del rango pedido

Distribución Muestral de Medias

b) $$z = \frac{172 - 174.5}{1.36} = -1.83$$



Nota: Es lo mismo si se toman los valores originales con la distribución para la media y desviación estándar correspondientes, pero esto sólo se puede efectuar con software y no con tablas.

Por lo tanto la respuesta es: (0.0330)(200)= 7 medias muestrales

Distribución muestral de la media

Cuando la Varianza poblacional, σ2 , es desconocida

Sabemos que la media se distribuye como una normal, si tipificamos

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

En el caso de desconocer σ, ésta se sustituye por la varianza muestral S o S_n o por la cuasivarianza muestral $\widehat{S}$ o S_{n-1} dando lugar al estadístico que denominaremos T.

$$T = \frac{\overline{X} - \mu}{\hat{S} / \sqrt{n}} = \frac{\overline{X} - \mu}{S / \sqrt{n-1}}$$

- Este estadístico tiene una distribución denominada **t-student** con ***n-1*** **grados de libertad** y se utiliza cuando el **tamaño de la muestra es pequeño,** generalmente por **debajo de *n*=30**
- Para **Muestras grandes, n > 30:** T → N (0,1)

Distribución muestral de la media

Ejemplo:

Se está realizando un estudio sobre la calidad del aire en una zona. Uno de los indicadores de la calidad del aire es el número medio de microgramos de partículas en suspensión por metro cúbico.

Supongamos que la variable X: ”Número de microgramos de partículas”, está normalmente distribuida. Se hacen **16 mediciones**, en las que se obtiene una cuasidesviación típica de 10.8585 unidades. Obtener la probabilidad de que la media muestral no difiera de la media poblacional en más de 8 unidades.

Distribución muestral de la media

$$T = \frac{\overline{X} - \mu}{\hat{S} / \sqrt{n}} = \frac{\overline{X} - \mu}{\frac{10{,}8585}{\sqrt{16}}} \rightarrow t_{15}$$

$$P\left(\left| \overline{X} - \mu \right| \leq 8 \right) = P\left(-8 \leq \overline{X} - \mu \leq 8 \right) =$$

$$= P\left(\frac{-8}{\frac{10.8585}{\sqrt{16}}} \leq \overline{X} - \mu \leq \frac{8}{\frac{10.8585}{\sqrt{16}}} \right) = P(-2.947 \leq t_{15} \leq 2.947) =$$

$$= 1 - 2P\left(t_{15} \geq 2.947\right) = 1 - 2 \times 0.005 = 1 - 0.01 = 0.99$$

Distribución muestral de la media

Ejemplo: Se hacen **36 mediciones** en las que se obtiene una cuasidesviación típica de 12 unidades. Obtener la probabilidad de que la media muestral no difiera de la media poblacional en más de 5 unidades.

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{12/\sqrt{36}} = \frac{\overline{X} - \mu}{2} \to t_{35} \cong N(0;1)$$

$$P\left(\left|\overline{X} - \mu\right| \leq 5\right) = P\left(-5 \leq \overline{X} - \mu \leq 5\right) =$$

$$= P\left(\frac{-5}{2} \leq \frac{\overline{X} - 5}{2} \leq \frac{5}{2}\right) = P(-2.5 \leq Z \leq 2.5) =$$

$$= 1 - 2 \times 0.00621 = 0.98758$$

Distribución muestral de la media

En Resumen:

Si X → N(μ, σ) el estimador puntual de la media μ es la media muestral $\bar{X}$

a. La media muestral $\bar{X}$ → N(μ, σ /$\sqrt{n}$) si σ es conocido

b. La media muestral $\bar{X}$ → N(μ, S_{n-1} /$\sqrt{n}$) si σ es desconocido y n> 30

siendo S_{n-1} la cuasivarianza muestral (a veces $\hat{S}$)

o bien $\bar{X}$ → N(μ, S_n / $\sqrt{n-1}$) siendo S_n la varianza muestral.

c. Si σ es desconocido y n≤ 30 $\dfrac{\bar{X} - \mu}{S_{n-1} / \sqrt{n}}$ → t_{n-1}

o bien $\dfrac{\bar{X} - \mu}{S_n / \sqrt{n-1}}$ → t_{n-1}

Distribución muestral de proporciones

Se tiene una **población infinita,** distribuida como una **binomial,** se consideran las posibles muestras de tamaño *n* extraídas de la población y para cada muestra se determina el **estadístico proporción de éxitos** $\widehat{\boldsymbol{P}}$

La distribución de probabilidad de esta variable $\widehat{\boldsymbol{P}}$, se llama distribución muestral de las proporciones.

Si $\mu_{\bar{P}}$ y $\sigma_{\bar{P}}$ son la media y la desviación típica de $\hat{P}$, respectivamente, se verifica que:

- $\mu_{\bar{P}} = p$
- $\sigma_{\bar{P}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{p \cdot (1-p)}{n}}$
- A medida que n crece ($n \geq 30$), la distribución de $\hat{P}$ se aproxima a una normal (siempre que p no se acerque a 0 o a 1):

$$\hat{P} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Distribución muestral de proporciones

Ejemplo: Imaginemos que sabemos que la proporción del alumnado de nuestro centro que es favorable a realizar una huelga es del 60 %. Cuando elegimos a un alumno, y nos preguntamos si es favorable a la huelga, es como si realizáramos una prueba binomial con probabilidad de éxito $p = 0{,}6$.

Cuando elegimos muestras aleatorias de 70 alumnos, el número de ellos favorable a la huelga, deberá seguir una distribución $B\ (70;\ 0{,}6)$, o bien, la proporción de ellos que es favorable se debe distribuir según:

$$\hat{P} \sim N\left(0{,}6; \sqrt{\frac{0{,}6 \cdot 0{,}4}{70}}\right) = N(0{,}6;\ 0{,}058)$$

Es decir, las proporciones que vayamos encontrando para muestras de tamaño 70, se irán distribuyendo de forma "normal" alrededor del 60 %, con una desviación típica del 5,8 %.

Distribución muestral de proporciones

- **Ejemplo:**
  - Si tiramos una moneda no trucada 100 veces, ¿cuál es la probabilidad de que obtengamos más de 55 caras?

En una moneda no trucada la proporción de caras es 0,5, con lo que ***p*=0,5;** ***q*=0,5** y ***n*=100**

- La distribución muestral de proporciones se distribuye

$$\mu_p = p = 0{,}5 \; ; \qquad \sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}5(1-0{,}5)}{100}} = \frac{0{,}5}{10} = 0{,}05$$

es decir la proporción muestral se distribuye según una **N(0,5;0,05)**

Si llamamos p' a la proporción en la muestra tenemos que calcular la probabilidad **P(p'>0,55)** *(55 caras/100 lanzamientos)*

$$Z = \frac{p' - \mu_p}{\sigma_p} = \frac{0{,}55 - 0{,}5}{0{,}05} = 1$$

P(p'>0,55) = P(z>1) =1-P(z≤1) = 1-0,8413 =**0,1587 → 15,87%**

Distribución muestral de proporciones

Funcion de Distribucion Normal Estandar F.D. P[Z<z]; z→ N(0,1)										
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,10	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,20	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,30	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,40	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,50	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,60	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,70	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,80	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,90	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,00	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,10	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,20	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,30	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,40	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,50	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,60	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,70	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,80	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,90	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,00	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,10	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,20	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,30	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,40	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,50	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,60	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,70	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,80	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,90	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,00	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900

Tabla:

Función de distribución normal estándar o tipificada **N(0,1)**

DISTRIBUCIÓN MUESTRAL DE LA DIFERENCIA DE MEDIAS

Diferencia de medias muestrales con varianza conocida y muestras independientes

Si $\overline{X}_1$ y $\overline{X}_2$ son independientes y $\mu_{\overline{X}_1-\overline{X}_2}$ y $\sigma_{\overline{X}_1-\overline{X}_2}$ son la media y la desviación típica, , de $\overline{X}_1-\overline{X}_2$ Se cumple

- $\mu_{\overline{X}_1-\overline{X}_2}=\mu_1-\mu_2$
- $\sigma_{\overline{X}_1-\overline{X}_2}=\sqrt{\dfrac{\sigma_1^{\,2}}{n_1}+\dfrac{\sigma_2^{\,2}}{n_2}}$, siendo σ_1 y σ_2 las desviaciones típicas poblacionales.
- Si las poblaciones tienen una distribución normal, $\overline{X}_1-\overline{X}_2$ es normal.
- Si las poblaciones no tienen una distribución normal, a medida que n_1 y n_2 crecen ($n_1 \geq 30$ y $n_2 \geq 30$), la distribución de $\overline{X}_1-\overline{X}_2$ se aproxima a una normal.

$$\overline{X}_1-\overline{X}_2 \sim N\left(\mu_1-\mu_2, \sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}\right)$$

DISTRIBUCIÓN MUESTRAL DE LA DIFERENCIA DE MEDIAS

Ejemplo: La duración media, en años, de los frigoríficos de la marca *A* es 18, y la de los de la marca *B*, 16. Las desviaciones típicas son 3 y 5 años respectivamente. Se toman 75 frigoríficos de la marca *A* y 50 de la marca *B*, y se observa su duración media. ¿Cuál es la probabilidad de que la duración media de la muestra *A* supere en más de un año a la duración de la muestra *B*?

Consideramos la variable aleatoria $\overline{X_1} - \overline{X_2}$ que asigna a cada par formado por una muestra de *A* y una de *B* la diferencia de sus duraciones medias.

Los datos del enunciado son: $\mu_1 = 18$, $\mu_2 = 16$, $\sigma_1 = 3$ y $\sigma_2 = 5$, $n_1 = 75$ y $n_2 = 50$.

Como se cumple que $n_1 = 75 \geq 30$ y $n_2 = 50 \geq 30$, se tiene que:

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}}\right) = N\left(18-16, \sqrt{\frac{3^2}{75} + \frac{5^2}{50}}\right) = N\,(2;\, 0{,}787)$$

Así:

$$P\,(\overline{X_1} - \overline{X_2} > 1) = P\left(Z > \frac{1-2}{0{,}787}\right) = P\,(Z > -1{,}27) = P\,(Z < 1{,}27) = 0{,}8980$$

Distribución Muestral de la Varianza

Si X_1, X_2, …,X_n son las variables estadísticas para una muestra de tamaño n, sabemos que la varianza muestral es :

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}$$

Como ya hemos visto, el valor esperado de la varianza muestral no es la varianza de la población:

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

El **estimador insesgado** de la **varianza muestral** se denomina **cuasivarianza muestral :**

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

$$E(\hat{S}^2) = \sigma^2$$

Distribución Muestral de la Varianza

- Si tomamos muestras aleatorias de tamaño ***n*** de una población que tiene una **distribución normal,** entonces se verifica que:

$$\frac{n\,S_n^{\;2}}{\sigma^2} = \frac{(n-1)S^2_{\;n-1}}{\sigma^2} \rightarrow \quad \chi^2_{\;(n\text{-}1)}$$

- Es decir los estimadores puntuales de la varianza, varianza muestral y cuasivarianza muestral, cumplen que la expresión anterior sigue una distribución **chi-cuadrado** con **n - 1 grados de libertad.**

Ejemplo: Encontrar la probabilidad de que una muestra aleatoria de 25 observaciones, de una población normal con varianza igual a 6, tenga una varianza muestral mayor que 9,45.

Distribución Muestral de la Varianza

- Según la expresión anterior:

$$\frac{n\, Sn^2}{\sigma^2} \text{ se distribuye como una } \chi^2_{(n-1)}$$

n -1 grados de libertad, → 24 grados de libertad

Sustituyendo tenemos (25*9,45)/6 = 39,4

Por tanto, buscamos en la tabla el número obtenido en la fila correspondiente a 24 grados de libertad y nos da un área a la izquierda de 0,975.

Por lo que la

$$P(S_n^2 > 9{,}45) = 1 - P(S_n^2 < 9{,}45) = 1 - P(\chi^2_{(n-1)} < 39{,}4) = 1 - 0{,}975 = 0{,}025$$

es decir es del 2,5%

Distribución Muestral de Varianzas

- **Ejercicio:**

Una población se compone de los cinco números 2, 3, 6, 8, 11. Considerar todas las muestras posibles de tamaño dos que puedan extraerse con remplazamiento de esta población. Hallar:

a. la media y la desviación típica de la población
b. la media y la desviación típica de la distribución muestral de medias
c. la media y la desviación típica de la distribución muestral de varianzas
d. determinar la probabilidad de muestras para las cuales las varianzas muestrales son mayores que 7,2. Verificar este resultado con el resultado real.

Distribución Muestral de Varianzas

a) $$\bar{\mu} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$$

$$\sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5} = \frac{16+9+0+4+25}{5} = \frac{54}{5} = 10{,}8$$

$$\sigma = \sqrt{10{,}8} = 3{,}2863$$

b) Hay 25 muestras de tamaño 2 que pueden extraerse con remplazamiento de los cinco números (tabla A)

(2,2)	(2,3)	(2,6)	(2,8)	(2,11)
(3,2)	(3,3)	(3, 6)	(3,8)	(3,11)
(6,2)	(6,3)	(6, 6)	(6,8)	(6,11)
(8,2)	(8,3)	(8, 6)	(8,8)	(8,11)
(11,2)	(11,3)	(11,6)	(11,8)	(11,11)

Distribución Muestral de Varianzas

Las correspondientes medias muestrales son (tabla B):

2	2,5	4	5	6,5
2,5	3	4,5	5,5	7
4	4,5	6	7	8,5
5	5,5	7	8	9,5
6,5	7	8,5	9,5	11

y la media de la distribución muestral de medias es

$$\overline{\mu_x} = \frac{150}{25} = 6 = \mu$$

la varianza de la distribución muestral de medias se obtiene restando el valor de la media 6 de cada número de la tabla A, elevando al cuadrado cada diferencia, sumando los 25 números así obtenidos y dividiendo por 25:

$$S_{\bar{X}}{}^{2} = \frac{135}{25} = 5{,}4 \rightarrow S_{\bar{X}} = 2{,}3238$$

Lo que muestra que para poblaciones finitas en las que se efectúa muestreo con remplazamiento (o poblaciones infinitas)

$$S_{\bar{X}}{}^{2} = \frac{\sigma^2}{n} = \frac{10{,}8}{2} = 5{,}4$$

Distribución Muestral de Varianzas

c) Las varianzas muestrales correspondientes a cada una de las 25 muestras:

0,00	0,25	4,00	9,00	20,25
0,25	0,00	2,25	6,25	16,00
4,00	2,25	0,00	1,00	6,25
9,00	6,25	1,00	0,00	2,25
20,25	16,00	6,25	2,25	0,00

La media de la distribución muestral de varianza es

$$\mu_{S^2} = \frac{suma\ de\ todas\ de\ las\ varianzas\ de\ la\ \ tabla\ C}{25} = \frac{135}{25} = 5{,}4$$

Lo que muestra que: $$E(S^2) = \mu_{S^2} = \frac{n-1}{n}\sigma^2 = \frac{2-1}{2}\cdot 10{,}8 = 5{,}4$$

Este resultado muestra la conveniencia de definir una varianza corregida para las muestras: estimador insesgado $$\hat{S}^2 = \frac{n}{n-1}S^2; \quad E(\hat{S}) = \sigma^2$$

Distribución Muestral de Varianzas

- La varianza de la distribución muestral de varianzas, se obtiene restando la media 5,4 a cada uno de los 25 números de la tabla anterior, elevando al cuadrado estas diferencias (tabla D) , sumándolas y dividiendo el resultado por 25:

29,16	26,52	1,96	12,96	220,52
26,52	29,16	9,92	0,72	112,36
1,96	9,92	29,16	19,36	0,72
12,96	0,72	19,36	29,16	9,92
220,52	112,36	0,72	9,92	29,16

$$\sigma_{S^2}^2 = \frac{suma\ elementos\ tabla\ D}{n} = \frac{975,5}{25} = 39.03$$

$$\sigma_{S^2} = \sqrt{\sigma_{S^2}^2} = \sqrt{39.03} = \mathbf{6,25}$$

Distribución Muestral de Varianzas

d) Tenemos n =2, y $\sigma^2 = 10{,}8$ y la expresión $\frac{n\,S^2}{\sigma^2} = \frac{2S^2}{10{,}8} = \frac{S^2}{5{,}4}$

Se distribuye como una chi-cuadrado con 1 grado de libertad por tanto

$$P\left(S^2 \geq 7{,}2\right) = \quad P\left(\chi_1^{\,2} \geq \frac{7{,}2}{5{,}4} = 1{,}3\right) \cong 0{,}25 = 25\%$$

por tanto esperaríamos que alrededor del 25% de las muestras (25/4 = 6,25 ≅6) , aproximadamente 6 de las muestras tengan varianzas mayores a 7,2.

- En la tabla C con las varianzas muestrales correspondientes a cada una de las 25 muestras si contamos las varianzas mayores que 7,2 vemos que hay 6 que superan dicho valor.

0,00	0,25	4,00	9,00	20,25
0,25	0,00	2,25	6,25	16,00
4,00	2,25	0,00	1,00	6,25
9,00	6,25	1,00	0,00	2,25
20,25	16,00	6,25	2,25	0,00

INTERVALOS DE CONFIANZA

- Las estimaciones puntuales obtenidas a partir de una muestra diferirán del parámetro poblacional y, en consecuencia, quedará un margen de incertidumbre que se expresa en términos del error estándar **(EE) del estimador**. Así, resulta natural querer disponer de una medida del parámetro poblacional que incorpore tanto la estimación puntual como su error estándar. Esta medida es el **intervalo de confianza (I.C.),** que facilita un rango de valores dentro del cual se encontrará el verdadero valor del parámetro poblacional con un cierto grado de confianza.

- la ***estimación mediante intervalos de confianza***, consiste en determinar un posible rango de valores o intervalo, en los que pueda precisarse, con una determinada probabilidad, que el valor de un parámetro de la población se encuentra dentro de unos límites.

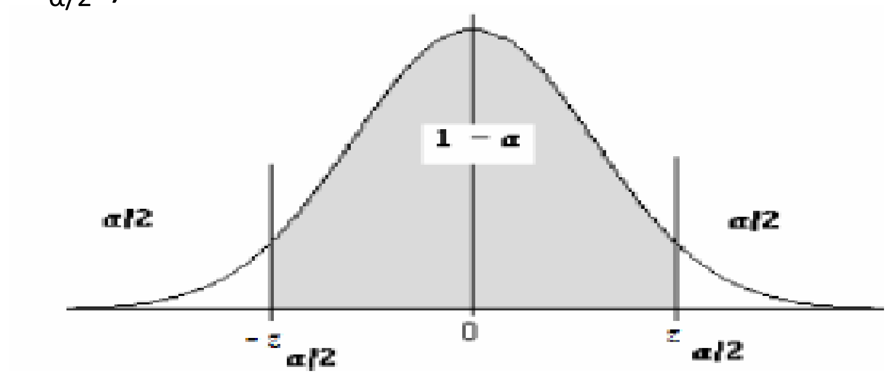
- A la probabilidad de que hayamos acertado al decir que el parámetro estaba contenido en dicho intervalo se la denomina ***nivel de confianza.***

INTERVALOS DE CONFIANZA

- Definimos ***Nivel de confianza*** como la "probabilidad" de que el intervalo calculado contenga al verdadero valor del parámetro.
- Se indica por **1-α** y habitualmente se da en porcentaje (1-α)% (Hablaremos de un nivel de confianza del 90%, del 95%, del 99%…).
- Hablamos de nivel de confianza y no de probabilidad ya que una vez extraída la muestra, el intervalo de confianza contendrá al verdadero valor del parámetro o no, lo que sabemos es que si repitiésemos el proceso con muchas muestras solo podríamos afirmar que el (1-α) % de los intervalos así construidos contendrían al verdadero valor del parámetro.
- A la probabilidad de equivocarnos se le denomina **nivel de significación**, y lo representamos por α nos informa de la probabilidad que existe de que el valor poblacional esté fuera de nuestro intervalo.
- Lógicamente, cuanto más pequeño sea α (es decir, cuanto más grande sea el nivel de confianza), la probabilidad de equivocarnos será menor, pero el intervalo que calcularemos será más grande y por tanto la precisión de la estimación será menor. Se trata pues de encontrar un equilibrio entre que, la probabilidad de equivocarnos no sea muy grande y que, el intervalo tampoco para obtener mayor precisión. Para ello se suelen prefijar niveles de confianza superiores al 90%.

INTERVALOS DE CONFIANZA

- Dado un nivel de confianza 1-α, implica que hay un α % de error, este error estará repartido en dos colas una a cada lado del área 1-α.
- llamaremos ***valor critico,*** al valor k que en una N(0,1) cumple que deja a la derecha, o a la izquierda, un área $\alpha/2$.
- Normalmente K se representa por $z_{\alpha/2}$ siendo Z → N(0,1) .
- Es decir, se cumple: $P(- z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1- \alpha$



Como vemos en el dibujo $P(Z \geq z_{\alpha/2}) = \alpha/2$ y $P(Z \leq - z_{\alpha/2}) = \alpha/2$
Es decir, $z_{\alpha/2}$ es el valor de una distribución Normal estándar que deja a su derecha una probabilidad de $\alpha/2$, por tanto :

$\mathbf{P(Z \leq z_{\alpha/2}) = 1- \alpha/2}$ valor que podemos buscar en la tabla de la N(0,1) a la izquierda

INTERVALOS DE CONFIANZA

- La expresión más general para hacer estimaciones calculando intervalos de confianza es sumar y restar al estimador muestral z veces el error estándar (EE) del estimador:

  Parámetro ∈ (estimador ± z x EE del estimador)

  donde z es el valor critico correspondiente de la distribución normal o la que usemos ya que por ejemplo a veces, en vez de la distribución normal, se usa la distribución t de Student, debido a que no disponemos de la desviación estándar poblacional sino sólo la desviación estándar muestral.
- Veremos a continuación los principales I.C. y sus expresiones concretas.

INTERVALO DE CONFIANZA PARA LA MEDIA POBLACIONAL

Fijado un nivel de confianza de 1-α, queremos **dos valores** tales que la probabilidad de que la media de la población μ, se encuentre entre ellos sea precisamente 1-α

Si nos fijamos en la definición de valor critico:

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1-\alpha$$

De donde $$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

Despejando: $$P\left(-z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \overline{X}-\mu \leq z_{\alpha/2} \cdot \sigma/\sqrt{n}\right) = 1-\alpha$$

Y por tanto: $$P\left(\overline{X}-z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \overline{X}+z_{\alpha/2} \cdot \sigma/\sqrt{n}\right) = 1-\alpha$$

INTERVALO DE CONFIANZA PARA LA MEDIA POBLACIONAL

Es decir, el **intervalo de confianza para el parámetro μ, de una población N (μ, σ)** a un **nivel de confianza (1-α)** viene dado por:

$$\left(\bar{X} - z_{\alpha/2}\ \sigma/\sqrt{n}\ ; \bar{X} + z_{\alpha/2}\ \sigma/\sqrt{n}\right)$$

Si **σ es desconocida**, se sustituye por la cuasidesviacion típica o desviación típica de la muestra

$$\left(\bar{X} - z_{\alpha/2} S_{n-1}/\sqrt{n}\ ; \bar{X} + z_{\alpha/2} S_{n-1}/\sqrt{n}\right)$$

Recordamos que $S_{n-1} = S_n\ \left(\sqrt{n}/\sqrt{n-1}\right)$

$$\left(\bar{X} - z_{\alpha/2} S_n/\sqrt{n-1}\ ; \bar{X} + z_{\alpha/2} S_n/\sqrt{n-1}\right)$$

Donde $z_{\alpha/2}$ es el valor critico, valor de una Normal estándar que deja a su derecha una probabilidad de α/2

Nota : En lo sucesivo trabajaremos con S , donde S = S_{n-1} = S_n $\left(\sqrt{n}/\sqrt{n-1}\right)$
Nota: *tenemos que tener en cuenta que,* ***o bien n > 30 , o bien la distribución de partida es normal,*** *pues sólo así conocemos la distribución de las medias muestrales que es en lo que nos basamos para calcular el intervalo de confianza.*

INTERVALO DE CONFIANZA PARA LA MEDIA POBLACIONAL

Si la varianza σ es desconocida y n≤30

Partiendo de una población Normal, en estas condiciones la variable aleatoria se distribuye como una t-Student con *n-1* grados de libertad de la forma,

$$\frac{\bar{X} - \mu}{S\,/\sqrt{n}} \rightarrow t_{n-1}$$

Construimos el IC

$$P\left(-t_{\alpha/2;\ n-1} < \frac{\bar{X} - \mu}{S\ /\ \sqrt{n}} < t_{\alpha/2;\ n-1}\right) = P\left(-t_{\alpha/2;\ n-1}\frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{\alpha/2;\ n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Por tanto el IC sería:

$$\left(\bar{X} - t_{\alpha/2;\ n-1}\frac{S}{\sqrt{n}};\ \bar{X} + t_{\alpha/2;\ n-1}\frac{S}{\sqrt{n}}\right)$$

donde $t_{\alpha/2;\ n-1}$ es el valor critico, valor de una distribución t de Student con n – 1 grados de libertad que deja a su derecha una probabilidad de α/2.

INTERVALO DE CONFIANZA PARA LA PROPORCIÓN

INTERVALO DE CONFIANZA PARA LA PROPORCIÓN P DE UNA BINOMIAL (n, p)

Construimos el IC para la proporción poblacional, tipificando la variable

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right) = P\left(-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \hat{P} - p < z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

El IC de nivel (1-α) para la proporción poblacional p será:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

INTERVALO DE CONFIANZA PARA LA PROPORCION

Ejemplo: Se ha obtenido una muestra al azar de 150 vendedores de una Editorial para estimar la proporción de vendedores en la Editorial que no alcanza un límite de ventas mínimo establecido por la dirección. De entre los seleccionados, 50 no han conseguido llegar al límite de ventas mínimo establecido. Se pide el intervalo de confianza para la proporción de trabajadores en la Editorial que no alcanza el límite al 80 %.

Solución:

La proporción de la muestra es 50/150 = 0,333

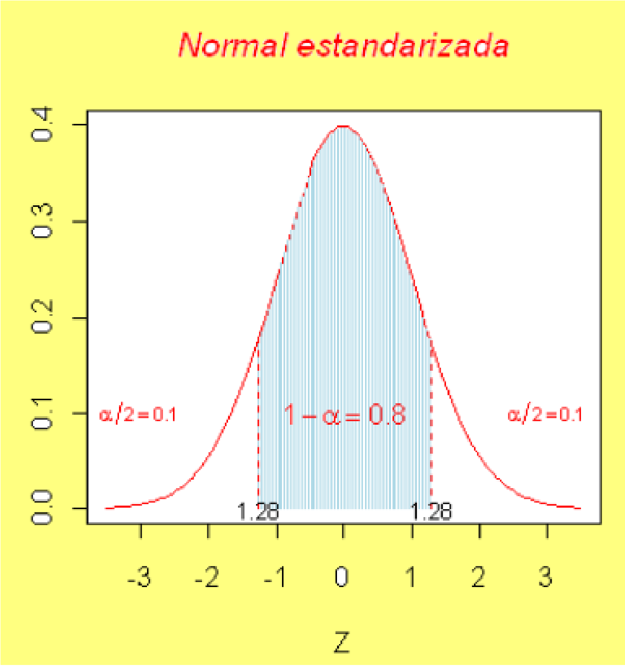
$1-\alpha = 0,8 \rightarrow \alpha/2 = 0,1$

Luego Los cuantiles de orden 0.1 y 0.9 para el nivel de confianza dado son -1.28 y 1.28, respectivamente

Sustituyendo en la expresión del IC

$$0{,}333 \pm 1{,}28\sqrt{\frac{0{,}333(1-0{,}333)}{150}}$$

$$(0{,}28, 0{,}38)$$



INTERVALO DE CONFIANZA PARA LA VARIANZA

INTERVALO DE CONFIANZA PARA LA VARIANZA σ^2 DE UNA DISTRIBUCIÓN NORMAL

El IC de nivel (1-α) será:

$$\left(\frac{(\boldsymbol{n-1})S^2}{\chi^2_{\alpha/2;\,(n-1)}};\ \frac{(\boldsymbol{n-1})S^2}{\chi^2_{(1-\alpha/2);\,(n-1)}}\right)$$

Donde $\chi^2_{\alpha/2;\,(n-1)}$ es el valor critico, valor de una distribución ji-cuadrado con $n - 1$ grados de libertad que deja a su derecha una probabilidad de $\alpha/2$.

INTERVALO DE CONFIANZA PARA LA VARIANZA

Ejemplo:

Se ha obtenido una muestra de 15 vendedores de una editorial para estimar el valor medio de las ventas por trabajador en la empresa. La media y varianza de la muestra (en miles de euros) son 5 y 2, respectivamente. Hallar el intervalo de confianza para la varianza de las ventas por trabajador en la editorial al 90 %.

El IC de nivel (1-α) será: $\left(\frac{(\boldsymbol{n-1})S^2}{\chi^2_{\alpha/2;\,(n-1)}};\ \frac{(\boldsymbol{n-1})S^2}{\chi^2_{(1-\alpha/2);\,(n-1)}}\right)$

$\alpha/2 = 0{,}05 \rightarrow 1-\alpha/2 = 0{,}95$; los grados de libertad son 14

$\chi^2_{\alpha/2;\,(n-1)} = 6{,}571$ y $\chi^2_{(1-\alpha/2);\,(n-1)} = 23{,}685$

como nos dan la varianza muestral tenemos $S^2 = (n/n-1)\, S_n^2 = (15/14)*2 = 2{,}143$

Sustituyendo en la formula tenemos

$$\frac{14\cdot 2{,}143}{23{,}685} \leq 2 \leq \frac{14\cdot 2{,}143}{6{,}571} \rightarrow (1{,}27;4{,}57)$$

INTERVALO DE CONFIANZA PARA LA VARIANZA

Distribución Chi-cuadrado

Percentiles de la distribución chi-cuadrado para ≤χ2												
χ2 / n	0,005	0,01	0,025	**0,05**	0,1	0,25	0,75	0,9	**0,95**	0,975	0,995	0,999
1	0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	1,3233	2,7055	3,8415	5,0239	7,8794	10,8276
2	0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	2,7726	4,6052	5,9915	7,3778	10,5966	13,8155
3	0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	4,1083	6,2514	7,8147	9,3484	12,8382	16,2662
4	0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	5,3853	7,7794	9,4877	11,1433	14,8603	18,4668
5	0,4117	0,5543	0,8312	1,1455	1,6103	2,6746	6,6257	9,2364	11,0705	12,8325	16,7496	20,5150
6	0,6757	0,8721	1,2373	1,6354	2,2041	3,4546	7,8408	10,6446	12,5916	14,4494	18,5476	22,4577
7	0,9893	1,2390	1,6899	2,1673	2,8331	4,2549	9,0371	12,0170	14,0671	16,0128	20,2777	24,3219
8	1,3444	1,6465	2,1797	2,7326	3,4895	5,0706	10,2189	13,3616	15,5073	17,5345	21,9550	26,1245
9	1,7349	2,0879	2,7004	3,3251	4,1682	5,8988	11,3888	14,6837	16,9190	19,0228	23,5894	27,8772
10	2,1559	2,5582	3,2470	3,9403	4,8652	6,7372	12,5489	15,9872	18,3070	20,4832	25,1882	29,5883
11	2,6032	3,0535	3,8157	4,5748	5,5778	7,5841	13,7007	17,2750	19,6751	21,9200	26,7568	31,2641
12	3,0738	3,5706	4,4038	5,2260	6,3038	8,4384	14,8454	18,5493	21,0261	23,3367	28,2995	32,9095
13	3,5650	4,1069	5,0088	5,8919	7,0415	9,2991	15,9839	19,8119	22,3620	24,7356	29,8195	34,5282
14	4,0747	4,6604	5,6287	6,5706	7,7895	10,1653	17,1169	21,0641	23,6848	26,1189	31,3193	36,1233
15	4,6009	5,2293	6,2621	7,2609	8,5468	11,0365	18,2451	22,3071	24,9958	27,4884	32,8013	37,6973
16	5,1422	5,8122	6,9077	7,9616	9,3122	11,9122	19,3689	23,5418	26,2962	28,8454	34,2672	39,2524
17	5,6972	6,4078	7,5642	8,6718	10,0852	12,7919	20,4887	24,7690	27,5871	30,1910	35,7185	40,7902
18	6,2648	7,0149	8,2307	9,3905	10,8649	13,6753	21,6049	25,9894	28,8693	31,5264	37,1565	42,3124
19	6,8440	7,6327	8,9065	10,1170	11,6509	14,5620	22,7178	27,2036	30,1435	32,8523	38,5823	43,8202
20	7,4338	8,2604	9,5908	10,8508	12,4426	15,4518	23,8277	28,4120	31,4104	34,1696	39,9968	45,3147
21	8,0337	8,8972	10,2829	11,5913	13,2396	16,3444	24,9348	29,6151	32,6706	35,4789	41,4011	46,7970
22	8,6427	9,5425	10,9823	12,3380	14,0415	17,2396	26,0393	30,8133	33,9244	36,7807	42,7957	48,2679
23	9,2604	10,1957	11,6886	13,0905	14,8480	18,1373	27,1413	32,0069	35,1725	38,0756	44,1813	49,7282
24	9,8862	10,8564	12,4012	13,8484	15,6587	19,0373	28,2412	33,1962	36,4150	39,3641	45,5585	51,1786
25	10,5197	11,5240	13,1197	14,6114	16,4734	19,9393	29,3389	34,3816	37,6525	40,6465	46,9279	52,6197
26	11,1602	12,1981	13,8439	15,3792	17,2919	20,8434	30,4346	35,5632	38,8851	41,9232	48,2899	54,0520
27	11,8076	12,8785	14,5734	16,1514	18,1139	21,7494	31,5284	36,7412	40,1133	43,1945	49,6449	55,4760
28	12,4613	13,5647	15,3079	16,9279	18,9392	22,6572	32,6205	37,9159	41,3371	44,4608	50,9934	56,8923
29	13,1211	14,2565	16,0471	17,7084	19,7677	23,5666	33,7109	39,0875	42,5570	45,7223	52,3356	58,3012
30	13,7867	14,9535	16,7908	18,4927	20,5992	24,4776	34,7997	40,2560	43,7730	46,9792	53,6720	59,7031
40	20,7065	22,1643	24,4330	26,5093	29,0505	33,6603	45,6160	51,8051	55,7585	59,3417	66,7660	73,4020
50	27,9907	29,7067	32,3574	34,7643	37,6886	42,9421	56,3336	63,1671	67,5048	71,4202	79,4900	86,6608
60	35,5345	37,4849	40,4817	43,1880	46,4589	52,2938	66,9815	74,3970	79,0819	83,2977	91,9517	99,6072
70	43,2752	45,4417	48,7576	51,7393	55,3289	61,6983	77,5767	85,5270	90,5312	95,0232	104,2149	112,3169
80	51,1719	53,5401	57,1532	60,3915	64,2778	71,1445	88,1303	96,5782	101,8795	106,6286	116,3211	124,8392
90	59,1963	61,7541	65,6466	69,1260	73,2911	80,6247	98,6499	107,5650	113,1453	118,1359	128,2989	137,2084
100	67.3276	70.0649	74.2219	77.9295	82.3581	90.1332	109.1412	118.4980	124.3421	129.5612	140.1695	149.4493

Estadística

Tema 6: Contraste de Hipótesis

Contenido

Hipótesis Estadísticas

Una hipótesis estadística (H) es una proposición acerca de una característica de la población de estudio.

Por ejemplo:

- "la variable X toma valores en el intervalo (a, b)", "el valor de θ es 2", "la distribución de X es normal", etc.

- Una compañía recibe un gran cargamento de piezas. Solo acepta el envío si no hay más de un 5% de piezas defectuosas. ¿Cómo tomar una decisión sin verificar todas las piezas?
- Se quiere saber si una propuesta de reforma legislativa es acogida de igual forma por hombres y mujeres. ¿Cómo se puede verificar esa conjetura?

Los dos últimos ejemplos tienen algo en común:

- Se formula la hipótesis sobre la población.
- Las conclusiones sobre la validez de la hipótesis se basarán en la **información de una muestra**.

Hipótesis Estadísticas

Tipos de hipótesis estadísticas

Hipótesis paramétricas: Una hipótesis paramétrica es una proposición sobre los valores que toma un parámetro.

- Hipótesis simple: aquella que especifica un único valor para el parámetro. Ejemplos: "$H : \theta = 0$", "$H : \theta = -23$", etc.
- Hipótesis compuesta: aquella que especifica un intervalo de valores para el parámetro. Ejemplos: "$H : \theta \geq 0$", "$H : 1 \leq \theta \leq 4$", etc.
  - Hipótesis unilateral: "$H : \theta \leq 4$", '$H : 0 < \theta$", etc.
  - Hipótesis bilateral: "$H : \theta \neq 4 \Leftrightarrow H : \theta < 4$ y $\theta > 4$"

Hipótesis no paramétricas: Una hipótesis no paramétrica es una proposición sobre cualquier otra característica de la población.
Ejemplos: "$H : X \sim N$", "$H : X$ independiente Y ", etc. **(no se estudia en este curso)**

Contraste de Hipótesis

Un **contraste de hipótesis** (también denominado **test de hipótesis o prueba de significación**) es un procedimiento para decidir si una propiedad que suponemos se da en una población estadística es compatible con lo observado en una muestra de dicha población.

Es decir son los procedimientos para aceptar o rechazar una hipótesis que se emite acerca de un parámetro u otra característica de la población.

ETAPAS DEL PROCESO

- El investigador formula una hipótesis sobre un parámetro poblacional, por ejemplo que toma un determinado valor
- Se selecciona una muestra de la población
- Se comprueba si los datos están o no de acuerdo con la hipótesis planteada, es decir se **compara la observación con la teoría**
    i. Si lo observado es incompatible con lo teórico entonces el investigador no acepta la hipótesis planteada y deberá proponer una nueva teoría.
    ii. Si lo observado es compatible con lo teórico entonces el investigador aceptara la hipótesis planteada como cierta.

Identificación de Hipótesis

Seguro que te ha pasado que te sientes la cabeza como un bombo, sensación extraña en la piel, algún escalofrío y ganas de nada. Estar en el sofá. Claros indicios que tienes fiebre, ¿verdad? Entonces piensas. Voy a comprobarlo. Te pones el termómetro y efectivamente marca 38ºC. Ya sabes lo que toca. Antitérmico y sofá.

Esto es precisamente **un contraste de hipótesis**. Tan sencillo cómo esto.

Tu **tienes una intuición** y quieres investigar si esta intuición es cierta. **Tu hipótesis de investigación** es: “tengo fiebre”. También **llamada hipótesis del investigador** o **alternativa** (**H1**). Es la que se quiere corroborar. La que no es habitual. Es un estado raro. Curioso. Nuevo.

Por el contrario tienes **la hipótesis nula**. ¿Por que nula?. Es la contraria. “NO tengo fiebre”, es la que **no te gustaría aceptar si eres un investigador** ya que quieres comprobar si tus datos dan la razón a tus suposiciones, intuiciones etc.

https://conceptosclaros.com/contraste-hipotesis/

Identificación de hipótesis

Mas Ejemplos

- *Decidir la inocencia o culpabilidad de una persona en un país en el que se sigue el principio de presunción de inocencia:*

Como se quiere evitar condenar a una persona inocente, sólo se hará cuando haya una fuerte evidencia de su culpabilidad, cuando esté demostrada la culpabilidad. En caso de duda, se primará la inocencia frente a la culpabilidad. Por tanto, en la terminología propuesta sería:

$\mathbf{H_0}$: Inocente

$\mathbf{H_1}$: Culpable

- *Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura:*

Desde el punto de vista inicial del profesorado, un estudiante no sabe la asignatura mientras no demuestre lo contrario; es decir, el examen ha de presentar pruebas suficientes de que conoce la asignatura. En general, en caso de duda o de falta de datos, se supondrá que el estudiante no sabe la asignatura, es decir suspenso frente a que si la sabe, aprobado. Por tanto, en la terminología propuesta sería:

$\mathbf{H_0}$: *El estudiante NO sabe la asignatura: suspenso*

$\mathbf{H_1}$: *El estudiante SI sabe la asignatura: aprobado*

Contraste de Hipótesis

Llamamos **hipótesis nula**, y la representamos por $\mathbf{H_0}$, a la hipótesis que mantendremos a no ser que los datos indiquen claramente que hay que adoptar otra posición.

- Es una idea es similar a la presunción de inocencia en un juicio.
- La hipótesis nula siempre contiene los signos "=", "≤" o "≥".
- La hipótesis nula, **se rechaza o no se rechaza**.

Llamamos **hipótesis alternativa**, y la representamos por $\mathbf{H_1}$, es la opción a contrastar, la posición contraria es decir, la negación de la hipótesis nula.

- Es generalmente la hipótesis que se quiere verificar.
- La hipótesis alternativa nunca contiene los signos "=", "≤" o "≥".
- La hipótesis alternativa **puede aceptarse o no aceptarse**.

> - Si aceptamos $\mathbf{H_1}$ implica rechazar $\mathbf{H_0}$
> - Si no aceptamos $\mathbf{H_1}$ implica no rechazar $\mathbf{H_0}$

Identificación de hipótesis

- **Hipótesis nula H_o**
  - La que contrastamos
  - Los datos pueden refutarla
  - No debería ser rechazada sin una buena razón.
- **Hipótesis H_I**
  - Niega a H_0 (H_I es la hipótesis que nos interesa).
  - Los datos pueden mostrar evidencia a favor
  - No debería ser aceptada sin **una gran evidencia a favor**.

$$\begin{cases} H_0 : & p = 50\% \\ H_1 : & p \neq 50\% \end{cases}$$

$= , \leq , \geq$

$\neq , < , >$

Contraste de Hipótesis

<u>Errores</u>

Cuando trabajamos con el método del contraste de hipótesis podemos cometer dos tipos de errores:

- Si se rechaza la hipótesis (**H_0**) cuando debería ser no rechazada, es decir se acepta la hipótesis (**H_1**), estamos ante un **error de tipo I**, **error $\boldsymbol{\alpha}$ o nivel de significación**
- Si no se rechaza la hipótesis (**H_0**), que debería ser rechazada, es decir no se acepta la hipótesis (**H_1**), estamos ante un **error de tipo II o error $\boldsymbol{\beta}$**

Para que un contraste de hipótesis sea bueno se debe diseñar de forma que se minimicen los errores de decisión

	Rechazar H_0	**No rechazar H_0**
H_0 cierta	**Error de tipo I**	Decisión correcta
H_1 cierta	Decisión correcta	**Error de tipo II**

Contraste de Hipótesis

En los ejemplos anteriores:

1.- Decidir la inocencia o culpabilidad de una persona en un estado en el que se sigue el principio de presunción de inocencia:

		Rechazar H_0	**No rechazar H_0**
H_0: *Inocente* H_1: *Culpable*	H_0 **cierta**	**Error de tipo I:** Se condena a una persona que es inocente	Se absuelve a una persona que es inocente
	H_1 **cierta**	Se condena a una persona que es culpable	**Error de tipo II:** Se absuelve a una persona que es culpable

Contraste de Hipótesis

En los ejemplos anteriores:

2.- Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura:

		Rechazar H_0	**No rechazar H_0**
H_0: *No sabe la asignatura* H_1: *Sí sabe la asignatura*	**H_0 cierta**	**Error de tipo I:** Se aprueba a un estudiante que NO sabe la asignatura	Se suspende a un estudiante que no sabe la asignatura
	H_1 cierta	Se aprueba a un estudiante que sí sabe la asignatura	**Error de tipo II:** Se suspende a un estudiante que SÍ sabe la asignatura

Contraste de Hipótesis

Nivel de Significación y Potencia

Llamaremos **nivel de significación, $\boldsymbol{\alpha}$** , a la probabilidad de cometer un error de tipo I, es decir, probabilidad de rechazar H_0 cuando es cierta

$$\alpha = P(\text{ rechazar } H_0 \,/\, H_0 \text{ es cierta}).$$

Llamaremos **potencia del contraste** al valor de **1- $\boldsymbol{\beta}$** , siendo β la probabilidad de cometer un error de tipo II, es decir,

$$\beta = P\,(\text{ NO Rechazar } H_0 \,/\, H_0 \text{ es falsa})$$

En general, se fija de antemano un ***nivel de confianza (1- $\boldsymbol{\alpha}$)*** expresado normalmente en %, que *"asegure"* un error de tipo I admisible (haciendo mínima la probabilidad de "condenar a un inocente") y de entre todos los contrastes con dicho nivel de confianza se elige el de mayor potencia.

α y $\boldsymbol{\beta}$ están inversamente relacionadas. Sólo pueden disminuirse las dos, aumentando el tamaño muestral n.

Nota : El estudio de la potencia de un test se escapa al nivel de este curso, así que daremos por hecho que los contrastes que planteemos cumplen la condición de equilibrio adecuada.

Región de Aceptación y Región Crítica

Ahora necesitamos un criterio para saber **con cuál de las dos hipótesis nos quedamos**

Para rechazar la hipótesis nula, es necesario que las evidencias sean muy fuertes; es decir, puede que haya cambios debidos al azar, en cuyo caso el cambio no es significativo, y no cambiamos , pero puede que los cambios sean debidos a otras causas. En este último caso es cuando el **cambio es significativo y rechazaremos la hipótesis nula**.

- Primero fijamos un cierto intervalo dentro del cual es normal que haya cambios, es decir, una región tal que si el parámetro (por ejemplo media o proporción) se mantiene en dicho intervalo, nos seguimos quedando con H_0, pues esas pequeñas variaciones son debidas al azar. Ese intervalo o región se denomina **región de aceptación**, y será mayor o menor dependiendo del nivel de confianza **(1-** α), que precisemos.

- La región que quede fuera de la región de aceptación indica que en este caso los cambios no se pueden atribuir al azar, y por tanto hemos de rechazar H_0 y aceptar H_1. Tal región se llama **región crítica o de rechazo**.

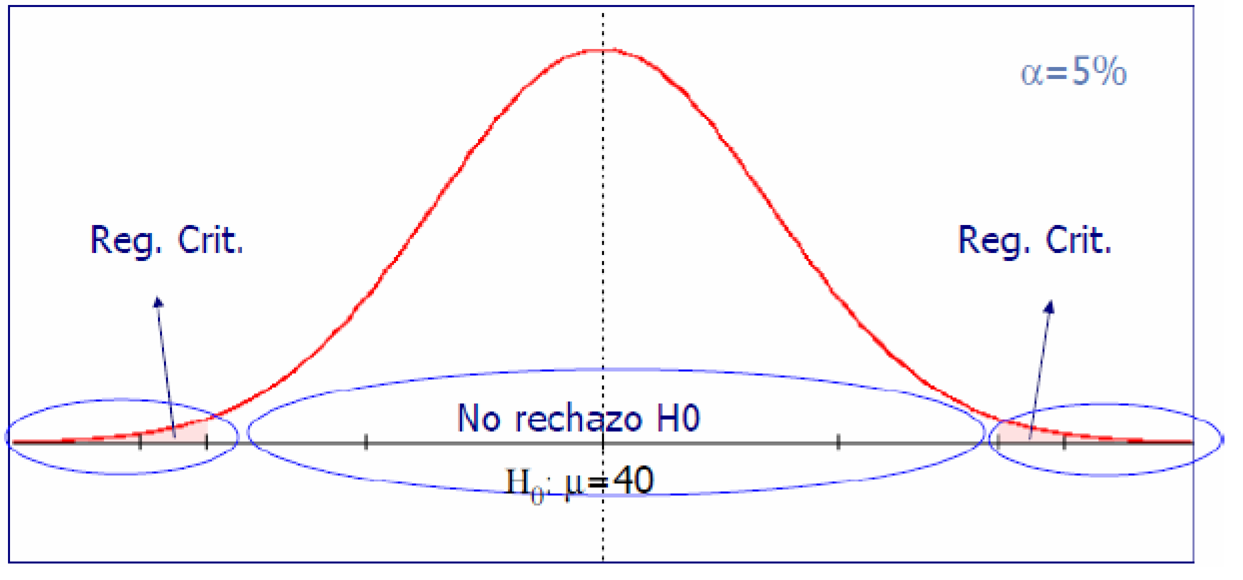
Región de Aceptación y Región Crítica

Así, gráficamente, si nuestro test de hipótesis viene planteado por ejemplo como sigue:

Fijamos un nivel de confianza del 95 % es decir $(1-\alpha) = 0,95$

El nivel de significación $\alpha = 0,05$ o del 5%

$$\begin{cases} H_0 & : \mu = 40\% \\ H_1 & : \mu \neq 40\% \end{cases}$$



Regla de Decisión

La **regla de decisión** quedara definida de acuerdo a la región crítica (o equivalentemente a la región de aceptación). A esta región le corresponde un determinado **nivel de significación**.

Por otro lado, la **información contenida en la muestra** se resume mediante un **estadístico** (esto es, un valor que es función de la muestra) cuya distribución de probabilidad esta relacionada con la hipótesis en estudio y **sea conocida**. También puede ser denominado **Función de decisión.** Por tanto definiremos la **región crítica en función del estadístico empleado**.

Resumiendo, contraste de hipótesis H_0 frente a H_1:

- α = nivel de significación,
- T = estadístico de test o estadístico de contraste
- $W\alpha$ = región crítica, subconjunto del espacio muestral definido a partir del estadístico T (tb RC)
- Regla de decisión:
  - si la muestra pertenece a $W\alpha$, se rechaza H_0 $\rightarrow$ se acepta H_1
  - si la muestra no pertenece a $W\alpha$, no se acepta H_1 $\rightarrow$ no se rechaza H_0

Contraste de Hipótesis

Pasos para la construcción de un contraste

1. Identificar el parámetro de interés ($\mu \rightarrow$ media poblacional)

2. Establecer las **hipótesis H_0 y H_1.**

3. Fijar un **nivel de significación α**

4. Determinar el **estadístico del contraste bajo H_0** (igual que en la construcción del intervalo de confianza), este estadístico proporcionara una medida de discrepancia entre la hipótesis nula y la información muestral. Esta medida estará en función de la diferencia del valor que especifica H_0 para el parámetro y el estimador muestral del parámetro, y tendrá distribución conocida.

Contraste de Hipótesis

Pasos para la construcción de un contraste

5. Establecer las **regiones de aceptación y rechazo**. Mediante el valor crítico. En este paso determinamos la discrepancia máxima que estamos dispuestos a admitir para aceptar H_0. Este valor dependerá de la distribución del estadístico de contraste bajo H_0, del nivel de significación α especificado y del tipo de hipótesis alternativa que tengamos.

6. Calcular el **valor** que toma el estadístico del contraste para la **muestra seleccionada.**

7. **Decidir** si se debe o no rechazar H_0 e interpretar la decisión tomada. Si el estadístico de contraste observado (empírico) cae en la región de rechazo, rechazamos H_0, en caso contrario, mantendremos como cierta H_0.

Nota: *Se denomina región de aceptación a la región que conduce a no rechazar H_0 y región de rechazo a la región que conduce a rechazar H_0.*

Contraste de Hipótesis para la Media μ (con σ conocida)

Contraste Bilateral

- Hipótesis

$$\left.\begin{array}{ll} H_0 \equiv & \mu = \mu_0 \\ H_1 \equiv & \mu \neq \mu_0 \end{array}\right\}$$

- Estadístico de contraste

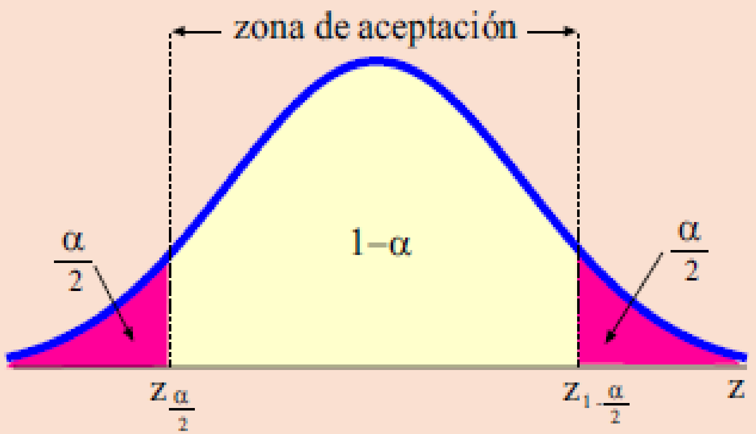
$$\mathbf{z} = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Valor crítico para un nivel α

$$z_{1-\alpha/2}$$

- Criterio de decisión :

Región de aceptación y región crítica a un nivel de confianza $1 - \alpha$



Si $z_{\alpha/2} < Z < z_{1-\alpha/2}$ no se puede rechazar la hipótesis nula

En caso contrario Se acepta H_1

Zona de aceptación

$$\mu \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Contraste de Hipótesis para la Media μ (con σ conocida)

Ejemplo: Los gastos corrientes por empleado de los distintos departamentos de una empresa siguen una distribución normal con desviación típica de 300 €. En una muestra de 16 departamentos, se ha obtenido un gasto medio por empleado de 1.350 €. Determina, para un nivel de confianza del 99%, si el gasto corriente medio por empleado en la empresa es de 1.280€

$$H_0: \mu = \mu_0 = 1.280€$$

$$H_1: \mu \neq \mu_0 = 1.280€$$

Estadístico para el contraste $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \rightarrow N(0,1)$

Región de aceptación para el nivel de confianza dado

$\alpha = 0{,}01 \rightarrow$ valor critico: $z_{\alpha/2} = z_{0,005} = z_{0,995} = 2{,}58$

es decir $\rightarrow$ La región de aceptación es (–2,58; 2,58)

$Z = \dfrac{1350 - 1280}{300 / \sqrt{16}} = 0{,}93$. Como $-2{,}58 < 0{,}93 < 2{,}58$,

con una probabilidad del 99%, **no se puede rechazar la hipótesis nula**

Contraste de Hipótesis para la Media μ (con σ conocida)

- **Contrastes Unilaterales**

Hipótesis $\left.\begin{array}{ll} H_0 \equiv & \mu \geq \mu_0 \\ H_1 \equiv & \mu < \mu_0 \end{array}\right\}$

Estadístico de contraste

$$\mathbf{z} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

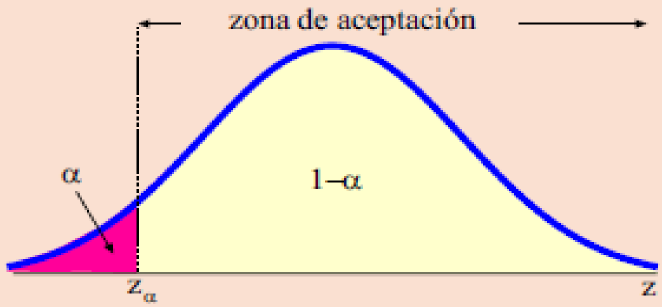
Valor crítico para un nivel α

$$z_\alpha$$

Criterio de decisión:

Si Z > z_α no se puede rechazar la H_0

Si Z < z_α Se rechaza H_0



Hipótesis $\left.\begin{array}{ll} H_0 \equiv & \mu \leq \mu_0 \\ H_1 \equiv & \mu > \mu_0 \end{array}\right\}$

Estadístico de contraste

$$\mathbf{z} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

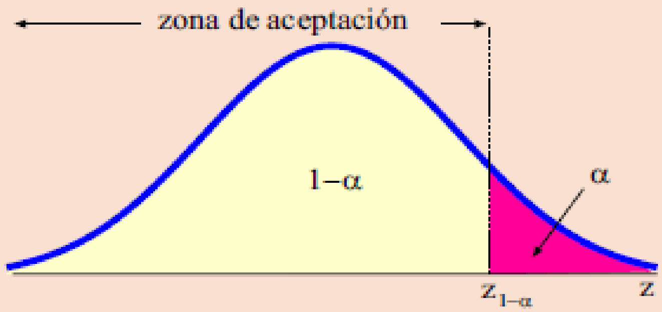
Valor crítico para un nivel α

$$z_{1-\alpha}$$

Criterio de decisión:

Si Z < $z_{1-\alpha}$ no se puede rechazar la H_0

Si Z > $z_{1-\alpha}$ Se rechaza H_0



Contraste de Hipótesis para la Media μ (con σ conocida)

Ejemplo: Una empresa garantiza que unas cuerdas que fabrica soportan, a lo sumo, un peso medio de 150 kg con una desviación típica de 12 kg. Para verificar esta afirmación, se toma una muestra de 64 cuerdas, y se obtiene un peso medio de 152 kg. ¿Se puede afirmar, con un nivel de confianza del 95%, que la afirmación de la empresa es verdadera?

H_0: $\mu \leq 150$

H_1: $\mu > 150$

Estadístico para el contraste $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \rightarrow N(0,1)$

Región de aceptación para el nivel de confianza dado $1 - \alpha = 0{,}95$

$\rightarrow z_\alpha = z_{0,05} = z_{0,95} = 1{,}65$ es decir $\rightarrow$ La región de aceptación es $(-\infty; 1{,}65)$

$Z = \dfrac{152-150}{12 / \sqrt{64}} = 1{,}33$. Como $1{,}33 \in (-\infty; 1{,}65)$ con una probabilidad del 95%,

no se puede rechazar la hipótesis nula

Contraste de Hipótesis para la Media μ (con σ desconocida y n≥30)

Se realizan de la misma forma que antes, sustituyendo la varianza por la cuasivarianza muestral, por ejemplo el contraste bilateral sería :

- Hipótesis

$$\left.\begin{array}{ll} H_0 \equiv & \mu = \mu_0 \\ H_1 \equiv & \mu \neq \mu_0 \end{array}\right\}$$

- Estadístico de contraste

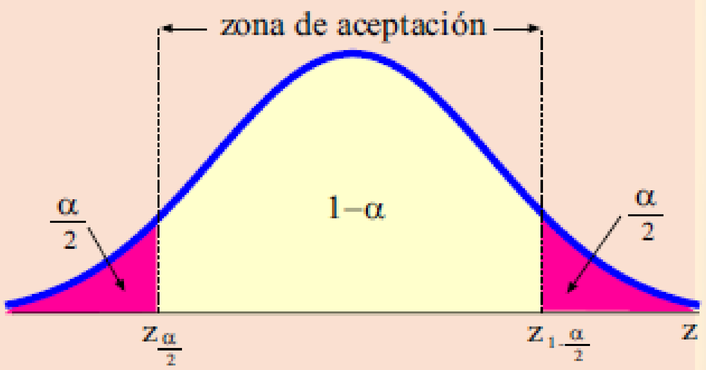
$$z = \frac{\overline{x} - \mu_0}{\hat{s}/\sqrt{n}}$$

- Valor crítico para un nivel α

$$z_{1-\alpha/2}$$

- Criterio de decisión :

Región de aceptación y región crítica a un nivel de confianza $1-\alpha$



Si $z_{\alpha/2} < Z < z_{1-\alpha/2}$ no se puede rechazar la hipótesis nula

En caso contrario Se acepta H_1

Recuerda que la cuasidesviación es $\hat{s} = \sqrt{\frac{n}{n-1}}\, s$

Contraste de Hipótesis

Para la media con Varianza de la población desconocida y muestras pequeñas ($n \leq 30$) Donde S es la cuasivarianza muestral (S_n) :

Partiendo de una población Normal, en estas condiciones el estadístico de contraste se distribuye como una t-Student con *n-1* grados de libertad

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \rightarrow t_{n-1}$$

A) $\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}$ RECHAZO H_0 si $\left| \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} \right| > t_{n-1;\frac{\alpha}{2}}$

B) $\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}$ RECHAZO H_0 si $\dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} > t_{n-1;\alpha}$

C) $\begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}$ RECHAZO H_0 si $\dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} < -t_{n-1;\alpha}$

Contraste de Hipótesis para la Media μ (con σ conocida)

Ejemplo: Se escoge a 17 individuos al azar y se les mide resultando que su estatura media es de 1,71 metros con varianza muestral de 0,02. Contrastar la hipótesis de que la estatura media nacional sea de 1.75 metros si utilizamos un nivel de la significación del 5%. Se supone normalidad

$H_0: \mu = 1{,}75$
$H_1: \mu \neq 1{,}75$

Estadístico para el contraste $T = \dfrac{\bar{X} - \mu}{S \,/\, \sqrt{n}} \rightarrow t_{n-1}$

Región de aceptación para el nivel de confianza dado $1 - \alpha = 0{,}95$

$\rightarrow t_{\alpha/2} = t_{0,025} = t_{0,975} = 2{,}12$ es decir $\rightarrow$ La región de aceptación es (-2,12; 2,12)

Como nos dan la desviación muestral (Sn) recordamos que: $S = S_{n-1} = S_n \left(\sqrt{n} / \sqrt{(n-1)}\right)$

$T = \dfrac{171-175}{0{,}02 \,/\, \sqrt{16}} = -8.$ Como $-8 \notin (-2{,}12;\ 2{,}12)$ con una probabilidad del 95%,

Se rechaza la hipótesis nula

Contraste de la proporción de una población binomial

Muestras Grandes

- Hipótesis

$$\left.\begin{array}{l} H_0 \equiv \quad p = p_0 \\ H_1 \equiv \quad p \neq p_0 \end{array}\right\}$$

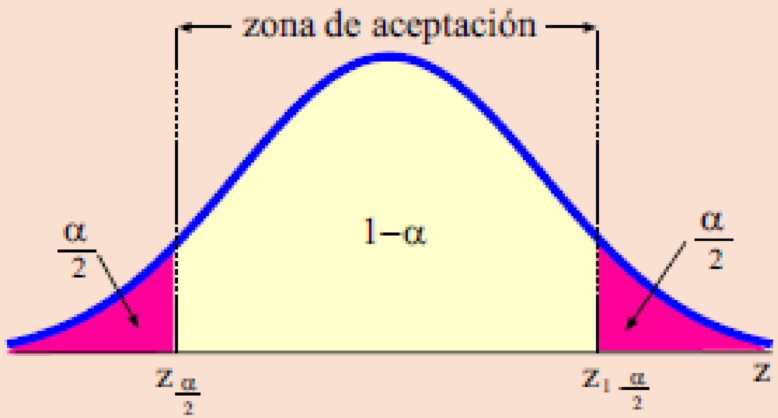
- Estadístico de contraste

$$\mathbf{z} = \frac{\hat{p} - p_0}{\sqrt{\frac{p\,q}{n}}}$$

- Valor crítico para un nivel α

$$z_{1-\alpha/2}$$

Región de aceptación y región crítica a un nivel de confianza $1-\alpha$



- Criterio de Decisión:

  Si $z_{\alpha/2} < Z < z_{1-\alpha/2}$ no se puede rechazar la hipótesis nula

  En caso contrario Se acepta H_1

Contraste de la proporción de una población binomial

Al lanzar 5000 veces una moneda al aire salieron 3000 caras. ¿Se puede aceptar, con un nivel de significación del 0.05, que la moneda está trucada?

Solución: La proporción de la muestra es $\hat{p} = \dfrac{3000}{5000} = 0.6$

- Hipótesis

$$\left.\begin{array}{ll} H_0 \equiv & p = 0.5 \\ H_1 \equiv & p \neq 0.5 \end{array}\right\}$$

- Estadístico de contraste

$$\mathbf{z} = \frac{\hat{p} - p}{\sqrt{\frac{p\cdot q}{n}}} = \frac{0.6 - 0.5}{\sqrt{\frac{0.5\cdot 0.5}{5000}}} = 14.14$$

- Valor crítico con $\alpha = 0.05$

$$z_{1-\alpha/2} = z_{0.975} = 1.96$$

- Decisión para bilateral

$$|z| = 14.14 > 1.96$$

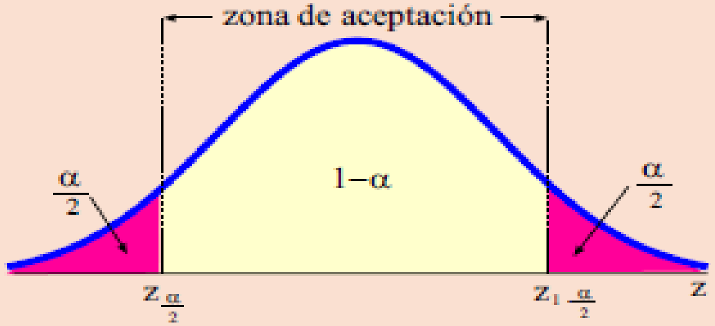
Se rechaza la H_0.

- Zona de aceptación

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p\cdot q}{n}}$$

$$0.5 \pm 1.96 \cdot \sqrt{\frac{0.5 \cdot 0.5}{5000}}$$

$$(0.486; 0.514)$$



- $0.6 \notin (0.486; 0.514)$ se rechaza H_0

Contraste para DIFERENCIAS DE MEDIAS

Diferencia de medias para muestras independientes con distribución normal y varianzas de las poblaciones conocidas :

Las Hipótesis son: H_0: $\mu_x - \mu_y = \mu_0$ H_1: $\mu_x - \mu_y \neq \mu_0$

$$Z = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \rightarrow N(0,1)$$

Si H_0 es cierta: α es la probabilidad de rechazarla y $1 - \alpha$ es la probabilidad de no rechazarla. Siendo α el nivel de significación del contraste.

$$Sabemos\ que{:}\ p\left(-Z\alpha_{/2} < \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} < Z\alpha_{/2}\right) = 1 - \alpha$$

Por lo que no se puede rechazar H_0 con probabilidad *p* cuando: $-Z\alpha_{/2} < Z < Z\alpha_{/2}$

Contraste de Hipótesis

Ejercicio: En la Universidad A, 400 alumnos con una media de 16 errores por examen, Universidad B, 200 alumnos, con una media de 15 errores. Desviaciones típicas poblacionales son de 40 para la Universidad A y 20 para la Universidad B ¿Se puede decir que el nivel de aprendizaje de estadística en ambas Universidades es el mismo con $\alpha = 4\%$?:

Las poblaciones siguen una distribución normal y conocemos las varianzas poblacionales. Las Hipótesis son: H_0: $\mu_x - \mu_y = 0$ $\quad H_1$: $\mu_x - \mu_y \neq 0$

Contraste sobre la variable $N(0,1)$): $\dfrac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} = \dfrac{16-\overline{15}-(0)}{\sqrt{\frac{40^2}{400}+\frac{20^2}{200}}} = 0{,}408$

Siendo α = 4% el nivel de significación del contraste→ $Z_{\alpha/2}$= 2,06

Como $(-2{,}06 < 0{,}408 < 2{,}06)$

Con probabilidad 96% No se rechaza la hipótesis nula

Contraste de la proporción de una población binomial

EJERCICIOS

1. En una determinada población juvenil, el peso, en kg sigue una distribución normal N(50, 10). Si se extrae una muestra aleatoria de 25 jóvenes y para un nivel de significación del 5%, ¿en qué condiciones se rechazaría la hipótesis de que la media de la población es de 50 kg?

2. Se realiza un sondeo electoral en que se obtiene que de 1600 electores seleccionados mediante un muestreo aleatorio simple 960 se muestran favorables al sí ¿Es aceptable la hipótesis de que el porcentaje de votos afirmativos será del 65% con un nivel se significación del 5%?: