

Probabilidades y Estadística Cs. de la Computación

Introducción

Breve reseña histórica:

La teoría de Probabilidades comienza a partir de una disputa entre jugadores en 1654. Los dos matemáticos que participaron de tales discusiones fueron Blaise Pascal y Pierre de Fermat, y su intercambio de correspondencia sentó las bases de la teoría de Probabilidades. Un matemático holandés, Christian Huygens tomó contacto con esa correspondencia y escribió el primer libro sobre Probabilidades en 1657, el cual trataba fundamentalmente sobre problemas relacionados con los juegos de azar.

Durante el siglo XVIII la teoría se desarrolló y se enriqueció con los aportes de Jacob Bernoulli y Abraham de Moivre. En 1812 Pierre de Laplace introdujo una serie de nuevas ideas y técnicas matemáticas en su libro *Theorie Analytique des Probabilités* y fundamentalmente sacó a la teoría del marco exclusivo de los juegos de azar y aplicó las ideas a muchos problemas científicos y prácticos. Algunas de las importantes aplicaciones desarrolladas en el siglo XIX fueron: teoría de errores, matemática actuarial y mecánica estadística.

Una de las dificultades para el desarrollo de la teoría matemática de las probabilidades fue llegar a una definición de probabilidad matemáticamente rigurosa, pero al mismo tiempo amplia para permitir su aplicación a un amplio rango de fenómenos. En el siglo XX se llegó a una definición axiomática de las Probabilidades (Kolmogorov, 1933).

¿Porqué estudiar Probabilidades y Estadística en Ciencias de la Computación?:

Posibles preguntas que queremos responder:

- ¿Cuál es el máximo número de terminales que pueden estar conectadas en un servidor antes de que el tiempo medio de espera se haga inaceptable?
- En una base de datos, ¿Cómo deberían ser guardados los datos para minimizar el tiempo medio de acceso?

Los sistemas de computación no son determinísticos. Pensemos, por ejemplo, en el delay en el envío de paquetes, comunicaciones en una red, equilibrio de "carga" en servidores, requerimientos de memoria, etc.

¿Para qué sirven las Probabilidades? Si bien estamos frente a procesos aleatorios, no son necesariamente "caóticos", en el sentido que podemos descubrir un patrón de comportamiento que pueda ser modelado.

Veamos un ejemplo de uso frecuente.

Compresión de archivos: El código ASCII contiene 256 caracteres, cada uno de los cuáles se representa con un número consistente en 8 dígitos binarios, por ejemplo, á se representa por $160 \equiv 10100000$.

Para simplificar el problema, supongamos que contamos con sólo 4 caracteres: A, B, C y D. Para representarlos necesitamos 2 bits. Por ejemplo, podríamos representarlos así:

A → 00
B → 01
C → 10
D → 11

Si un texto constara de n caracteres necesitaríamos $2n$ bits para guardarlo. Esta cantidad de bits es *determinística*.

Supongamos que sabemos que ciertas letras aparecen con más frecuencia que otras, por ejemplo, supongamos que sabemos que las frecuencias con que aparecen las 4 letras en un texto son:

A 0.70 (70%)
B 0.12 (12%)
C 0.10 (10%)
D 0.08 (8%)

El método de codificación de Huffman utiliza la información disponible sobre la frecuencias de aparición de los caracteres y asigna códigos de longitud variable. Por ejemplo, podríamos asignar a los 4 caracteres de nuestro ejemplo los siguientes códigos:

A → 1
B → 00
C → 011
D → 010

¿Cuánto espacio (en bits) ocuparía ahora un texto de n caracteres? No lo sabemos, pero podemos suponer que tendremos en promedio:

0.70 n veces A's
0.12 n veces B's
0.10 n veces C's
0.08 n veces D's

y el número de bits requerido sería:

$$0.70 n * (1) + 0.12 n *(2) + 0.10 n * (3) + 0.08 n * (3) = 1.48 n.$$

Como se observa, el método produce una disminución del espacio promedio requerido para almacenar un texto.

Probabilidad

El término Probabilidad se refiere al estudio del azar y la incertidumbre. En aquellas situaciones en las cuáles se puede producir uno de varios resultados posibles, la Teoría de la Probabilidad provee métodos para cuantificar la chance de ocurrencia de cada uno de ellos.

Ejemplos:

- Se arroja un dado dos veces y se registra la suma de puntos. ¿Cuál es la probabilidad de que se obtenga una suma mayor que 10?
- En un juego de ruleta, ¿cuál es la probabilidad de ganar apostando a primera columna?
- En un juego de ruleta, ¿cuál es la ganancia esperada apostando repetidamente a primera columna?
- ¿Cuál es la probabilidad de que un servidor que atiende a 20 terminales se sature en un determinado momento?
- Dada la información disponible, ¿cuál es la probabilidad de que llueva el próximo fin de semana?

Definiciones:

Experimento: Es cualquier proceso o acción que genera observaciones y que puede ser repetible. Por ejemplo, arrojar un dado, seleccionar un individuo y registrar su peso y su altura, seleccionar una muestra de productos elaborados por una empresa para hacer un control de calidad, seleccionar un día al azar y registrar el número de veces que se satura un servidor.

Espacio muestral asociado a un experimento: Es el conjunto de todos los resultados posibles del experimento. Lo notaremos S .

Ejemplos:

- 1) Se arroja una moneda una vez.
 $S = \{\text{cara, ceca}\}$ ó $S = \{1, 0\}$ ó $S = \{\text{éxito, fracaso}\}$
- 2) Se arroja una moneda dos veces.
 $S = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$
- 3) Se arroja una moneda hasta que aparece por primera vez una cara.
 $S = \{(1), (0, 1), (0, 0, 1), (0, 0, 0, 1), \dots\} = \{(x_1, x_2, \dots, x_n) / n \in \mathbb{N}, x_i = 0 \text{ si } i < n, x_n = 1\}$
- 4) Se registra el tiempo transcurrido desde que se intenta la conexión a un servidor hasta que la conexión se efectiviza.

$$S = \mathcal{R}^+ = (0, \infty)$$

Si el sistema tiene un time-out en el tiempo t_0 , tendríamos $S = (0, t_0)$.

Como se observa, un espacio muestral puede ser *finito*, como en los ejemplos 1) y 2), *infinito numerable*, como en el ejemplo 3) o *infinito no numerable*, como en el ejemplo 4).

Sucesos o eventos: No sólo estamos interesados en resultados individuales de un experimento sino que pueden interesarnos colecciones o conjuntos de ellos. Se denomina suceso o evento a cualquier subconjunto del espacio muestral. Si S es finito o infinito numerable, cualquier subconjunto es un evento. Si S es infinito "casi todo" subconjunto de S es un evento. Los eventos los designaremos en general con las primeras letras del abecedario en mayúscula: A, B, C,...

Evento elemental o simple: consiste de un único resultado individual.

Evento compuesto: consiste de más de un evento elemental.

Ejemplos: En los ejemplos anteriores, posibles eventos son

- 1) $A = \text{"sale cara"} = \{\text{cara}\} = \{1\}$.
- 2) $A = \text{"número de caras es menor o igual que 1"} = \{(1,0), (0,1), (0,0)\}$.
- 3) $A = \text{"número de tiros requeridos es menor o igual que 5"} = \{(x_1, x_2, \dots, x_n) \in S \mid n \leq 5\}$.
 $B = \text{"número de tiros requeridos es par"} = \{(x_1, x_2, \dots, x_n) \in S \mid n = 2k, k \in \mathbb{N}\}$.
- 4) $A = \text{"el tiempo es mayor de 10 minutos"} = (10, \infty)$ (en el caso de un sistema sin time-out)

Relación con Teoría de conjuntos: Como un evento o suceso es un conjunto, valen las mismas relaciones que en teoría de conjuntos.

S es un subconjunto de S denominado *suceso cierto o seguro*.

\emptyset es un subconjunto de S denominado *suceso imposible*.

$A \cup B$ es el suceso *unión*. Ocurre cuando A ocurre ó B ocurre.

$A \cap B$ es el suceso *intersección*. Ocurre cuando ocurre A y ocurre B.

A^c ó \bar{A} es el *opuesto* o *complemento* de A. Ocurre cuando no ocurre A.

$A - B = A \cap B^c$ es el suceso *diferencia*. Ocurre cuando ocurre A y no ocurre B.

Se dice que A *está contenido* en B o que A implica B y se denota $A \subseteq B$ si la realización de A conduce a la realización de B, es decir si todo elemento de A pertenece a B.

Dos sucesos A y B se dicen *mutuamente excluyentes* o *disjuntos* si $A \cap B = \emptyset$.

Recordemos algunas propiedades:

$$\begin{aligned} \text{Asociatividad: } A \cup B \cup C &= (A \cup B) \cup C = A \cup (B \cup C) \\ A \cap B \cap C &= (A \cap B) \cap C = A \cap (B \cap C) \end{aligned}$$

$$\begin{aligned} \text{Conmutatividad: } A \cup B &= B \cup A \\ A \cap B &= B \cap A \end{aligned}$$

$$\begin{aligned} \text{Distributividad: } (A \cup B) \cap C &= (A \cap C) \cup (B \cap C) \\ (A \cap B) \cup C &= (A \cup C) \cap (B \cup C) \end{aligned}$$

$$\text{Leyes de De Morgan: } \left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c \quad \text{y} \quad \left(\bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

Interpretación intuitiva de la Probabilidad: Supongamos que se repite n veces un mismo experimento aleatorio en forma independiente y bajo las mismas condiciones. Sea n_A el número de veces que ocurre el suceso A en las n repeticiones. Se denomina *frecuencia relativa* de A en la secuencia de n repeticiones a

$$fr(A) = \frac{n_A}{n}$$

La evidencia empírica muestra que cuando n crece, $fr(A)$ tiende a estabilizarse alrededor de un número que llamaremos $P(A)$.

¿Qué propiedades tiene la frecuencia relativa?

$$1) \quad fr(A) = \frac{n_A}{n} \geq 0$$

$$2) \quad fr(S) = \frac{n_S}{n} = \frac{n}{n} = 1$$

$$3) \text{ Si } A \cap B = \emptyset \Rightarrow fr(A \cup B) = \frac{n_{A \cup B}}{n} = \frac{n_A + n_B}{n} = \frac{n_A}{n} + \frac{n_B}{n} = fr(A) + fr(B)$$

La definición axiomática de Probabilidad, que daremos a continuación, es consistente con la idea intuitiva que se tiene de ella.

Axiomas de Probabilidad: Dado un experimento aleatorio y un espacio muestral asociado S , a cada evento A se le asociará un número que notaremos $P(A)$ y que llamaremos probabilidad del evento A . Esta asignación debe satisfacer los siguientes axiomas:

A1. $P(A) \geq 0$ para todo evento A .

A2. $P(S) = 1$

A3a. Si A_1, A_2, \dots, A_n es una colección finita de sucesos mutuamente excluyentes, es decir que $A_i \cap A_j = \emptyset \forall i \neq j$, entonces

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

A3b. Si $A_1, A_2, \dots, A_n, \dots$ es una colección infinita numerable de sucesos mutuamente excluyentes, es decir si $A_i \cap A_j = \emptyset \forall i \neq j$, entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Ejemplo: Consideremos el ejemplo en que se arroja una moneda una vez, para el cual el espacio muestral es $S = \{\text{cara}, \text{ceca}\}$. Si denominamos $E_1 = \{\text{cara}\}$ y $E_2 = \{\text{ceca}\}$ a los dos eventos elementales, como $P(S) = 1 = P(E_1) + P(E_2)$, entonces $P(E_2) = 1 - P(E_1)$. Por lo tanto, cualquier asignación de probabilidades de la forma: $P(E_1) = p$ y $P(E_2) = 1 - p$ con $0 \leq p \leq 1$, satisface los axiomas.

Propiedades de la Probabilidad:

1) $P(A^c) = 1 - P(A)$ para todo suceso A

Dem: $1 = P(S) \underset{A2}{=} P(A \cup A^c) \underset{A3a}{=} P(A) + P(A^c) \Rightarrow P(A^c) = 1 - P(A)$

En la tercera igualdad usamos el axioma 3 pues $A \cap A^c = \emptyset$.

$$2) P(\emptyset) = 0$$

$$\text{Dem: } P(\emptyset) \underset{P1}{=} 1 - P(\emptyset^c) = 1 - P(S) \underset{A2}{=} 1 - 1 = 0$$

$$3) \text{ Si } A \subseteq B \Rightarrow P(A) \leq P(B) \text{ y } P(B - A) = P(B) - P(A)$$

Dem: Si $A \subseteq B \Rightarrow B = A \cup (B - A)$ y éstos dos eventos son excluyentes. Por el axioma A3a

$$P(B) = P(A) + P(B - A)$$

Dado que, por el axioma A1, $P(B - A) \geq 0$, resulta $P(B) \geq P(A)$ y, despejando, se obtiene la segunda afirmación.

$$4) \text{ Dados dos sucesos cualesquiera } A \text{ y } B, P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Dem: $A \cup B = A \cup (B - A) = A \cup (B \cap A^c)$ y estos dos eventos son excluyentes, entonces, por el axioma A3a,

$$P(A \cup B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \quad (1)$$

Por otra parte, $B = (B \cap A) \cup (B \cap A^c)$ y estos dos eventos son disjuntos, entonces

$$P(B) = P(B \cap A) + P(B \cap A^c) \Rightarrow P(B \cap A^c) = P(B) - P(B \cap A) \quad (2)$$

De (1) y (2) resulta que $P(A \cup B) = P(A) + P(B) - P(B \cap A)$ como queríamos demostrar.

$$5) \text{ Dados dos sucesos cualesquiera } A \text{ y } B, P(A \cup B) \leq P(A) + P(B).$$

Dem: Esta propiedad se deduce inmediatamente de la propiedad anterior y del axioma A1.

Ejercicios: a) Demostrar, usando la propiedad 4) que, dados tres sucesos cualesquiera, A_1, A_2 y A_3 ,

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\ - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

b) Probar, usando inducción que, dados A_1, A_2, \dots, A_n sucesos cualesquiera,

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

Asignación de probabilidades: Supongamos que el espacio muestral S asociado con cierto experimento es finito o infinito numerable. En este caso, una manera simple de trabajar es asignar probabilidades a los sucesos elementales, ya que cualquier suceso A será unión de sucesos elementales y éstos son obviamente mutuamente excluyentes.

Designando E_i a los sucesos elementales de S , $S = \bigcup_{i=1}^{\infty} E_i$ (la unión podría ser finita si el espacio muestral fuese finito). Si conocemos $p_i = P(E_i) \geq 0 \forall i$, de manera que

$\sum_{i=1}^{\infty} p_i = 1$, entonces dado cualquier suceso A , su probabilidad se puede obtener sumando las probabilidades de los elementales que lo componen, es decir:

$$P(A) = \sum_{E_i \subset A} p_i$$

Ejemplos: 1) Se arroja un dado equilibrado. En este caso, $S = \{1, 2, 3, 4, 5, 6\}$ y, por suponerlo equilibrado, los sucesos elementales $E_i = \{i\}$ para $i = 1, \dots, 6$ tienen probabilidad $p_i = 1/6$. Si deseamos calcular la probabilidad del suceso $A =$ "el resultado es par", usando que

$$A = E_2 \cup E_4 \cup E_6 \quad \text{se obtiene} \quad P(A) = P(E_2) + P(E_4) + P(E_6) = 1/2$$

2) Supongamos ahora que se arroja un dado en el cual la probabilidad de las caras pares es el doble que la probabilidad de las caras impares, o sea que, si llamamos p a la probabilidad de cada cara impar,

$$P(E_1) = P(E_3) = P(E_5) = p \quad \text{y} \quad P(E_2) = P(E_4) = P(E_6) = 2p$$

Como la suma de las probabilidades debe ser igual a 1,

$$\sum_{i=1}^6 P(E_i) = 3p + 6p = 9p = 1 \quad \Rightarrow \quad p = \frac{1}{9}$$

$$\text{y, en este caso, } P(A) = P(E_2) + P(E_4) + P(E_6) = 3 \frac{2}{9} = \frac{2}{3}.$$

3) Arrojamus una moneda equilibrada hasta obtener cara. ¿Cuál es la probabilidad de que la cara sea obtenida en un número par de lanzamientos?

Si representamos el espacio muestral tal como lo hicimos más arriba, tendríamos

$$A = \{(0,1), (0,0,0,1), (0,0,0,0,0,1), \dots\}$$

Veremos más adelante que en las condiciones de este experimento es razonable asumir que

$$P(\text{obtener cara en el } k\text{-ésimo lanzamiento}) = \left(\frac{1}{2}\right)^k$$

Por lo tanto:

$$P(A) = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^{2k} = \sum_{k=1}^{\infty} \left(\frac{1}{4}\right)^k = \frac{1}{1 - \frac{1}{4}} - 1 = \frac{4}{3} - 1 = \frac{1}{3}$$

ya que si $0 < p < 1$, entonces

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$$

Espacios de equiprobabilidad: Sea un experimento aleatorio cuyo espacio muestral asociado S es finito y sea $n = \# S$ (el símbolo $\#$ representa el cardinal del conjunto). Diremos que el espacio es de equiprobabilidad si los n sucesos elementales tienen igual probabilidad, es decir si

$$P(E_i) = p \quad \forall i$$

$$\text{Como } 1 = P(S) = \sum_{i=1}^n P(E_i) = \sum_{i=1}^n p = np \quad \Rightarrow \quad p = \frac{1}{n} = \frac{1}{\#S}.$$

$$\text{Dado cualquier suceso } A, \quad P(A) = \sum_{E_i \subset A} P(E_i) = \sum_{E_i \subset A} \frac{1}{n} = \frac{\#A}{\#S}.$$

Ejemplos: 1) De una urna que contiene 2 bolillas blancas y 3 rojas se extraen 2 bolillas con reposición.

- ¿Cuál es la probabilidad de que se extraiga al menos una bolilla roja?
- ¿Cuál es la probabilidad de que la primera bolilla extraída sea roja y la segunda blanca?

Supondremos que las bolillas están numeradas, de manera de poder considerar que se trata de un espacio de equiprobabilidad, entonces $S = \{(x_1, x_2) / x_i \in \{R_1, R_2, R_3, B_1, B_2\}\}$ y su cardinal es $\#S = 5 \cdot 5 = 25$

a) $P(A) = 1 - P(A^c)$ siendo $A^c = \{(x_1, x_2) \in S / x_i \in \{B_1, B_2\}\}$. Como $\#A^c = 2 \cdot 2 = 4$, resulta $P(A^c) = \frac{4}{25} \Rightarrow P(A) = 1 - \frac{4}{25} = \frac{21}{25}$.

b) $B = \{(x_1, x_2) \in S / x_1 \in \{R_1, R_2, R_3\}, x_2 \in \{B_1, B_2\}\}$. Como $\#B = 3 \cdot 2 = 6 \Rightarrow P(B) = \frac{6}{25}$.

2) Consideremos el ejemplo 1) pero suponiendo ahora que las extracciones se realizan *sin reposición*.

En este caso, $S = \{(x_1, x_2) / x_i \in \{R_1, R_2, R_3, B_1, B_2\}, x_1 \neq x_2\} \Rightarrow \#S = 5 \cdot 4 = 20$.

a) $P(A) = 1 - P(A^c)$ siendo $A^c = \{(x_1, x_2) \in S / x_i \in \{B_1, B_2\}\}$. Como $\#A^c = 2 \cdot 1 = 2$, resulta $P(A^c) = \frac{2}{20} = \frac{1}{10} \Rightarrow P(A) = 1 - \frac{1}{10} = \frac{9}{10}$.

b) $B = \{(x_1, x_2) \in S / x_1 \in \{R_1, R_2, R_3\}, x_2 \in \{B_1, B_2\}\}$. Como $\#B = 3 \cdot 2 = 6 \Rightarrow P(B) = \frac{6}{20}$.

Observación: ¿Qué pasaría si en los ejemplos anteriores eligiésemos como espacio muestral $S = \{(B, B), (B, R), (R, B), (R, R)\}$, denotando B : bolilla blanca y R : bolilla roja? ¿Sería razonable suponer equiprobabilidad?

Probabilidad condicional

Consideremos una urna que contiene 4 bolillas rojas y 5 blancas. De las 4 bolillas rojas, 2 son lisas y 2 rayadas y de las 5 bolillas blancas, 4 son lisas y una sola es rayada. Supongamos que se extrae una bolilla y, sin que la hayamos mirado, alguien nos dice que la bolilla es roja, ¿cuál es la probabilidad de que la bolilla sea rayada?

Sean los sucesos A: “la bolilla es rayada” y B: “la bolilla es roja”. Obviamente, sin ninguna información previa, $P(A) = 2/9 = 1/3$ y $P(B) = 4/9$.

Sin embargo, como sabemos que la bolilla es roja, la probabilidad de que sea rayada es $1/2$, ya que, de las rojas la mitad es lisa y la mitad rayada. Observemos, que al ocurrir B, el espacio muestral se *reduce*.

En general, dado un experimento y su espacio muestral asociado, queremos determinar cómo afecta a la probabilidad de A el hecho de saber que ha ocurrido otro evento B.

Definición: Sean A y B eventos tales que $P(B) > 0$, la probabilidad del evento A condicional a la ocurrencia del evento B es

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ejemplos: 1) En el ejemplo anterior, $P(B) = 4/9$ y

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/9}{4/9} = \frac{2}{4} = \frac{1}{2}.$$

2) Consideremos una población en la que cada individuo es clasificado según dos criterios: es o no portador de HIV y pertenece o no a cierto grupo de riesgo que denominaremos R. La correspondiente tabla de probabilidades es:

	Portador (A)	No portador (A ^c)	
Pertenece a R (B)	0.003	0.017	0.020
No pertenece a R (B ^c)	0.003	0.977	0.980
	0.006	0.994	1.000

En esta población, la probabilidad de que un individuo sea portador es $P(A) = 0.006$ y la probabilidad de que sea portador y pertenezca al grupo de riesgo R es $P(A \cap B) = 0.003$.

Dado que una persona seleccionada al azar pertenece al grupo de riesgo R, ¿cuál es la probabilidad de que sea portador?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.003}{0.020} = 0.150$$

es decir que 150 de cada 1000 individuos del grupo de riesgo R, son “probablemente” portadores de HIV.

Calculemos ahora la probabilidad de que una persona sea portadora de HIV, dado que no pertenece al grupo de riesgo R.

$$P(A | B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{0.003}{0.980} = 0.00306$$

es decir que sólo 3 de cada 1000 individuos no pertenecientes al grupo de riesgo R, son “posibles” portadores de HIV.

Propiedades de la Probabilidad condicional: Dado un suceso B fijo tal que $P(B) > 0$, $P(\bullet|B)$ es una probabilidad, en el sentido que satisface los axiomas de probabilidad y por lo tanto todas las propiedades que se deducen a partir de ellos. Por ejemplo:

A1. $P(A|B) \geq 0$ para todo suceso A.

A2. $P(S|B) = 1$.

$$\text{Dem: } P(S | B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Ejercicios: 1) Verificar que $P(\bullet|B)$ satisface el axioma **A3a**.

2) Verificar que $P((A_1 \cup A_2) | B) = P(A_1 | B) + P(A_2 | B) - P((A_1 \cap A_2) | B)$

Regla del producto: Dados dos sucesos A y B, tales que $P(B) > 0$,

$$P(A \cap B) = P(A|B) P(B)$$

Si además, $P(A) > 0$,

$$P(A \cap B) = P(B | A) P(A)$$

Ejemplo: En el ejemplo presentado al comienzo, supongamos ahora que se extraen dos bolillas *sin reposición*. ¿Cuál es la probabilidad de extraer una bolilla roja y una blanca, en ese orden?

Sean C: “la primera bolilla es roja” y D: “la segunda bolilla es blanca”. debemos calcular $P(C \cap D)$. Aplicando la regla del producto

$$P(C \cap D) = P(C) P(D|C) = \frac{4}{9} \frac{5}{8} = \frac{20}{72} = \frac{5}{18}.$$

La regla del producto es especialmente útil cuando el experimento consta de varias etapas ya que se puede generalizar. Así, por ejemplo, si $P(A_1) > 0$ y $P(A_1 \cap A_2) > 0$, se tiene

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2 | A_1) P(A_3 | (A_1 \cap A_2))$$

y se extiende a n sucesos.

Ejemplo: En el mismo ejemplo, ¿cómo podemos obtener la probabilidad de que la segunda bolilla extraída sea blanca (suceso D)?. Sabemos calcular, usando la regla del producto la probabilidad de que la segunda sea blanca y la primera sea roja. Hemos visto que esta probabilidad es $P(C \cap D) = 5/18$. Del mismo modo podemos obtener la probabilidad de que ambas bolillas sean blancas (suceso $(D \cap C^c)$). Esta probabilidad es

$$P(C^c \cap D) = P(C^c) P(D|C^c) = \frac{5}{9} \frac{4}{8} = \frac{20}{72} = \frac{5}{18}.$$

Si ahora observamos que el suceso D puede escribirse como

$$D = (D \cap C) \cup (D \cap C^c)$$

se obtiene

$$P(D) = P(D \cap C) + P(D \cap C^c) = \frac{5}{18} + \frac{5}{18} = \frac{5}{9}. \quad (1)$$

¿Cómo podemos obtener ahora la probabilidad de que la primera bolilla haya sido roja (suceso C) sabiendo que la segunda fue blanca (suceso D)? La probabilidad requerida es

$$P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{5/18}{5/9} = \frac{1}{2}. \quad (2)$$

Los resultados (1) y (2) son ejemplos de aplicación de los dos Teoremas que veremos a continuación: el Teorema de la Probabilidad Total y el Teorema de Bayes, respectivamente.

Definición: Una colección de eventos A_1, A_2, \dots, A_k constituye una *partición* del espacio muestral S si

1. $A_i \cap A_j = \emptyset \quad \forall i \neq j$

2. $P(A_i) > 0 \quad \forall i$

3. $\bigcup_{i=1}^k A_i = S$

Teorema de la probabilidad total: Sea A_1, A_2, \dots, A_k una partición del espacio muestral S y sea B un suceso cualquiera,

$$P(B) = \sum_{i=1}^k P(B | A_i) P(A_i)$$

Dem:

$$B = B \cap S = B \cap \left(\bigcup_{i=1}^k A_i \right) = \bigcup_{i=1}^k (B \cap A_i)$$

Como $(B \cap A_i) \cap (B \cap A_j) = \emptyset \quad \forall i \neq j$, entonces

$$P(B) = P\left(\bigcup_{i=1}^k (B \cap A_i) \right) = \sum_{i=1}^k P(B \cap A_i) = \sum_{i=1}^k P(B | A_i) P(A_i).$$

Teorema de Bayes: Sea A_1, A_2, \dots, A_k una partición del espacio muestral S y sea B un suceso cualquiera tal que $P(B) > 0$,

$$P(A_j | B) = \frac{P(B | A_j) P(A_j)}{\sum_{i=1}^k P(B | A_i) P(A_i)}$$

Dem:

$$P(A_j | B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B | A_j) P(A_j)}{\sum_{i=1}^k P(B | A_i) P(A_i)}$$

En el numerador se aplicó la regla del producto y en el denominador el Teorema de la probabilidad total.

El Teorema de Bayes describe cómo es posible “revisar” la probabilidad inicial de un evento o probabilidad a priori ($P(A_i)$) para reflejar la información adicional que nos provee la ocurrencia de un evento relacionado. La probabilidad revisada se denomina probabilidad a posteriori.

Ejemplo: Supongamos que cierta prueba para detectar la presencia de una enfermedad en un individuo, da resultado positivo (detecta la presencia de la enfermedad) en un individuo enfermo con probabilidad 0.99 y en un individuo sano con probabilidad 0.02 (*falso positivo*). Por lo tanto, dicha prueba no detecta la enfermedad en un individuo sano con probabilidad 0.98 y no la detecta en un individuo enfermo con probabilidad 0.01 (*falso negativo*). Es decir que si denotamos A: “la persona padece esa enfermedad” y B: “la prueba es positiva”,

$$P(B | A) = 0.99 \quad P(B | A^c) = 0.02$$

$$P(B^c | A) = 0.01 \quad P(B^c | A^c) = 0.98$$

Se supone, en base a estudios previos, que la incidencia de esa enfermedad en cierta población es 0.001, es decir que la probabilidad a priori de A es 0.001. Se selecciona al azar un individuo de esa población, se le aplica la prueba y el resultado es positivo, ¿cuál es la probabilidad de que en efecto padezca la enfermedad?

Debemos calcular la probabilidad a posteriori de A, $P(A|B)$:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | A^c)P(A^c)} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} = 0.0472$$

Por lo tanto, la probabilidad de que esté enfermo, habiendo sido positivo el resultado de la prueba es aproximadamente 0.05.

Las probabilidades a posteriori dependen fuertemente de las probabilidades a priori. Si se aplica la prueba a individuos de una población en la cual la incidencia de la enfermedad es mucho mayor, también aumentará la probabilidad a posteriori.

Verifique ésto, suponiendo ahora que $P(A) = 0.01$.

Más adelante, desarrollaremos otro ejemplo de aplicación de estos Teoremas.

Independencia

La definición de probabilidad condicional nos permite “revisar” la probabilidad $P(A)$ asignada a un suceso, cuando se sabe que otro suceso B ha ocurrido. Hay casos en los que $P(A | B) \neq P(A)$, mientras que en otros $P(A | B) = P(A)$, es decir que la ocurrencia del suceso B no altera la probabilidad de ocurrencia de A.

Ejemplo: De una urna que contiene 4 bolillas negras y 6 blancas se extraen dos bolillas *sin reposición*, ¿cuál es la probabilidad de que la segunda bolilla sea blanca, sabiendo que la primera es negra?

Denominando A : “la segunda bolilla es blanca” y B : “la primera bolilla es negra”,

$$P(A|B) = \frac{6}{9} = \frac{2}{3}.$$

Por otra parte,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = \frac{6}{9} \frac{4}{10} + \frac{5}{9} \frac{6}{10} = \frac{54}{90} = \frac{6}{10} = \frac{3}{5}$$

y, por lo tanto, $P(A|B) \neq P(A)$, es decir que la ocurrencia del suceso B modifica la probabilidad del suceso A .

Observemos que la probabilidad de que la segunda bolilla sea blanca coincide con la probabilidad de que la primera lo sea.

Ejercicio: Verificar que, en cambio, si las extracciones se realizan con reposición, $P(A) = P(A|B)$.

Diremos que los eventos A y B son *independientes* si la información acerca de la ocurrencia o no de uno de ellos no afecta la probabilidad de ocurrencia del otro,

Definición: Los eventos A y B son independientes si

$$P(A \cap B) = P(A)P(B)$$

Si la igualdad no se cumple, decimos que A y B son dependientes.

Proposición: Supongamos $P(B) > 0$, A y B son independientes si y sólo si $P(A|B) = P(A)$.

Dem: (\Rightarrow) Si $P(B) > 0 \Rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}$ está bien definida, pero por ser A y B independientes, $P(A \cap B) = P(A)P(B)$, entonces

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

(\Leftarrow) Aplicando la regla del producto, si $P(B) > 0$, $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$.

Observación: Si $P(B) = 0$, como $A \cap B \subseteq B$, $P(A \cap B) = 0$, y por lo tanto la igualdad $P(A \cap B) = P(A)P(B)$ siempre se satisface.

Ejemplo: De un mazo de 40 cartas españolas, se extrae una carta al azar. Consideremos los siguientes sucesos:

A: "la carta es copa o espada"

B: "la carta no es copa"

C: "la carta es copa u oro"

$$P(A) = \frac{20}{40} = \frac{1}{2} \quad P(B) = \frac{30}{40} = \frac{3}{4} \quad P(C) = \frac{20}{40} = \frac{1}{2}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3} \neq P(A), \text{ entonces } A \text{ y } B \text{ no son independientes.}$$

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{1/4}{1/2} = \frac{1}{2} = P(A), \text{ entonces } A \text{ y } C \text{ son independientes.}$$

Propiedades: 1) Si los sucesos A y B son excluyentes, es decir si $A \cap B = \emptyset$ y si $P(A) > 0$, $P(B) > 0$, entonces A y B no son independientes.

Dem: En efecto, en este caso, $0 = P(A \cap B) \neq P(A)P(B)$.

2) Si $P(B) = 0$, entonces B es independiente de cualquier suceso A tal que $P(A) > 0$.

Dem: Como $A \cap B \subseteq B$, $P(A \cap B) = 0$ y por lo tanto $P(A \cap B) = P(A)P(B)$, es decir que A y B son independientes.

3) Si $A \subseteq B$, $P(A) > 0$ y $P(B) < 1$, A y B no son independientes.

Dem: Como $A \subseteq B \Rightarrow A \cap B = A \Rightarrow P(A \cap B) = P(A) \neq P(A)P(B)$. Luego, A y B no son independientes.

4) Si A y B son sucesos independientes, A y B^c también lo son.

Dem:

$$P(A) = P(A \cap B) + P(A \cap B^c) \Rightarrow P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) =$$

$$P(A)(1 - P(B)) = P(A)P(B^c).$$

Ejercicio: Demostrar que si A y B son sucesos independientes, A^c y B^c también lo son.

Independencia de más de dos eventos: La definición de independencia de dos eventos puede extenderse a más de dos.

Definición: Los eventos A_1, A_2, \dots, A_n son independientes si para todo $k = 2, \dots, n$ y para todo conjunto de índices $\{i_1, i_2, \dots, i_k\}$ tales que $1 \leq i_1 < i_2 < \dots < i_k \leq n$, se verifica

$$P(A_{i_1} \cap A_{i_2} \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \dots P(A_{i_k})$$

Es decir que es necesario verificar $\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - n - 1$ condiciones.

Observación: Si los sucesos A_1, A_2, \dots, A_n son independientes, entonces son independientes de a pares pero la recíproca no es cierta.

Ejemplos: 1) Sea $S = \{w_1, w_2, w_3, w_4\}$ un espacio de equiprobabilidad y consideremos $n = 3$ y los sucesos

$$A = \{w_1, w_4\}$$

$$B = \{w_2, w_4\}$$

$$C = \{w_3, w_4\}$$

$$P(A) = P(B) = P(C) = \frac{1}{2}.$$

Además,

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$

$$P(A \cap C) = \frac{1}{4} = P(A)P(C)$$

$$P(B \cap C) = \frac{1}{4} = P(B)P(C)$$

es decir, que los sucesos son independientes de a pares. Sin embargo,

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C)$$

y, por lo tanto, los sucesos A, B y C no son independientes.

2) Veamos un ejemplo también para el caso $n = 3$, en el cual se satisface la factorización de $P(A \cap B \cap C)$ y no se cumple para alguna de las intersecciones dobles. Sea $S = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8\}$ un espacio de equiprobabilidad y consideremos los sucesos

$$A = \{w_1, w_2, w_3, w_4\}$$

$$B = \{w_1, w_2, w_7, w_8\}$$

$$C = \{w_1, w_5, w_6, w_7\}$$

Como antes, $P(A) = P(B) = P(C) = \frac{1}{2}$. Además,

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$

$$P(B \cap C) = \frac{1}{4} = P(B)P(C)$$

$$P(A \cap C) = \frac{1}{8} \neq P(A)P(C)$$

Se observa que no se satisface una de las igualdades, pero sí se satisface

$$P(A \cap B \cap C) = \frac{1}{8} = P(A)P(B)P(C).$$

Finalmente, veremos un ejemplo en el que utilizamos los diferentes conceptos y propiedades estudiadas en esta Sección.

Ejemplo: Muchos sistemas de computación trabajan con enormes bases de datos, como por ejemplo, sistemas de tarjetas de crédito o sistemas de reservas de pasajes aéreos. Debido al volumen de datos involucrado, la velocidad de acceso al sistema depende de las características de las unidades de almacenamiento utilizadas, como así también de las redes de comunicación conectadas a la base de datos. Nos concentraremos en el primer aspecto, es decir en el problema del almacenamiento.

Consideremos unidades de almacenamiento consistentes en discos planos, cada uno de los cuáles está compuesto por un conjunto de anillos concéntricos denominados “pistas”. Cada pista está a su vez subdivida en áreas de almacenamiento denominadas “sectores”. El acceso al disco se realiza mediante una cabeza lectora/grabadora que se puede mover hacia adelante o hacia atrás a lo largo de un brazo fijo. El disco rota bajo ese brazo y la cabeza lee o modifica un dato cuando el correspondiente sector pasa bajo ella. Consideremos un disco que consiste de 76 pistas, numeradas de 0 a 75, con 8 sectores cada una, numerados de 0 a 7.

Supongamos que, en el momento en que se debe acceder a un dato que se encuentra en el sector 2 de la pista 51, la cabeza se encuentra sobre la pista 22. Entonces, debe moverse en primer lugar hasta la pista 51 (este movimiento se llama búsqueda o *seek*) y luego debe esperar hasta que el sector 2 pase bajo ella (este período de tiempo se denomina retardo rotacional o *rotational delay*).

Si el cabezal se mueve por ejemplo a una velocidad de 3.2 milisegundos (ms) por pista, la búsqueda del ejemplo demandaría $(3.2)(51-22) = (3.2)(29) = 92.8$ ms. Si además suponemos que el disco realiza una rotación completa en 30 ms, el retardo rotacional

puede demorar entre 0 y 30 ms, con un promedio de 15 ms. Por último, supongamos que el acceso concreto al dato demora 1.2 milisegundos.

Este sistema es de naturaleza probabilística o aleatoria. Las demandas de acceso arriban en tiempos aleatorios y se demandan datos aleatorios, en el sentido de que no sabemos con anticipación qué dato se va a requerir.

Analicemos el siguiente ejemplo. Supongamos que las probabilidades de que una demanda de acceso corresponda a cada una de las 76 pistas son iguales y que accesos sucesivos son independientes. Supongamos también que la cabeza lectora/grabadora se encuentra sobre la pista 20, ¿cuál es la probabilidad de que el tiempo total de búsqueda (seek) para las dos siguientes demandas de acceso sea a lo sumo 50 ms?

Sea A el suceso "la búsqueda combinada demora a lo sumo 50 ms" y definamos, para cada $i = 0, 1, \dots, 75$, los sucesos

T_i : "el primero de los dos accesos siguientes corresponderá a un dato que está sobre la pista i "

Entonces

$$P(A) = \sum_{i=0}^{75} P(A \cap T_i) = \sum_{i=0}^{75} P(A|T_i)P(T_i) \quad (3)$$

Como se observa, debemos calcular $P(A|T_i)$, es decir debemos calcular la probabilidad de que la búsqueda combinada demore a lo sumo 50 ms dado que el primer acceso es a la pista i , para $i = 0, 1, \dots, 75$. Por ejemplo, ¿cómo calcularíamos $P(A|T_{26})$?

Si la primera búsqueda nos lleva a la pista 26, demandará $(26-20)(3.2)$ ms = 19.2 ms, por lo tanto la búsqueda total llevará a lo sumo 50 ms si la segunda búsqueda demora a lo sumo 30.8 ms. Como en 30.8 ms se pueden recorrer a lo sumo 9 pistas $(30.8/3.2)$, no podemos ir más allá de la pista $26-9=17$ o de la pista $26+9=35$. En otras palabras $P(A|T_{26})$ será la probabilidad de que el segundo pedido de acceso se refiera a un dato que está entre las pistas 17 y 35 inclusive. Dado que suponemos que todas las pistas son equiprobables,

$$P(A|T_{26}) = \frac{19}{76} = \frac{1}{4}.$$

Del mismo modo, se calculan todas las probabilidades condicionales requeridas en (3) y se obtiene el valor de $P(A)$ pedido.

Variabes aleatorias discretas

Al realizar un experimento generalmente estamos interesados en alguna función del resultado más que en el resultado en sí mismo. Así, por ejemplo, al arrojar un dado dos veces podríamos estar interesados sólo en la suma de los puntos obtenidos y no en el par de valores que dio origen a ese valor de la suma. Esa cantidad de interés, o más formalmente esa función a valores reales definida sobre el espacio muestral se denomina *variable aleatoria*. *Variable* porque toma distintos valores y *aleatoria* porque el valor observado no puede ser predicho antes de la realización del experimento, aunque sí se sabe cuáles son sus posibles valores.

Dado que el valor de una variable aleatoria (en adelante lo abreviaremos v.a.) es determinado por el resultado de un experimento, podremos asignar probabilidades a los posibles valores o conjuntos de valores de la variable.

Ejemplo: Se arroja dos veces un dado equilibrado. Un espacio muestral asociado es:

$$S = \{(x_1, x_2) / x_i \in \{1,2,3,4,5,6\}\}$$

Posibles v.a. asociadas con este experimento son:

X: "número de caras pares"

Y: "máximo puntaje"

Z: "suma de puntos"

Definición: Sea S un espacio muestral asociado con un experimento aleatorio. Una variable aleatoria X es una función que asocia a cada elemento $w \in S$ un número real $X(w)=x$, es decir

$$X : S \rightarrow \mathfrak{R}$$

Como se observa, en general representaremos a las v.a. con letras mayúsculas: X , Y , Z , etc. y sus valores con letras minúsculas, es decir $X(w)=x$ significa que x es el número real asociado al resultado $w \in S$ a través de X .

Ejemplos: 1) Volviendo al ejemplo anterior,

$X((2,5)) = 1$	$Y((2,5)) = 5$	$Z((2,5)) = 7$
$X((1,3)) = 0$	$Y((1,3)) = 3$	$Z((1,3)) = 4$
$X((2,2)) = 2$	$Y((2,2)) = 2$	$Z((2,2)) = 4$

2) Se arroja una moneda equilibrada 3 veces,

$$X = \begin{cases} 1 & \text{si el número de caras es impar} \\ 0 & \text{en caso contrario} \end{cases}$$

3) Se arroja una moneda equilibrada hasta que se obtiene la primera cara,

X: “número de tiros necesarios”

4) A partir del instante en que se intenta la conexión a un servidor, se registra el tiempo que demora en concretarse la misma,

X: “tiempo requerido para la conexión”.

En los ejemplos 1), 2) y 3) las v.a. toman un número finito o infinito numerable de valores, mientras que en el ejemplo 4) la v.a. X toma valores en un conjunto infinito no numerable, el intervalo $(0, \infty)$ o un intervalo $(0, M)$ si existe un tiempo máximo (“time out”).

Notación: Indicaremos con R_X el rango de la v.a. X, es decir el conjunto de valores posibles de la v.a. X.

Ejemplos: En los ejemplos anteriores,

$$\begin{aligned} 1) \quad R_X &= \{0, 1, 2\} \\ R_Y &= \{1, 2, 3, 4, 5, 6\} \\ R_Z &= \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} \end{aligned}$$

$$2) \quad R_X = \{0, 1\}$$

$$3) \quad R_X = \{1, 2, 3, \dots\} = \mathbb{N}$$

$$4) \quad R_X = (0, \infty) \text{ ó } (0, M) \text{ si existe un “time out”}$$

Definición: Una v.a. es *discreta* si toma un número finito o infinito numerable de valores.

Ejemplo: En el caso del ejemplo 1), ¿cómo calcularíamos la probabilidad de que la v.a. Z tome el valor 7, suponiendo que los lanzamientos son independientes?

$$P(Z = 7) = P(\{(x_1, x_2) \in S / Z((x_1, x_2)) = 7\}) = P(\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}) = \frac{6}{36} = \frac{1}{6}.$$

Definición: La **función de probabilidad puntual o de masa** de la v.a. discreta X, se define para todo x como

$$p_X(x) = P(X = x) = P(\{w \in S / X(w) = x\})$$

Se cumplen las siguientes propiedades:

$$p_X(x) \geq 0 \quad \forall x$$

$$\sum_{x \in R_X} p_X(x) = 1$$

La función de probabilidad puntual de una v.a. X nos dice cómo se distribuye la probabilidad total entre los distintos valores de X , y se determina a partir de la probabilidad de los sucesos asociados a cada valor de X .

Ejemplos: 1) Hallemos la función de probabilidad puntual de la v.a. X : “número de caras pares al arrojar dos veces un dado equilibrado”. Recordemos que $R_X = \{0, 1, 2\}$.

$$p_X(0) = P(X = 0) = P(\{(x_1, x_2) \in S / x_1, x_2 \in \{1, 3, 5\}\}) = \frac{9}{36} = \frac{1}{4}$$

$$p_X(1) = P(X = 1) =$$

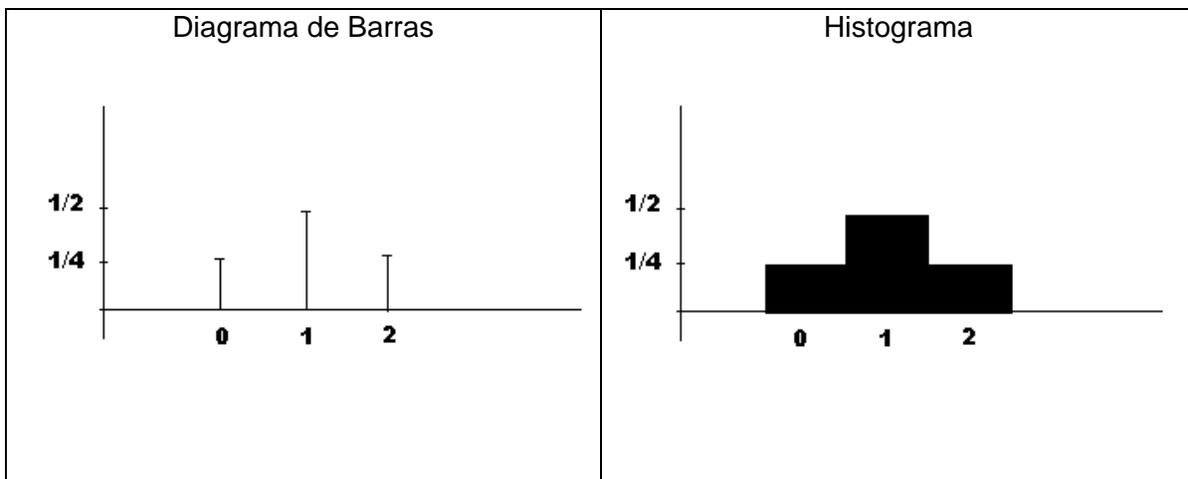
$$= P(\{(x_1, x_2) \in S / x_1 \in \{1, 3, 5\}, x_2 \in \{2, 4, 6\}\} \cup \{(x_1, x_2) \in S / x_1 \in \{2, 4, 6\}, x_2 \in \{1, 3, 5\}\}) = \frac{18}{36} = \frac{1}{2}$$

$$p_X(2) = P(X = 2) = P(\{(x_1, x_2) \in S / x_1, x_2 \in \{2, 4, 6\}\}) = \frac{9}{36} = \frac{1}{4}$$

Podemos resumir esta información en una tabla de la forma:

x	0	1	2
$p_X(x)$	1/4	1/2	1/4

o mediante un gráfico en el cual, para cada valor de x se construye una barra o un rectángulo centrado en x , cuya altura es proporcional a $p_X(x)$



Definición: La **función de distribución acumulada** de una v.a. discreta X con función de probabilidad puntual $p_X(x)$ se define para todo $x \in \mathfrak{R}$, como

$$F_X(x) = P(X \leq x) = \sum_{y \leq x, y \in R_X} p_X(y)$$

Es decir que $F_X(x)$ es la probabilidad de que la v.a. X tome valores menores o iguales que x .

Ejemplo: Volviendo al ejemplo 1), hallemos la función de distribución acumulada de la v.a. X , cuya función de probabilidad puntual es

x	0	1	2
$p_X(x)$	1/4	1/2	1/4

Si $x < 0$ $F_X(x) = P(X \leq x) = 0$

$x = 0$ $F_X(0) = P(X \leq 0) = p_X(0) = \frac{1}{4}$

$0 < x < 1$ $F_X(x) = P(X \leq x) = p_X(0) = \frac{1}{4}$

$x = 1$ $F_X(1) = P(X \leq 1) = p_X(0) + p_X(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$

$1 < x < 2$ $F_X(x) = P(X \leq x) = p_X(0) + p_X(1) = \frac{3}{4}$

$x = 2$ $F_X(2) = P(X \leq 2) = p_X(0) + p_X(1) + p_X(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

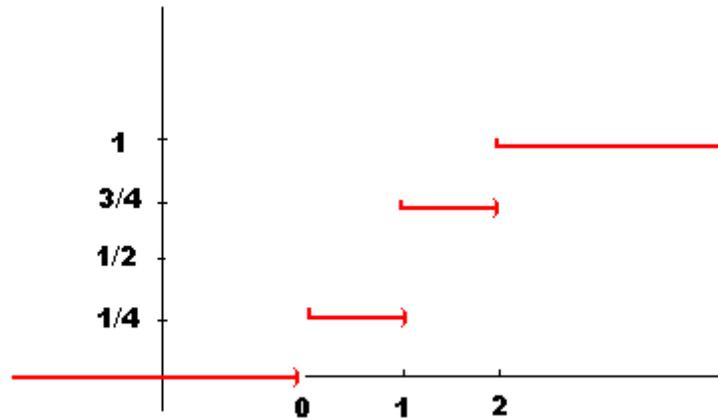
$x > 2$ $F_X(x) = P(X \leq 2) = p_X(0) + p_X(1) + p_X(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

Resumiendo:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{4} & \text{si } 0 \leq x < 1 \\ \frac{3}{4} & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

¿Cómo es $F_X(x)$?

Observamos que se trata de una función escalera, no decreciente que toma valores entre 0 y 1.



Propiedades de la función de distribución acumulada:

- i) $\forall x \in \mathfrak{R}, F_X(x) \in [0,1]$.
- ii) $F_X(x)$ es monótona no decreciente, es decir que si $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$.
- iii) $F_X(x)$ es continua a derecha, es decir $\lim_{h \rightarrow 0^+} F_X(x+h) = F_X(x)$.
- iv) $\lim_{x \rightarrow \infty} F_X(x) = 1$ y $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- v) En cada punto x , el valor del salto es la probabilidad puntual, es decir

$$p_X(x) = F_X(x) - F_X(x^-)$$

donde $x^- = \lim_{h \rightarrow 0^+} (x-h)$ (límite por la izquierda). En particular si X toma valores $x_1 < x_2 < \dots$, entonces $p_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ para todo $i \geq 2$ y $p_X(x_1) = F_X(x_1)$.

Dem: Daremos sólo demostraciones heurísticas de estas propiedades. Demostraciones rigurosas pueden encontrarse, por ejemplo, en S. Ross (1988) o B. James (1981).

i) Obvio, ya que $F_X(x) = P(X \leq x) = P(\{w \in S / X(w) \leq x\})$ y toda probabilidad toma valores entre 0 y 1.

ii) Consideremos el suceso

$$A = \{w / X(w) \leq x_2\} = \{w / X(w) \leq x_1\} \cup \{w / x_1 < X(w) \leq x_2\} = A_1 \cup A_2$$

Como $A_1 \cap A_2 = \emptyset$, $P(A) = P(A_1) + P(A_2)$, es decir

$$P(X \leq x_2) = P(X \leq x_1) + P(x_1 < X \leq x_2) \geq P(X \leq x_1)$$

y, por lo tanto,

$$F_X(x_2) \geq F_X(x_1)$$

iii) Recordemos que una función $g(x)$ es continua a derecha en x si $\lim_{h \rightarrow 0^+} g(x+h) = g(x)$.

Por lo tanto, la continuidad a derecha de $F_X(x)$ en todo x resulta de su definición:

$$F_X(x) = P(X \leq x).$$

$$\text{iv) } \lim_{x \rightarrow \infty} F_X(x) = \lim_{x \rightarrow \infty} P(X \leq x) = \lim_{x \rightarrow \infty} P\{w / X(w) \leq x\} = P(S) = 1$$

$$\lim_{x \rightarrow -\infty} F_X(x) = \lim_{x \rightarrow -\infty} P(X \leq x) = \lim_{x \rightarrow -\infty} P(w / X(w) \leq x) = P(\emptyset) = 0$$

$$\text{v) } p_X(x) = P(X = x) = P(X \leq x) - P(X < x) = F_X(x) - F_X(x^-)$$

Proposición: Sean a y b tales que $a \leq b$, entonces

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a^-)$$

$$P(a < X < b) = F_X(b^-) - F_X(a)$$

$$P(a \leq X < b) = F_X(b^-) - F_X(a^-)$$

Dem: Demostremos la primera igualdad

$$\begin{aligned} P(a < X \leq b) &= P(X \in (a, b]) = P(X \in (-\infty, b]) - P(X \in (-\infty, a]) \\ &= P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) \end{aligned}$$

Ejercicio: Demostrar las siguientes 3 igualdades, usando por ejemplo que

$$P(a \leq X \leq b) = P(a < X \leq b) + P(X = a)$$

y aplicando la propiedad v) de las funciones de distribución acumuladas.

Ejemplo: Volviendo al ejemplo 1), y usando la función de distribución calculada antes, calculemos $P(1 \leq X \leq 2)$ y $P(X = 1)$.

$$P(1 \leq X \leq 2) = F_X(2) - F_X(1^-) = 1 - \frac{1}{4} = \frac{3}{4}$$

$$P(X = 1) = F_X(1) - F_X(1^-) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}.$$

Ejemplo: Un experimento tiene sólo dos resultados posibles, que denominaremos Éxito y Fracaso. El experimento se repite en forma independiente hasta que se obtiene el primer éxito. Sea $p = P(\text{Éxito})$, $0 < p < 1$, y definamos la v.a. $X = \text{“número de repeticiones hasta obtener el primer éxito”}$. Como ya hemos visto, $R_X = N$.

Hallemos la función de probabilidad puntual de la v.a. X .

$$\begin{aligned} p_X(1) &= p \\ p_X(2) &= (1-p)p \\ p_X(3) &= (1-p)^2 p \\ &\dots \\ p_X(k) &= (1-p)^{k-1} p \\ &\dots \end{aligned}$$

Entonces,

$$p_X(k) = (1-p)^{k-1} p \quad \forall k \in N.$$

Verifiquemos que en efecto esta función satisface las dos propiedades

$$p_X(x) \geq 0 \quad \forall x$$

$$\sum_{x \in R_X} p_X(x) = 1$$

Dado que $0 < p < 1$, la primer propiedad obviamente se satisface. Respecto a la segunda,

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{j=0}^{\infty} (1-p)^j = p \frac{1}{1-(1-p)} = 1$$

donde hemos usado que la suma de la serie geométrica $\sum_{i=0}^{\infty} q^i = \frac{1}{1-q}$, si $|q| < 1$.

Hallemos la función de distribución acumulada de la v.a. X .

$$\begin{aligned}
 x < 1 & \quad F_X(x) = 0 \\
 1 \leq x < 2 & \quad F_X(x) = p \\
 2 \leq x < 3 & \quad F_X(x) = p + p(1-p) \\
 3 \leq x < 4 & \quad F_X(x) = p + p(1-p) + p(1-p)^2 \\
 & \dots \\
 k \leq x < k+1 & \quad F_X(x) = p \sum_{j=1}^k (1-p)^{j-1} = p \sum_{i=0}^{k-1} (1-p)^i = p \frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k \\
 & \dots
 \end{aligned}$$

Hemos usado que la suma parcial de una serie geométrica es $\sum_{i=0}^n q^i = \frac{1-q^{n+1}}{1-q}$.

Recordemos que la función de distribución debe estar definida para todo $x \in \mathfrak{R}$, entonces

$$F_X(x) = \begin{cases} 0 & \text{si } x < 1 \\ 1 - (1-p)^{[x]} & \text{si } x \geq 1 \end{cases}$$

donde $[a]$ denota la parte entera de a .

Ejercicio: Verificar que esta función de distribución satisface las propiedades enunciadas antes.

Parámetro de una función de probabilidad: En el ejemplo anterior la probabilidad de Éxito la designamos p donde $0 < p < 1$. Variando este valor obtenemos diferentes funciones de probabilidad que constituyen lo que se denomina una *familia* de distribuciones. El valor p se denomina *parámetro* de la distribución.

En el caso del ejemplo, la familia obtenida se denomina Geométrica de parámetro p y diremos que $X \sim G(p)$. Por ejemplo, si el experimento hubiese consistido en arrojar un dado equilibrado hasta obtener el primer as, $X \sim G(1/6)$ y si hubiese consistido en arrojar una moneda equilibrada hasta obtener la primera cara, $X \sim G(1/2)$.

Esperanza o valor esperado de una v.a. discreta:

Una empresa proveedora de servicio de Televisión Satelital tiene 20000 clientes en cierta zona, cada uno de los cuáles puede optar por contratar de 1 a 5 paquetes de señales (el abono básico consiste en un solo paquete y cada uno de los otros paquetes incluye grupos de señales temáticas o premium). Supongamos que, entre los 20000 clientes, la distribución del número de paquetes X contratados es la siguiente:

x	1	2	3	4	5
número de clientes	7500	5500	3500	2000	1500
proporción	37.5%	27.5%	17.5%	10.0%	7.5%

Si interesa el número promedio de paquetes contratados, o sea el valor promedio de X en la población, deberíamos calcular:

$$\frac{1 \cdot 7500 + 2 \cdot 5500 + 3 \cdot 3500 + 4 \cdot 2000 + 5 \cdot 1500}{20000} = \frac{44500}{20000} = 2.225$$

Observemos que, si no hubiésemos conocido los números de clientes que contratan cada número de paquetes ni el total de la población, sino sólo las proporciones de cada número (o su probabilidad) hubiésemos podido obtener el valor promedio, ya que dicho número puede escribirse en la forma:

$$\begin{aligned} 1 \cdot \frac{7500}{20000} + 2 \cdot \frac{5500}{20000} + 3 \cdot \frac{3500}{20000} + 4 \cdot \frac{2000}{20000} + 5 \cdot \frac{1500}{20000} = \\ = 1 \cdot 0.375 + 2 \cdot 0.275 + 3 \cdot 0.175 + 4 \cdot 0.10 + 5 \cdot 0.075 \end{aligned}$$

Esto motiva la siguiente definición.

Definición: Sea X una v.a. discreta que toma valores en R_X con función de probabilidad puntual $p_X(x)$, la **esperanza o valor esperado de X** se define como

$$E(X) = \mu_X = \sum_{x \in R_X} x p_X(x)$$

siempre que $\sum_{x \in R_X} |x| p_X(x) < \infty$. Si la serie de los valores absolutos diverge, la esperanza no puede definirse y decimos que no existe.

Ejemplos: 1) Sea X : "número de caras pares al arrojar dos veces un dado equilibrado". Como

x	0	1	2
$p_X(x)$	1/4	1/2	1/4

entonces,
$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

2) Sea X una v.a. que toma sólo dos valores que designaremos 1 y 0 (Éxito y Fracaso) con la siguiente función de probabilidad puntual

x	1	0
$p_X(x)$	α	$1-\alpha$

siendo $0 < \alpha < 1$. Una v.a. de este tipo se dice que es una v.a. de tipo *Bernoulli* y su esperanza es:

$$E(X) = 1 \cdot \alpha + 0 \cdot (1 - \alpha) = \alpha$$

3) Veamos un ejemplo en que no existe $E(X)$. Sea X una v.a. con la siguiente función de probabilidad puntual

$$p_X(x) = \begin{cases} \frac{6}{\pi^2} \frac{1}{x^2} & \text{si } x \in \mathbb{N} \\ 0 & \text{en otro caso} \end{cases}$$

En primer lugar, observemos que $p_X(x)$ es una función de probabilidad puntual, ya que

$$\sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}$$

y, por lo tanto la suma de las probabilidades es 1. Calculemos la esperanza de X ,

$$E(X) = \sum_{x=1}^{\infty} x \frac{6}{\pi^2} \frac{1}{x^2} = \frac{6}{\pi^2} \sum_{x=1}^{\infty} \frac{1}{x} = \infty$$

4) Consideremos nuevamente un experimento que tiene sólo dos resultados posibles y que se repite en forma independiente hasta que se obtiene el primer éxito. Si $p = P(\text{Éxito})$, $0 < p < 1$, y si definimos la v.a. $X = \text{“número de repeticiones hasta obtener el primer éxito”}$, hemos demostrado que su función de probabilidad puntual está dada por

$$p_X(k) = (1 - p)^{k-1} p \quad \forall k \in \mathbb{N}$$

Calculemos la esperanza de X .

$$E(X) = \sum_{k=1}^{\infty} k p (1 - p)^{k-1} = p \sum_{k=1}^{\infty} k (1 - p)^{k-1} = -p \sum_{k=1}^{\infty} \frac{\partial}{\partial p} (1 - p)^k$$

Como la serie de potencias involucrada en la última igualdad es convergente, la derivada de la suma es la suma de las derivadas, entonces

$$E(X) = -p \frac{\partial}{\partial p} \left(\sum_{k=1}^{\infty} (1-p)^k \right) = -p \frac{\partial}{\partial p} \left(\frac{1}{1-(1-p)} - 1 \right) = -p \frac{\partial}{\partial p} \left(\frac{1}{p} - 1 \right) = -p \left(-\frac{1}{p^2} \right) = \frac{1}{p}.$$

y por lo tanto hemos demostrado que $E(X) = \frac{1}{p}$.

Interpretación de la esperanza: $E(X)$ es el centro de gravedad de la función de probabilidad puntual. Es decir que si imaginamos que sobre cada valor posible de X , x_i , colocamos una masa $p_X(x_i)$, el punto de equilibrio del sistema es $E(X)$. En este sentido, podemos decir que $E(X)$ es una medida del “centro” de la distribución.

Otra interpretación de $E(X)$ está relacionada con un resultado que estudiaremos más adelante, denominado “ley de los grandes números”. Imaginemos que se repite indefinidamente un experimento aleatorio y que en cada repetición nuestra v.a. X toma diferentes valores. Se ha demostrado que el promedio de los resultados obtenidos tiende a estabilizarse en un número que es $E(X)$, si es que ésta existe.

Esperanza de una función de una v.a. discreta: Volvamos al ejemplo considerado al comienzo del párrafo dedicado a la esperanza. Sea la v.a. X : número de paquetes de programas contratado por un cliente seleccionado al azar y consideremos su función de probabilidad puntual:

x	1	2	3	4	5
$p_X(x)$	0.375	0.275	0.175	0.100	0.075

Supongamos que el costo del servicio (Y) es función del número de paquetes contratado, según la siguiente fórmula:

$$Y = 30(X + 1)$$

¿Cuál es el valor esperado del costo pagado por cliente? Es decir, ¿cuál es $E(Y)$?

A partir de la función de probabilidad puntual de X , podemos obtener la de función de probabilidad de Y ya que, por un lado $R_Y = \{60, 90, 120, 150, 180\}$ y, por ejemplo, $P(Y=120)=P(X=3)=0.175$. Entonces,

y	60	90	120	150	180
$p_Y(y)$	0.375	0.275	0.175	0.100	0.075

y, $E(Y) = 60 \cdot 0.375 + 90 \cdot 0.275 + 120 \cdot 0.175 + 150 \cdot 0.10 + 180 \cdot 0.075 = 96.75$.

Observemos que, $E(Y) = \sum_{x=1}^5 h(x) p_X(x)$, siendo $h(x) = 30(x+1)$.

Proposición: Si X es discreta y toma valores x_1, x_2, \dots , entonces $h(X)$ es discreta con valores y_1, y_2, \dots , siendo $y_j = h(x_i)$ para al menos un valor de i .

Proposición: Si la v.a. X tiene función de probabilidad puntual $p_X(x)$ para todo $x \in R_X$, entonces la esperanza de cualquier función real $h(X)$, está dada por

$$E(h(X)) = \sum_{x \in R_X} h(x) p_X(x)$$

si la serie es absolutamente convergente, o sea si $\sum_{x \in R_X} |h(x)| p_X(x) < \infty$.

Dem: Sea $Y = h(X)$, entonces

$$E(Y) = \sum_j y_j p_Y(y_j) = \sum_j y_j \left[\sum_{i/h(x_i)=y_j} p_X(x_i) \right] = \sum_j \sum_{i/h(x_i)=y_j} y_j p_X(x_i) = \sum_i h(x_i) p_X(x_i).$$

Propiedades de la esperanza:

1) (*Linealidad*) Si a y b son constantes reales, $E(aX + b) = aE(X) + b$.

Dem: Sea $h(X) = aX + b$, entonces

$$E(h(X)) = E(aX + b) = \sum_{x \in R_X} (ax + b) p_X(x) = a \sum_{x \in R_X} x p_X(x) + b \sum_{x \in R_X} p_X(x) = aE(X) + b.$$

2) Si X es una v.a. tal que $P(X=c)=1$, entonces $E(X)=c$.

Dem: $E(X) = c p_X(c) = c$.

Varianza de una v.a. discreta:

Consideremos las siguientes funciones de probabilidad:

x	2	3	4
$p_X(x)$	1/3	1/3	1/3

y	1	2	3	4	5
$p_Y(y)$	1/12	5/12	2/12	1/12	3/12

z	3
$p_Z(z)$	1

Estas tres v.a. tienen la misma esperanza, sin embargo la forma de su distribución es muy diferente.

Ejercicio: Graficar las tres funciones de probabilidad puntual y verificar que $E(X)=E(Y)=E(Z)=3$.

Definiremos una medida de la variabilidad de una variable aleatoria alrededor de su esperanza.

Definición: Sea X una v.a. discreta con función de probabilidad puntual $p_X(x)$ y esperanza μ_X , la **varianza de X** , que se denotará $V(X)$, σ_X^2 ó σ^2 , es

$$V(X) = \sigma_X^2 = \sum_{x \in R_X} (x - \mu_X)^2 p_X(x) = E[(X - \mu_X)^2].$$

y el **desvío standard de X** , es $\sigma_X = +\sqrt{V(X)}$.

Ejemplos: 1) Calculemos la varianza y el desvío standard de las tres v.a. que acabamos de presentar, cuya esperanza es igual a 3.

$$V(X) = \sigma_X^2 = (2-3)^2 \frac{1}{3} + (3-3)^2 \frac{1}{3} + (4-3)^2 \frac{1}{3} = \frac{2}{3}$$

$$V(Y) = \sigma_Y^2 = (1-3)^2 \frac{1}{12} + (2-3)^2 \frac{5}{12} + (3-3)^2 \frac{2}{12} + (4-3)^2 \frac{1}{12} + (5-3)^2 \frac{3}{12} = \frac{22}{12} = \frac{11}{6}$$

$$V(Z) = \sigma_Z^2 = (3-3)^2 \cdot 1 = 0$$

2) Consideremos X : “número de caras pares al arrojar dos veces un dado equilibrado” cuya función de probabilidad puntual es

x	0	1	2
$p_X(x)$	1/4	1/2	1/4

y su esperanza es $E(X) = 1$, entonces

$$V(X) = (0-1)^2 \frac{1}{4} + (1-1)^2 \frac{1}{2} + (2-1)^2 \frac{1}{4} = \frac{1}{2}.$$

3) Sea X una v.a. Bernoulli con función de probabilidad puntual

x	1	0
$p_X(x)$	α	$1-\alpha$

con $0 < \alpha < 1$. Recordemos que $E(X) = \alpha$, entonces

$$V(X) = (1 - \alpha)^2 \alpha + (0 - \alpha)^2 (1 - \alpha) = \alpha(1 - \alpha) [(1 - \alpha) + \alpha] = \alpha(1 - \alpha).$$

Proposición: $V(X) = E(X^2) - (E(X))^2$.

Dem:

$$\begin{aligned} V(X) &= E((X - \mu_X)^2) = \sum_{x \in R_X} (x - \mu_X)^2 p_X(x) = \sum_{x \in R_X} (x^2 - 2\mu_X x + \mu_X^2) p_X(x) = \\ &= \sum_{x \in R_X} x^2 p_X(x) - 2\mu_X \sum_{x \in R_X} x p_X(x) + \mu_X^2 \sum_{x \in R_X} p_X(x) = E(X^2) - 2\mu_X E(X) + \mu_X^2 = \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2 = E(X^2) - (E(X))^2. \end{aligned}$$

Ejemplo: Consideremos nuevamente un experimento que tiene sólo dos resultados posibles y que se repite en forma independiente hasta que se obtiene el primer éxito. Si $p = P(\text{Éxito})$, $0 < p < 1$, hemos definido la v.a. $X = \text{“número de repeticiones hasta obtener el primer éxito”}$, cuya función de probabilidad puntual está dada por:

$$p_X(k) = (1 - p)^{k-1} p \quad \forall k \in \mathbb{N}$$

Hemos demostrado que $E(X) = \frac{1}{p}$. Demostraremos ahora que $V(X) = \frac{1-p}{p^2}$.

Calculemos $E(X^2)$.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty} [(k+1)k - k] p(1-p)^{k-1} = \\ &= \sum_{k=1}^{\infty} (k+1)kp(1-p)^{k-1} - \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \sum_{k=1}^{\infty} (k+1)kp(1-p)^{k-1} - E(X) = \\ &= p \left[\sum_{k=1}^{\infty} (k+1)k(1-p)^{k-1} \right] - \frac{1}{p} = p \left[\sum_{k=1}^{\infty} \frac{\partial^2}{\partial p^2} (1-p)^{k+1} \right] - \frac{1}{p} = \end{aligned}$$

$$\begin{aligned}
 &= p \frac{\partial^2}{\partial p^2} \left[\sum_{k=1}^{\infty} (1-p)^{k+1} \right] - \frac{1}{p} = p \frac{\partial^2}{\partial p^2} \left[\sum_{j=2}^{\infty} (1-p)^j \right] - \frac{1}{p} = \\
 &= p \frac{\partial^2}{\partial p^2} \left[\frac{1}{1-(1-p)} - 1 - (1-p) \right] - \frac{1}{p} = p \frac{\partial^2}{\partial p^2} \left[\frac{1}{p} - 2 + p \right] - \frac{1}{p} = \\
 &= p \frac{\partial}{\partial p} \left[-\frac{1}{p^2} + 1 \right] - \frac{1}{p} = p \frac{2}{p^3} - \frac{1}{p} = \frac{2}{p^2} - \frac{1}{p}
 \end{aligned}$$

Entonces,

$$V(X) = E(X^2) - (E(X))^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{(1-p)}{p^2}$$

como queríamos demostrar.

Propiedades de la varianza y del desvío standard:

1) $V(aX + b) = a^2V(X)$ y $\sigma_{aX+b} = |a|\sigma_X$.

Dem: Observemos que, en general,

$$V(h(X)) = \sum_{x \in R_X} (h(x) - E(h(X)))^2 p_X(x).$$

Entonces,

$$\begin{aligned}
 V(aX + b) &= \sum_{x \in R_X} (ax + b - E(aX + b))^2 p_X(x) = \sum_{x \in R_X} (ax + b - aE(X) - b)^2 p_X(x) = \\
 &= \sum_{x \in R_X} (ax - aE(X))^2 p_X(x) = a^2 \sum_{x \in R_X} (x - E(X))^2 p_X(x) = a^2V(X)
 \end{aligned}$$

y, por lo tanto, $\sigma_{aX+b} = |a|\sigma_X$.

En particular, observemos que $\sigma_{aX}^2 = a^2 \sigma_X^2$ y $\sigma_{X+b}^2 = \sigma_X^2$, y por lo tanto un cambio de escala afecta la varianza pero una traslación no la afecta.

2) Si X es una v.a. tal que $P(X=c) = 1$, entonces $V(X) = 0$.

Dem: Ejercicio.

VARIABLES ALEATORIAS DISCRETAS

Distribución Binomial:

Muchos experimentos aleatorios satisfacen las siguientes condiciones:

- El experimento consiste de n pruebas, siendo n fijo.
- Las pruebas son idénticas y en cada prueba hay sólo dos resultados posibles, que denominaremos Éxito (E) y Fracaso (F). Una prueba de este tipo se denomina **ensayo de Bernoulli**.
- Las pruebas son independientes, es decir que el resultado de una prueba no influye sobre el de las otras.
- La probabilidad de Éxito ($P(E)=p$) se mantiene constante en todas las pruebas.

Definición: Un experimento que satisface estos cuatro requerimientos se denomina **experimento Binomial**.

Ejemplos: 1) Se arroja una moneda n veces y se llama Éxito al suceso "sale cara".

2) Se arroja un dado equilibrado n veces y se llama Éxito al suceso "se obtiene un as".

3) Se arroja n veces un dardo a un blanco circular de radio R , el cuál contiene en el centro un círculo de radio $R/4$ y se denomina Éxito al suceso "el dardo impacta en el círculo central".

4) Se extraen 4 bolillas *con reposición* de una urna que contiene 5 bolillas blancas y 3 negras y se denomina Éxito al suceso "las 4 bolillas son blancas".

5) ¿Es el que sigue un experimento Binomial? Se extraen 2 bolillas *sin reposición* de una urna que contiene 5 bolillas blancas y 3 negras y se denomina Éxito al suceso "la bolilla extraída es blanca".

NO, no lo es ya que si denominamos B_i al suceso "la i -ésima bolilla extraída es blanca",

$$P(B_2 | B_1) = \frac{4}{7} \neq P(B_2) = \frac{5}{8}$$

y, por lo tanto no se verifica la tercera condición. En realidad tampoco se verifica la segunda ya que las pruebas no son idénticas (la composición de la urna varía). Observemos que, sin embargo la cuarta condición se satisface.

Variable aleatoria binomial: Consideremos un experimento binomial que consiste de n repeticiones y en el cual $P(E) = p$. Denominaremos v.a. binomial a la variable

X : número de éxitos en las n repeticiones.

Notación: $X \sim Bi(n, p)$.

Calculemos su función de probabilidad puntual. Para ello, observemos en primer lugar que $R_X = \{0, 1, 2, \dots, n\}$.

Sea $k \in R_X$, una secuencia posible con k éxitos y $n-k$ fracasos es:

$$\underbrace{E \dots E}_k \underbrace{F \dots F}_{n-k}$$

y su probabilidad, dada la independencia de las repeticiones, es $p^k (1-p)^{n-k}$. Pero, hay $\binom{n}{k}$ secuencias posibles conteniendo k éxitos, entonces

$$P(X = k) = p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k \in \{0, 1, \dots, n\}$$

Verifiquemos que $\sum_{k=0}^n p_X(k) = 1$. En efecto,

$$\sum_{k=0}^n p_X(k) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1.$$

Hemos usado la fórmula del Binomio de Newton: $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$.

Función de distribución: Si $X \sim \text{Bi}(n, p)$,

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k} & \text{si } 0 \leq x \leq n \\ 1 & \text{si } x > n \end{cases}$$

donde $\lfloor x \rfloor$ denota la parte entera de x .

Ejemplo: Supongamos que se arroja un dado equilibrado 10 veces y se llama Éxito al suceso "se obtiene un as". La v.a.

X : número de ases en los 10 tiros

tiene distribución Binomial de parámetros 10 y $1/6$, o sea $X \sim \text{Bi}(10, 1/6)$, entonces

$$P(X = 4) = \binom{10}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 = 0.054$$

$$P(3 \leq X \leq 5) = \sum_{k=3}^5 \binom{10}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{10-k} = F_X(5) - F_X(2) = 0.22$$

Esperanza y varianza de una variable aleatoria binomial: Sea $X \sim \text{Bi}(n, p)$,

$$E(X) = np \quad \text{y} \quad V(X) = np(1-p)$$

Dem: En el caso $n=1$, X es una v.a. Bernoulli y ya hemos demostrado que en este caso, $E(X)=p$ y $V(X) = p(1-p)$. Sea ahora $n>1$,

$$E(X) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} =$$

$$\sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} = np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} =$$

$$np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = np (p + (1-p))^{n-1} = np.$$

Recordemos que $V(X) = E(X^2) - (E(X))^2 = E(X^2) - n^2 p^2$.

$$\begin{aligned} E(X^2) &= \sum_{k=0}^n k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^n (k(k-1) + k) \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=2}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + E(X) \end{aligned}$$

$$= \sum_{k=2}^n k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} + np = \sum_{k=2}^n \frac{n!}{(k-2)!(n-k)!} p^k (1-p)^{n-k} + np$$

$$= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} (1-p)^{n-k} + np$$

$$\stackrel{(k-2)=j}{=} n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j (1-p)^{n-2-j} + np = n(n-1)p^2 (p + (1-p))^{n-2} + np$$

$$= n(n-1)p^2 + np$$

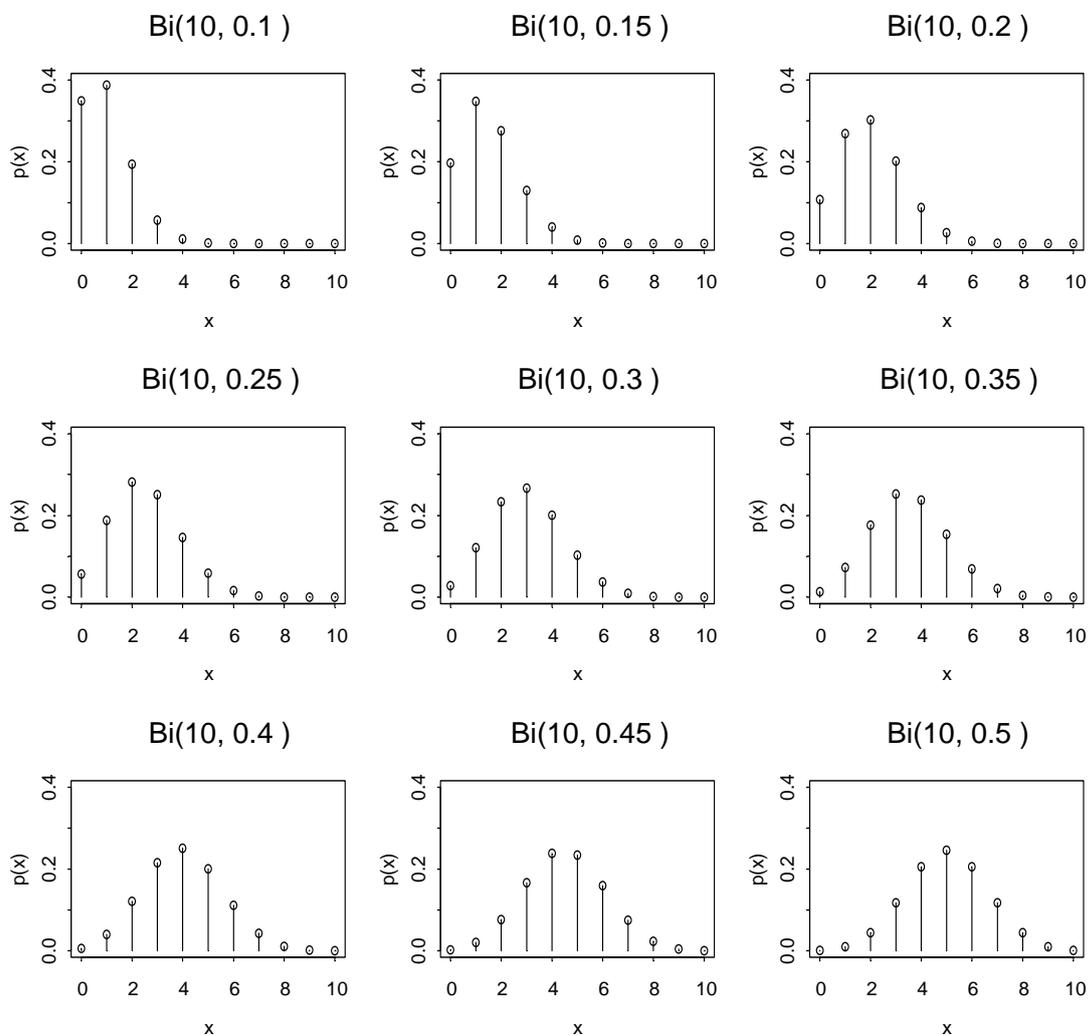
En realidad, para que la demostración anterior sea válida debe ser $n \geq 2$, pero es inmediato verificar que, si $n=1$, $E(X^2) = p$ y por lo tanto la expresión hallada es válida para todo n .

Finalmente,

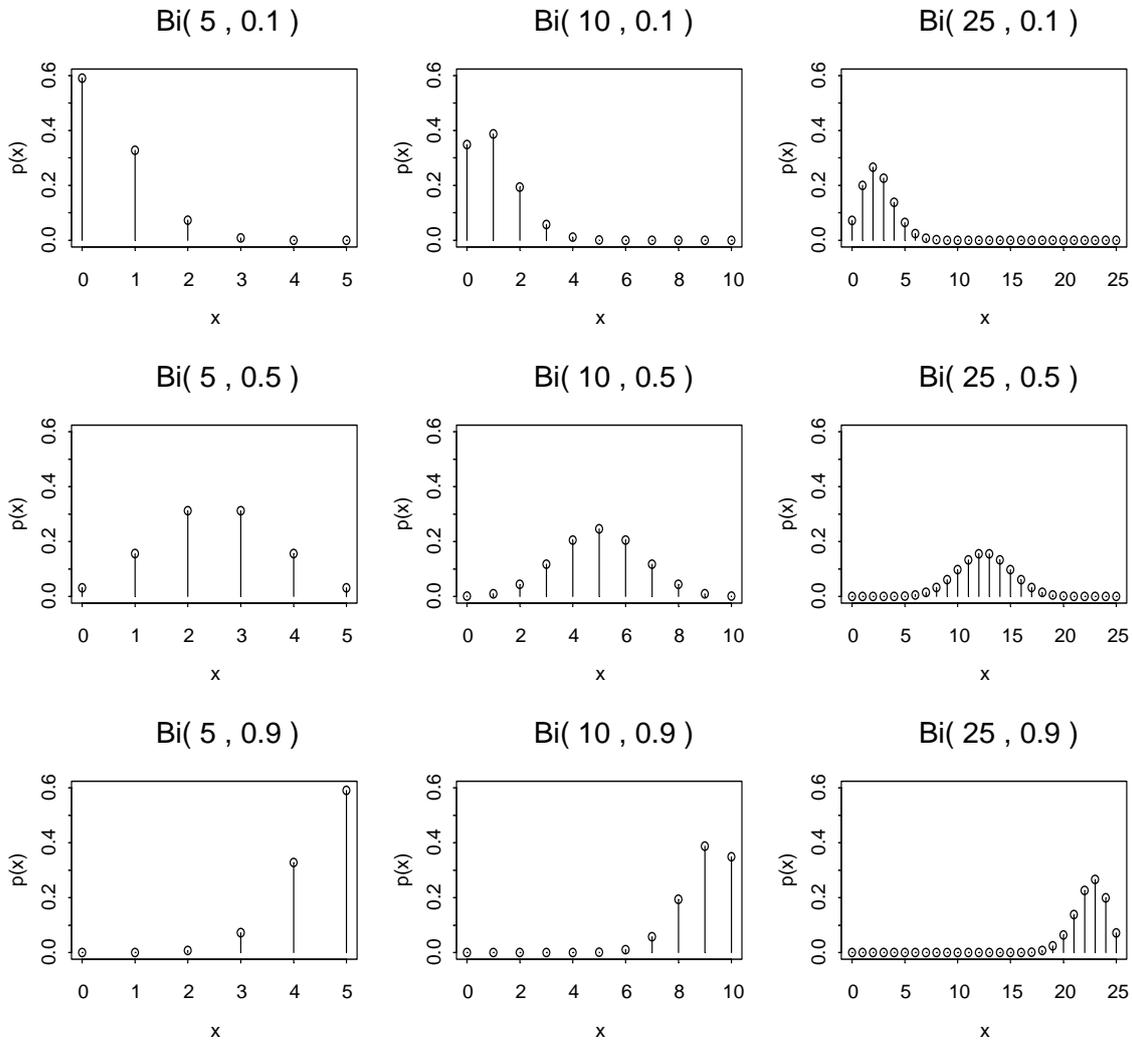
$$V(X) = E(X^2) - (E(X))^2 = n(n-1)p^2 + np - n^2 p^2 = -np^2 + np = np(1-p)$$

En el siguiente gráfico se muestra la función de probabilidad puntual correspondiente a la distribución Binomial para distintos valores de p y $n=10$. Puede observarse cómo la distribución se simetriza a medida que p tiende a 0.5.

¿Cómo serían los gráficos para valores de $p > 0.5$?



En el siguiente gráfico se muestra la función de probabilidad puntual correspondiente a la distribución Binomial para distintos valores de p y n .



Variable aleatoria Geométrica: Supongamos que se repite en forma independiente un **ensayo de Bernoulli** con probabilidad de Éxito ($P(E)=p$) constante en todas las pruebas. Se define la v.a.

X : número de repeticiones hasta obtener el primer Éxito.

Notación: $X \sim G(p)$.

Al estudiar en general las v.a. discretas, hemos probado que la función de probabilidad puntual de X está dada por

$$p_x(k) = (1-p)^{k-1} p \quad \forall k \in \mathbb{N}.$$

y su función de distribución acumulada por

$$F_x(x) = \begin{cases} 0 & \text{si } x < 1 \\ 1 - (1-p)^{[x]} & \text{si } x \geq 1 \end{cases}$$

donde $[x]$ denota la parte entera de x .

Esperanza y varianza de una variable aleatoria geométrica: Sea $X \sim G(p)$,

$$E(X) = \frac{1}{p} \quad \text{y} \quad V(X) = \frac{(1-p)}{p^2}$$

Dem: Lo hemos demostrado al estudiar en general la esperanza y la varianza de una v.a. discreta.

Proposición (Propiedad de Falta de Memoria): Sea $X \sim G(p)$ y sean n y m números naturales cualesquiera,

$$P(X > n + m | X > n) = P(X > m)$$

Dem: Ejercicio.

(Sugerencia: Demostrar que si $X \sim G(p)$, $P(X > k) = (1-p)^k$).

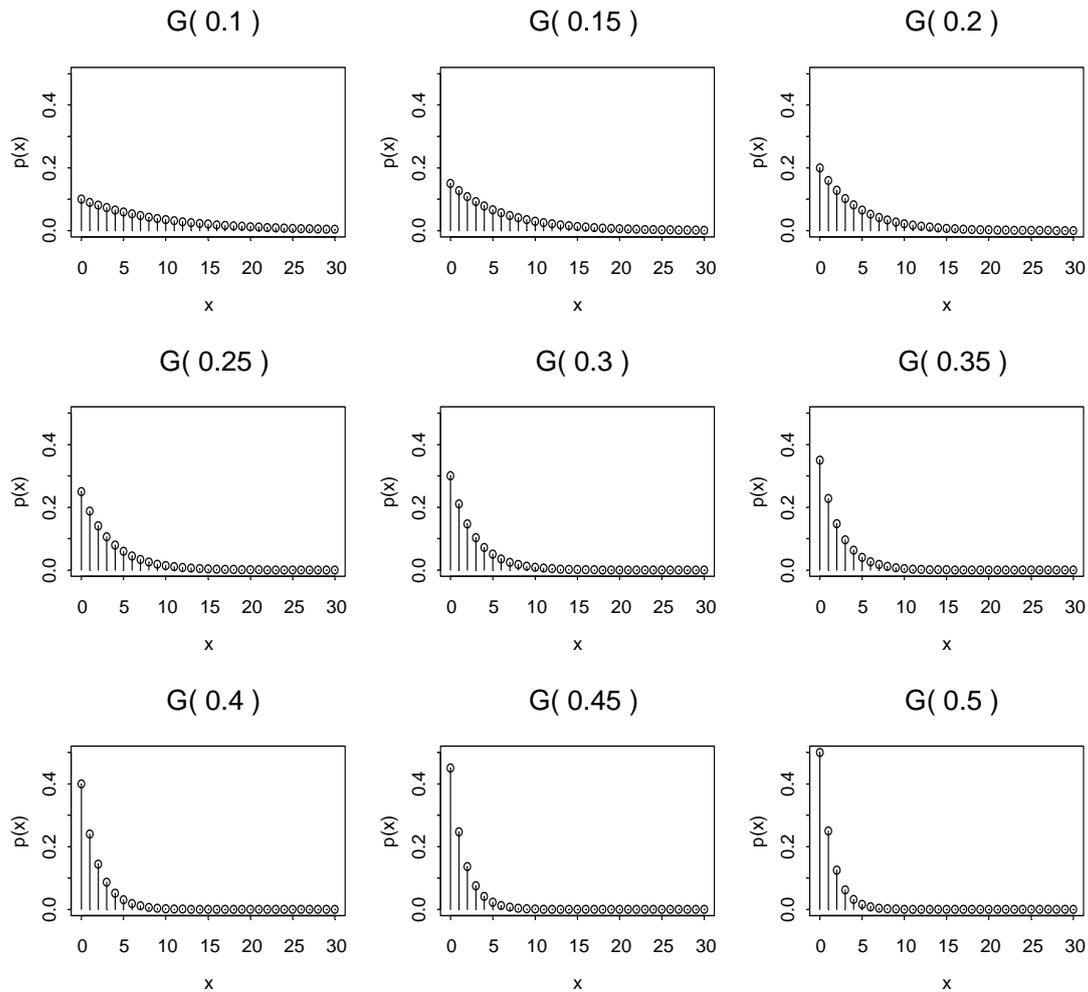
Ejemplo: Sea X : "número de tiros hasta obtener el primer as en una sucesión de tiros de un dado equilibrado", entonces $X \sim G(1/6)$.

$$P(X = 7) = \frac{1}{6} \left(\frac{5}{6} \right)^6 = 0.06$$

$$P(X \geq 6) = P(X > 5) = \left(\frac{5}{6} \right)^5 = 0.40$$

$$E(X) = \frac{1}{1/6} = 6 \quad V(X) = \frac{5/6}{(1/6)^2} = 30$$

En el siguiente gráfico se muestra la función de probabilidad puntual correspondiente a la distribución Geométrica para distintos valores de p .



Variable aleatoria Binomial Negativa: Supongamos que se repite en forma independiente un **ensayo de Bernoulli** con probabilidad de Éxito ($P(E)=p$) constante en todas las pruebas. Se define la v.a.

X : número de repeticiones hasta obtener el r -ésimo Éxito ($r \geq 1$).

Notación: $X \sim BN(r,p)$.

Esta v.a. es una generalización de la v.a. Geométrica, la cual corresponde al caso $r = 1$.

Observemos que $R_X = \{r, r+1, r+2, \dots\}$ y hallemos su función de probabilidad puntual.

Sea k un número natural, $k \geq r$. Para que sean necesarias k repeticiones para obtener el primer Éxito, el r -ésimo Éxito debe ocurrir en la repetición k y en las $(k-1)$ repeticiones previas debe haber exactamente $(r-1)$ Éxitos. Como las repeticiones son independientes la probabilidad de una configuración de ese tipo es $p^r(1-p)^{k-r}$, pero hay varias configuraciones de esta forma. ¿Cuántas? Tantas como formas de elegir entre las $(k-1)$ primeras repeticiones, aquellas donde ocurrirán los $(r-1)$ Éxitos, o sea $\binom{k-1}{r-1}$.

Por lo tanto la función de probabilidad puntual será:

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad \forall k \in \{r, r+1, r+2, \dots\}$$

Función de distribución: Si $X \sim BN(r,p)$,

$$F_X(x) = \begin{cases} 0 & \text{si } x < r \\ \sum_{k=r}^{\lfloor x \rfloor} \binom{k-1}{r-1} p^r (1-p)^{k-r} & \text{si } x \geq r \end{cases}$$

donde $\lfloor x \rfloor$ denota la parte entera de x .

Ejemplo: Se extraen con reposición bolillas de una urna que contiene 3 bolillas blancas y 7 rojas. Se define X : número de extracciones hasta obtener la cuarta bolilla roja.

$$X \sim BN(4, 7/10)$$

$$P(X = 5) = \binom{5-1}{4-1} \left(\frac{7}{10}\right)^4 \left(\frac{3}{10}\right) = 0.29$$

$$P(5 \leq X \leq 7) = \sum_{k=5}^7 \binom{k-1}{3} \left(\frac{7}{10}\right)^4 \left(\frac{3}{10}\right)^{k-4} = 0.49$$

Proposición: Sea $X \sim BN(r, p)$,

$$E(X) = \frac{r}{p} \quad V(X) = \frac{r(1-p)}{p^2}$$

Dem: Lo demostraremos más adelante usando que una v.a. Binomial Negativa puede expresarse como suma de v.a. Geométricas independientes.

Observación: Esta v.a. suele también definirse como el número de Fracazos antes de obtener el r -ésimo Éxito. Si la denotamos X , entonces su rango será

$$R_{X^*} = \{0, 1, 2, \dots\} = \mathbb{N} \cup \{0\}$$

y su función de probabilidad puntual: $p_{X^*}(x) = \binom{r+x-1}{x} p^r (1-p)^x$

En este caso,

$$E(X^*) = \frac{r(1-p)}{p} \quad \text{y} \quad V(X^*) = \frac{r(1-p)}{p^2}$$

Variable aleatoria Hipergeométrica: Supongamos que

- La población a ser muestreada consiste de N elementos o individuos (población finita)
- Cada elemento o individuo puede ser clasificado como Éxito o Fracaso y hay D Éxitos en la población.
- Se extrae de la población una *muestra* de n elementos o individuos, de forma tal que cualquier *subconjunto* de tamaño n tiene la misma probabilidad de ser elegido.

Sea X : número de éxitos en la muestra de tamaño n . Se dice que X tiene distribución Hipergeométrica de parámetros n , N y D y se denota

$$X \sim H(n, N, D)$$

Ejemplo: De una urna que contiene 3 bolillas blancas y 7 negras se extraen 4 bolillas *sin reposición* y se define X : número de bolillas blancas extraídas.

¿Cómo calcularíamos la probabilidad de que se extraigan 2 bolillas blancas ($X = 2$)?

Como todos los conjuntos de 4 bolillas tienen la misma probabilidad de ser extraídos, la probabilidad de uno cualquiera de ellos será $\frac{1}{\binom{10}{4}}$. Por otro lado hay $\binom{3}{2}\binom{7}{2}$ conjuntos

que contienen 2 bolillas blancas y 2 negras y, por lo tanto la probabilidad pedida será:

$$P(X = 2) = \frac{\binom{3}{2}\binom{7}{2}}{\binom{10}{4}} = \frac{3 \cdot 21}{210} = \frac{3}{10}.$$

Proposición: Si $X \sim H(n, N, D)$,

$$p_X(k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} \quad \max(0, n - (N - D)) \leq k \leq \min(n, D)$$

Dem: El número de subconjuntos distintos de tamaño n que se pueden extraer de una población de tamaño N es $\binom{N}{n}$. De esos conjuntos, hay $\binom{D}{k}\binom{N-D}{n-k}$ que contienen k Éxitos y $(n-k)$ Fracasos y se obtiene la función de probabilidad. El rango de valores posibles de k resulta de observar que se deben satisfacer tres condiciones:

$$0 \leq k \leq n \quad k \leq D \quad n - k \leq N - D$$

De las dos primeras se obtiene: $k \leq n, k \leq D \Leftrightarrow k \leq \min(n, D)$

De la primera y la tercera se obtiene: $k \geq 0, k \geq n - (N - D) \Leftrightarrow k \geq \max(0, n - (N - D))$.

Proposición: Si $X \sim H(n, N, D)$,

$$E(X) = n \frac{D}{N} \quad V(X) = \left(\frac{N-n}{N-1}\right) n \frac{D}{N} \left(1 - \frac{D}{N}\right)$$

Dem: Ejercicio opcional.

Observaciones: 1) El factor $\left(\frac{N-n}{N-1}\right)$ que aparece en la expresión de la varianza se denomina factor de corrección por población finita.

2) Si n es pequeño en relación a N , la hipergeométrica puede ser aproximada por la distribución Binomial de parámetros n y $p=D/N$. Observemos que, en este caso el factor de corrección finita es aproximadamente 1.

Límite de la función de probabilidad puntual de una v.a. Binomial:

Proposición: Sea $X \sim Bi(n,p)$ y supongamos que $n \rightarrow \infty$ y $p \rightarrow 0$, de manera que $n \cdot p = \lambda$ (fijo), entonces:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \longrightarrow \frac{e^{-\lambda} \lambda^k}{k!} \quad \forall k \in N_o = N \cup \{0\}$$

Dem:

$$\begin{aligned} p_X(k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \frac{\lambda^k}{k!} \\ &= \left[\frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n}\right] \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \frac{\lambda^k}{k!}. \end{aligned}$$

Observemos que:

$$1 \cdot \frac{n-1}{n} \dots \frac{n-k+1}{n} \xrightarrow{n \rightarrow \infty} 1$$

$$\left(1 - \frac{\lambda}{n}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}$$

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \xrightarrow{n \rightarrow \infty} 1$$

Entonces, $p_X(k) \longrightarrow \frac{e^{-\lambda} \lambda^k}{k!}$, como queríamos demostrar.

Esta proposición sugiere que la función de probabilidad puntual podría ser aproximada por la función de probabilidad límite, pero ¿cuándo se considera que n es grande y p es pequeño para que la aproximación sea buena?

Algunos autores sugieren $n \geq 100$, $p \leq 0.01$ y $np \leq 20$.

En la siguiente tabla se presentan a modo de ejemplo, algunos valores exactos de la probabilidad y su aproximación para el caso $X \sim Bi(100, 1/36)$

k	Prob. exacta (Binomial)	Aproximación
0	0.0598	0.0622
1	0.1708	0.1727
2	0.2416	0.2399
5	0.0857	0.0857
8	0.0049	0.0055
9	0.0014	0.0017
10	0.0004	0.0005

Como se observa, la aproximación es bastante buena, aún cuando no se cumple la condición $p \leq 0.01$.

Variable aleatoria Poisson: Una v.a. cuya función de probabilidad puntual es la obtenida en la proposición anterior, se dice que tiene distribución de Poisson de parámetro λ ($\lambda > 0$), y se nota $X \sim P(\lambda)$.

Es decir, $X \sim P(\lambda)$ si su función de probabilidad puntual está dada por:

$$p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \forall k \in N_o = N \cup \{0\}$$

Verifiquemos que es, en efecto, una función de probabilidad puntual:

Es obvio que $p_X(k) \geq 0 \quad \forall k$.

Por otra parte

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1,$$

ya que $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ es el desarrollo en serie de e^x .

Ejemplo: Sea X : "número de mensajes rechazados por segundo por un servidor", y supongamos que $X \sim P(5)$.

a) Calcular la probabilidad de que se rechacen exactamente 2 mensajes en un segundo.

$$P(X = 2) = \frac{e^{-5} 5^2}{2!} = 0.084$$

b) Calcular la probabilidad de que se rechacen a lo sumo 2 mensajes en un segundo.

$$P(X \leq 2) = \sum_{k=0}^2 \frac{e^{-5} 5^k}{k!} = e^{-5} \left(1 + 5 + \frac{5^2}{2} \right) = 0.125$$

Proposición: Si $X \sim P(\lambda)$, entonces

$$E(X) = \lambda \quad \text{y} \quad V(X) = \lambda$$

Dem:

$$E(X) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} = \lambda \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} = \lambda.$$

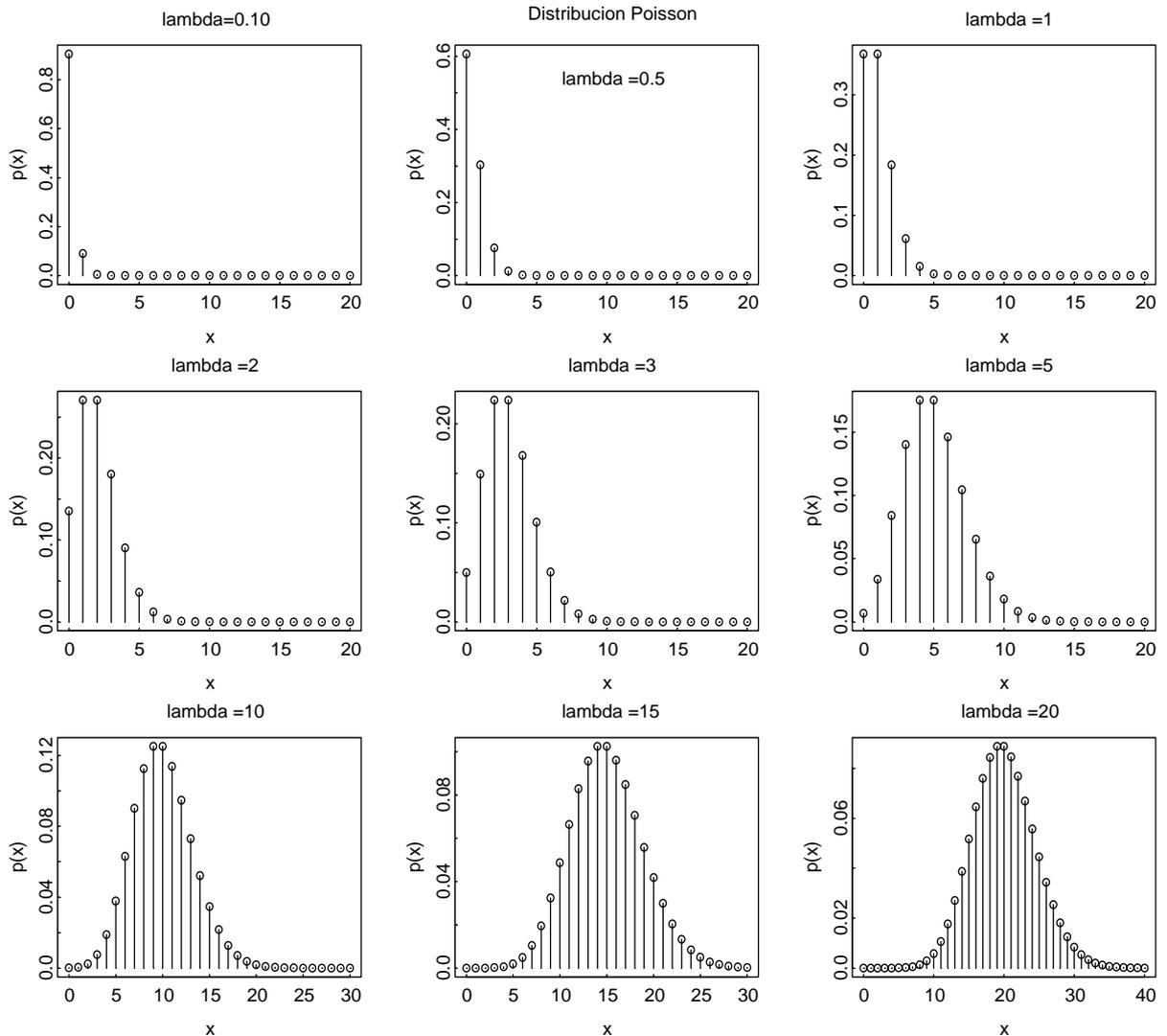
Por otra parte,

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=0}^{\infty} (k(k-1) + k) \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=2}^{\infty} k(k-1) \frac{e^{-\lambda} \lambda^k}{k!} + \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{e^{-\lambda} \lambda^{k-2}}{(k-2)!} + E(X) = \lambda^2 \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

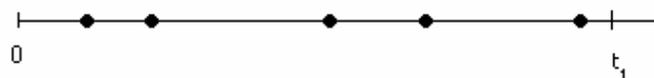
Entonces

$$V(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

En el siguiente gráfico se muestra la función de probabilidad puntual correspondiente a la distribución de Poisson para distintos valores de λ . En él puede observarse cómo la distribución se simetriza alrededor de λ a medida que este parámetro crece.



Proceso de Poisson: Una aplicación importante de la distribución de Poisson surge en relación con la ocurrencia de eventos a lo largo del tiempo, por unidad de área, por unidad de volumen, etc. En lo que sigue nos referiremos, sin pérdida de generalidad a ocurrencias de un evento a lo largo del tiempo, que podremos esquematizar en la forma:



A partir del instante 0 y hasta el momento t_1 ocurrieron 5 eventos.

Imaginemos que dividimos el intervalo $(0, t_1)$ en un número muy grande de pequeños subintervalos, de manera que se satisfacen las siguientes condiciones:

- La probabilidad de que ocurra un evento en un subintervalo pequeño es aproximadamente proporcional a la longitud del subintervalo.
- La probabilidad de que ocurra más de un evento en un subintervalo es despreciable con respecto a la probabilidad de que ocurra uno.
- La ocurrencia de un evento en un subintervalo es independiente de lo que ocurre en otro subintervalo disjunto.

En particular, si todos los intervalos son de igual longitud t_1/n , la v.a. X_{t_1} : “número de eventos que ocurren en el intervalo $(0, t_1)$ ” es “casi” binomial, siendo Éxito la ocurrencia de un evento en cada uno de los subintervalos y $p = P(\text{Éxito})$ =probabilidad de que ocurra un evento. Si el número de subintervalos es suficientemente grande y por lo tanto el p suficientemente pequeño, por el resultado límite que hemos probado, la variable X_{t_1} tiene distribución de Poisson.

Ejemplos: 1) Mensajes de correo electrónico que llegan a una casilla de correos.

- 2) Emisión de partículas por una sustancia radioactiva.
- 3) Accidentes que ocurren en un cruce de ruta.
- 4) Número de errores en una página de un libro.
- 5) Número de larvas de cierto insecto en un terreno.

Ejercicio: Para cada uno de estos ejemplos, discutir en que situaciones se verifican las tres condiciones enunciadas.

Definición: Supongamos que se observa la ocurrencia de un evento a lo largo del tiempo y que existe una cantidad positiva $\theta > 0$, tal que

- 1) La probabilidad de que ocurra exactamente un evento en un intervalo pequeño de longitud Δt es aproximadamente igual a $\theta \Delta t$, es decir:

$$P(\text{ocurra un evento en } \Delta t) = \theta \Delta t + o(\Delta t)$$

siendo $o(h)$ una función $g(h)$ tal que $\lim_{h \rightarrow 0} \frac{g(h)}{h} = 0$.

- 2) La probabilidad de que ocurra más de un evento en un intervalo pequeño de longitud Δt es despreciable cuando se la compara con la probabilidad de que ocurra un evento, es decir:

$$P(\text{ocurra más de un evento en } \Delta t) = o(\Delta t)$$

- 3) El número de eventos que ocurren en un intervalo es independiente del número de eventos que ocurren en otro intervalo disjunto.

Entonces, el número de ocurrencias del evento en un periodo de longitud t tiene distribución de Poisson de parámetro (θt) , es decir que la v.a. X_t : "número de ocurrencias del evento en el intervalo de longitud t " satisface

$$X_t \sim P(\theta t)$$

Observaciones: 1) ¿Cómo se interpreta la cantidad θ ?

Puede interpretarse como la tasa media a la cual ocurren los eventos en la unidad de tiempo. Se la suele llamar tasa media de ocurrencia o intensidad del Proceso de Poisson.

- 2) ¿Cuál es la diferencia entre un Proceso de Poisson y una v.a. con distribución Poisson?

La definición anterior, que en realidad es un teorema, da las condiciones bajo las cuáles ciertos experimentos aleatorios que producen como resultados eventos en el tiempo (o en longitud, área, volumen, etc) pueden ser modelados mediante la distribución de Poisson. Consideremos los ejemplos 1) a 5). Sólo bajo ciertas condiciones, satisfacen las propiedades de un Proceso de Poisson.

Ejemplo: Supongamos que el número de mensajes de correo electrónico que llegan a una casilla de correos sigue un proceso de Poisson de intensidad $\theta = 2$ mensajes / minuto.

- a) ¿Cuál es la probabilidad de que no se reciba ningún mensaje entre las 12 hs y las 12:03 hs?

Sea X_3 : "número de mensajes en un periodo de 3 minutos", $X_3 \sim P(2 \cdot 3) = P(6)$.

Entonces, $P(X_3=0) = e^{-6} = 0.002$

- b) ¿Cuál es el número esperado de mensajes en media hora?

Sea X_{30} : "número de mensajes en un periodo de 30 minutos"

$$X_{30} \sim P(2 \cdot 30) = P(60) \Rightarrow E(X_{30}) = 60$$

- c) ¿Cuál es la probabilidad de que no se reciba ningún mensaje entre las 13:30 hs y las 13:33 hs?

La respuesta es la misma del ítem a) porque la distribución depende sólo de la longitud del intervalo y no de su ubicación.

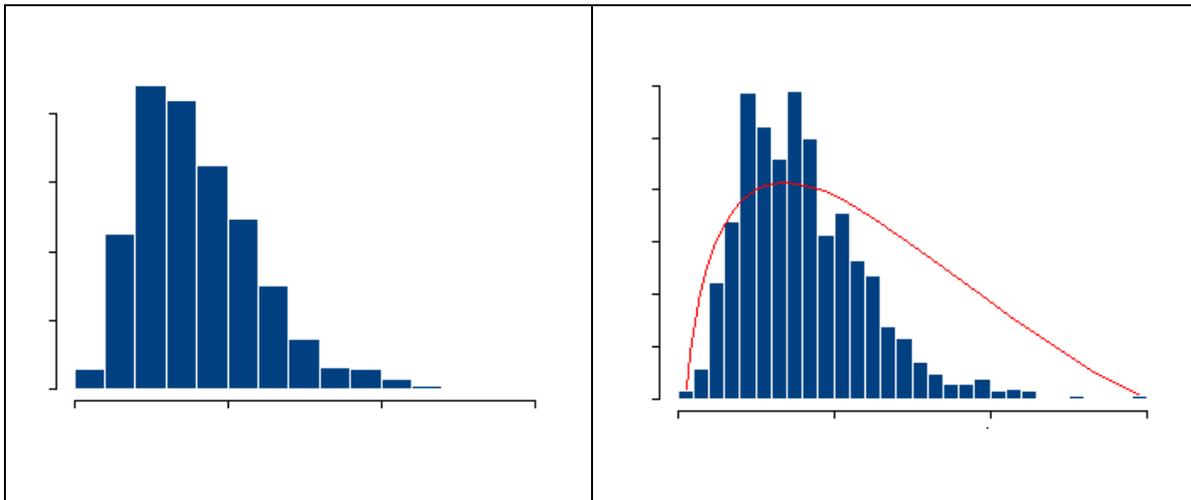
Variabes aleatorias continuas

Ejemplo: Con el fin de realizar un control de calidad en una fábrica de baterías, se mide el tiempo de duración de baterías elegidas al azar y se define la v.a.

X : tiempo de duración de una batería

La v.a. X es esencialmente continua (“tiempo”), siendo su rango el intervalo real $[0, \infty)$. pero supongamos que medimos la duración de la batería en días, es decir “discretizamos” el rango de la v.a. y se convierte en $N_0 = \mathbb{N} \cup \{0\}$. Por tratarse de una v.a. discreta, su función de probabilidad puntual puede representarse mediante un histograma con área total igual a 1. Si medimos la duración en horas, obtenemos un histograma con mayor número de intervalos de menor longitud cada uno, pero que sigue teniendo área total igual a 1.

Si continuamos aumentando la precisión de la medición (minutos, segundos, décimas de segundo, etc), obtenemos como límite de los histogramas una curva suave, y la probabilidad de que la duración de la batería se encuentre entre dos valores a y b ($a < b$) estará dada por el área bajo la curva entre a y b .



Definición: Una v.a. X es *continua* si existe una función

$$f : \mathbb{R} \rightarrow \mathbb{R}^+ = [0, \infty)$$

llamada función de densidad de la v.a. X tal que

$$P(X \in A) = \int_A f(x)dx \quad \forall A \subseteq \mathbb{R}$$

En particular, si $A = [a, b]$, entonces

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

y $P(X = a) = P(a \leq X \leq a) = 0 \quad \forall a \in \mathfrak{R}$.

Propiedad: Para que una función $f(x)$ sea una función de densidad, debe satisfacer

$$f(x) \geq 0 \quad \forall x \in \mathfrak{R}$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Observación: Notar que $f(x)$ no es una probabilidad, de hecho puede ser mayor que 1. Es simplemente el valor de una función en un punto.

Ejemplo: Sea

$$f(x) = \begin{cases} a x^2 & \text{si } 1 \leq x \leq 3 \\ 0 & \text{en otro caso} \end{cases}$$

Otra forma de expresar la densidad es $f(x) = a x^2 I_{[1,3]}(x)$, donde la función I se define como

$$I_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

a) Calcular el valor de la constante a .

$$\int_{-\infty}^{\infty} f(x) dx = 1 \Leftrightarrow \int_1^3 a x^2 dx = 1 \Leftrightarrow a \int_1^3 x^2 dx = 1 \Leftrightarrow a \left. \frac{x^3}{3} \right|_1^3 = 1 \Leftrightarrow a \frac{26}{3} = 1 \Leftrightarrow a = \frac{3}{26}.$$

b) Calcular $P(X \geq 2)$.

$$P(X \geq 2) = \int_2^{\infty} f(x) dx = \int_2^3 \frac{3}{26} x^2 dx = \left. \frac{3}{26} \frac{x^3}{3} \right|_2^3 = \frac{27-8}{26} = \frac{19}{26}.$$

Definición: La **función de distribución acumulada** de una v.a. continua X con función de densidad $f(x)$ se define para todo $x \in \mathfrak{R}$, como

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Ejemplo: En el ejemplo anterior, obtengamos la función de distribución acumulada de la v.a. X .

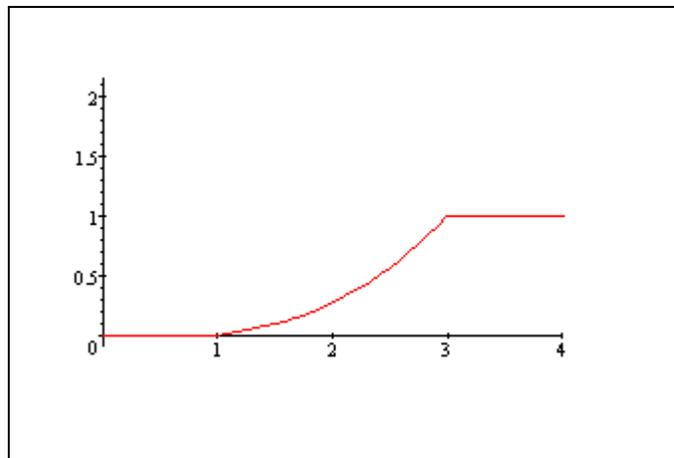
$$\text{Si } x < 1, F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0$$

$$\text{Si } 1 \leq x \leq 3, F(x) = \int_{-\infty}^x f(t) dt = \int_1^x \frac{3}{26} t^2 dt = \frac{3}{26} \frac{t^3}{3} \Big|_1^x = \frac{x^3 - 1}{26}$$

$$\text{Si } x > 3, F(x) = \int_{-\infty}^x f(t) dt = \int_1^3 \frac{3}{26} t^2 dt = 1$$

Resumiendo,

$$F(x) = \begin{cases} 0 & \text{si } x < 1 \\ \frac{x^3 - 1}{26} & \text{si } 1 \leq x \leq 3 \\ 1 & \text{si } x > 3 \end{cases}$$



Observamos que se trata de una función continua, no decreciente que toma valores entre 0 y 1.

Propiedades de la función de distribución acumulada: Sea X una v.a. continua,

- i) $\forall x \in \mathfrak{R}, F_X(x) \in [0,1]$.
- ii) $F_X(x)$ es monótona no decreciente, es decir que si $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$.
- iii) $F_X(x)$ es continua en todo punto.
- iv) $\lim_{x \rightarrow \infty} F_X(x) = 1$ y $\lim_{x \rightarrow -\infty} F_X(x) = 0$

Observemos que las propiedades i), ii) y iv) ya las hemos demostrado en general al considerar las v.a. discretas. Respecto a la propiedad iii), en el caso discreto probamos que la función de distribución es continua a derecha en todo punto, mientras que en este caso es continua en todo punto.

Proposición: Sean a y b tales que $a \leq b$, entonces

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = F(b) - F(a).$$

Dem: Resulta inmediatamente del hecho que, si X es continua, $P(X = x) = 0$

Proposición: Si X es una v.a. continua con función de densidad $f(x)$ y función de distribución acumulada $F(x)$, entonces en todo punto donde $F(x)$ es derivable,

$$F'(x) = \frac{\partial F(x)}{\partial x} = f(x)$$

Dem: Resulta del Teorema Fundamental del Cálculo Integral, y de la definición de $F(x)$.

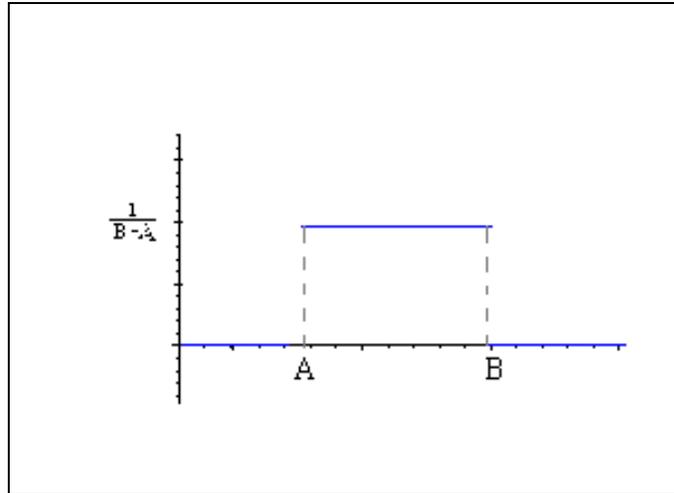
Distribución Uniforme:

Definición: Se dice que X tiene distribución Uniforme en el intervalo $[A,B]$, si su función de densidad es

$$f(x) = \frac{1}{B-A} I_{[A,B]}(x)$$

es decir, la densidad es constante sobre el intervalo $[A,B]$ y 0 fuera de él. A y B son los parámetros de la distribución.

Notación: $X \sim U(A,B)$.



Función de distribución: Hallemos la función de distribución acumulada de $X \sim U(A, B)$.

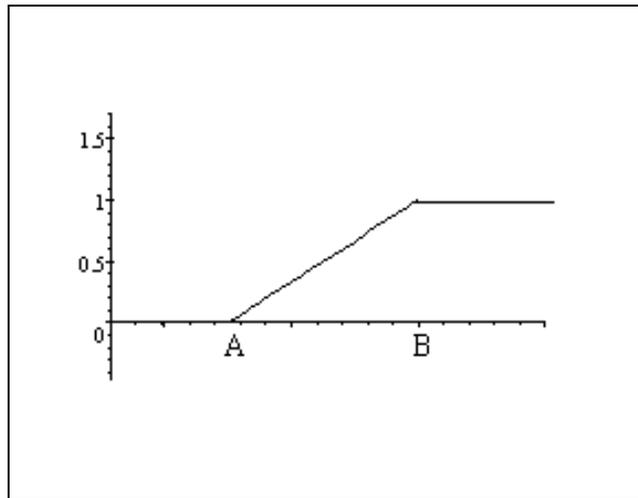
$$\text{Si } x < A \Rightarrow F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0.$$

$$\text{Si } A \leq x \leq B \Rightarrow F(x) = \int_{-\infty}^x f(t) dt = \int_A^x \frac{1}{B-A} dt = \frac{t}{B-A} \Big|_A^x = \frac{x-A}{B-A}.$$

$$\text{Si } x > B \Rightarrow F(x) = \int_{-\infty}^x f(t) dt = \int_A^B \frac{1}{B-A} dt = \frac{t}{B-A} \Big|_A^B = \frac{B-A}{B-A} = 1.$$

Resumiendo,

$$F(x) = \begin{cases} 0 & \text{si } x < A \\ \frac{x-A}{B-A} & \text{si } A \leq x \leq B \\ 1 & \text{si } x > B \end{cases}$$



Percentiles de una distribución continua: Sea X una v.a. continua con función de densidad $f(x)$ y función de distribución acumulada $F(x)$ y sea $0 < p < 1$. El percentil $(100 p)$ -ésimo de la distribución de X es el valor x_p tal que

$$F(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f(t) dt = p$$

Ejemplos: 1) Sea X con función de densidad $f(x) = \frac{3}{26} x^2 I_{[1,3]}(x)$. Su función de distribución está dada por

$$F(x) = \begin{cases} 0 & \text{si } x < 1 \\ \frac{x^3 - 1}{26} & \text{si } 1 \leq x \leq 3 \\ 1 & \text{si } x > 3 \end{cases}$$

Obtengamos el 25-percentil de esta distribución ($p = 0.25$). Buscamos $x_{0.25}$ tal que $F(x_{0.25}) = 0.25$.

$$F(x_{0.25}) = 0.25 \Leftrightarrow \frac{x_{0.25}^3 - 1}{26} = 0.25 \Leftrightarrow x_{0.25} = (0.25 \cdot 26 + 1)^{1/3} = 1.96$$

2) Sea $X \sim U(A, B)$. Su función de distribución está dada por

$$F(x) = \begin{cases} 0 & \text{si } x < A \\ \frac{x - A}{B - A} & \text{si } A \leq x \leq B \\ 1 & \text{si } x > B \end{cases}$$

Hallemos el 50-percentil de esta distribución ($p = 0.50$). Buscamos $x_{0.50}$ tal que $F(x_{0.50}) = 0.50$.

$$F(x_{0.50}) = 0.50 \Leftrightarrow \frac{x_{0.50} - A}{B - A} = 0.50 \Leftrightarrow x_{0.50} = 0.50(B - A) + A = \frac{A + B}{2}.$$

El 50-percentil se denomina **mediana** de la distribución.

Esperanza o valor esperado de una v.a. continua:

Definición: Sea X una v.a. continua con función de densidad $f(x)$, la **esperanza o valor esperado de X** se define como

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx$$

siempre que $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$. Si esta integral es ∞ , la esperanza no puede definirse y decimos que no existe.

Ejemplo: Sea $X \sim U(A, B)$,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_A^B x \frac{1}{B - A} dx = \frac{x^2}{2(B - A)} \Big|_A^B = \frac{B^2 - A^2}{2(B - A)} = \frac{A + B}{2}.$$

Proposición: Si la v.a. continua X tiene función de densidad $f(x)$, entonces la esperanza de cualquier función real $h(X)$, está dada por

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx$$

si $\int_{-\infty}^{\infty} |h(x)|f(x)dx < \infty$.

Propiedad (Linealidad): Si a y b son constantes reales, $E(aX + b) = aE(X) + b$.

Dem: Sea $h(X) = aX + b$, entonces

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx = \int_{-\infty}^{\infty} (ax + b)f(x)dx = a \int_{-\infty}^{\infty} x f(x)dx + b \int_{-\infty}^{\infty} f(x)dx = aE(X) + b.$$

Ejemplo: Dos especies compiten en una región para controlar una limitada cantidad de cierto recurso. sea X : proporción del recurso controlada por la especie 1. Supongamos que $X \sim U(0, 1)$, es decir

$$f(x) = \begin{cases} 1 & \text{si } x \in [0,1] \\ 0 & \text{si } x \notin [0,1] \end{cases}$$

Este modelo de asignación de recursos se denomina “broken stick” o “vara rota” ya que es análogo a quebrar una vara en un punto aleatorio. La especie que controla la mayoría del recurso, controla la cantidad.

$$\text{Sea } h(X) = \max(X, 1 - X) = \begin{cases} 1 - X & \text{si } 0 \leq X < \frac{1}{2} \\ X & \text{si } \frac{1}{2} \leq X \leq 1 \end{cases}$$

El valor esperado para la cantidad controlada por la especie que más controla es:

$$\begin{aligned} E(h(X)) &= \int_{-\infty}^{\infty} h(x)f(x)dx = \int_{-\infty}^{\infty} \max(x, 1 - x) f(x)dx = \int_0^{1/2} (1 - x)f(x)dx + \int_{1/2}^1 x f(x)dx = \\ &= \int_0^{1/2} (1 - x) dx + \int_{1/2}^1 x dx = \left(x - \frac{x^2}{2} \right) \Big|_0^{1/2} + \frac{x^2}{2} \Big|_{1/2}^1 = \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = 1 - \frac{1}{4} = \frac{3}{4}. \end{aligned}$$

Varianza de una v.a. continua:

Definición: Sea X una v.a. continua con esperanza μ_X y densidad $f(x)$, la **varianza de X** , que se denotará $V(X)$, σ_X^2 ó σ^2 , es

$$V(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx$$

y el **desvío standard de X**, es $\sigma_X = +\sqrt{V(X)}$.

Proposición: $V(X) = E(X^2) - (E(X))^2$.

Dem:

$$\begin{aligned} V(X) &= E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx = \int_{-\infty}^{\infty} (x^2 - 2x\mu_X + \mu_X^2) f(x) dx = \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu_X \int_{-\infty}^{\infty} x f(x) dx + \mu_X^2 \int_{-\infty}^{\infty} f(x) dx = E(X^2) - 2\mu_X \mu_X + \mu_X^2 = E(X^2) - \mu_X^2 \end{aligned}$$

como queríamos demostrar.

Ejemplos: Sea $X \sim U(A, B)$, hemos demostrado que $E(X) = \frac{A+B}{2}$, es decir el punto medio del intervalo. Hallemos la varianza de X . Como $V(X) = E(X^2) - (E(X))^2$, necesitamos calcular $E(X^2)$.

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_A^B x^2 \frac{1}{B-A} dx = \frac{x^3}{3(B-A)} \Big|_A^B = \frac{B^3 - A^3}{3(B-A)} = \frac{(B-A)(B^2 + AB + A^2)}{3(B-A)} = \\ &= \frac{(B^2 + AB + A^2)}{3} \end{aligned}$$

Entonces,

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 = \frac{(B^2 + AB + A^2)}{3} - \left(\frac{A+B}{2}\right)^2 = \\ &= \frac{4(B^2 + AB + A^2) - 3(A^2 + 2AB + B^2)}{12} = \frac{B^2 - 2AB + A^2}{12} = \frac{(B-A)^2}{12}. \end{aligned}$$

Por lo tanto, $V(X) = \frac{(B - A)^2}{12}$.

Propiedad de la varianza y del desvío standard: Sea X una v.a. continua con densidad $f(x)$,

$$V(aX + b) = a^2V(X) \quad \text{y} \quad \sigma_{aX+b} = |a|\sigma_X.$$

Dem: : Observemos que, en general,

$$V(h(X)) = \int_{-\infty}^{\infty} (h(x) - E(h(X)))^2 f(x) dx$$

entonces, si $h(x) = ax + b$,

$$\begin{aligned} V(aX + b) &= \int_{-\infty}^{\infty} [(ax + b) - E(ax + b)]^2 f(x) dx = \int_{-\infty}^{\infty} [ax + b - aE(X) - b]^2 f(x) dx = \\ &= \int_{-\infty}^{\infty} [ax - aE(X)]^2 f(x) dx = a^2 \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx = a^2V(X), \end{aligned}$$

como queríamos demostrar.

Obviamente, $\sigma_{aX+b} = |a|\sigma_X$.

VARIABLES ALEATORIAS CONTINUAS

Distribución Uniforme: Recordemos que X tiene distribución uniforme en el intervalo $[A, B]$, si su función de densidad es

$$f(x) = \frac{1}{B-A} I_{[A,B]}(x)$$

Notación: $X \sim U(A, B)$.

Su función de distribución acumulada está dada por:

$$F(x) = \begin{cases} 0 & \text{si } x < A \\ \frac{x-A}{B-A} & \text{si } A \leq x \leq B \\ 1 & \text{si } x > B \end{cases}$$

Esperanza y varianza de una variable aleatoria uniforme: Sea $X \sim U(A, B)$, hemos demostrado que

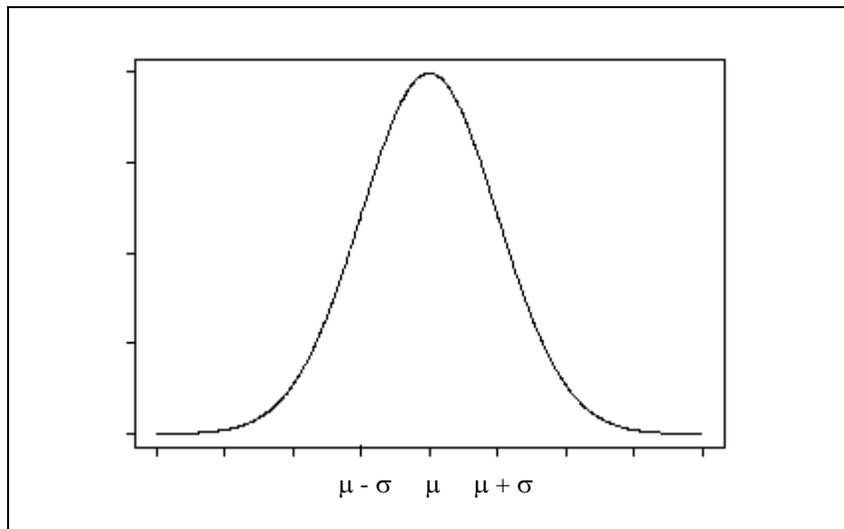
$$E(X) = \frac{A+B}{2} \quad \text{y} \quad V(X) = \frac{(B-A)^2}{12}.$$

Distribución Normal: Se dice que X tiene distribución Normal de parámetros μ y σ^2 ($\mu \in \mathfrak{R}, \sigma > 0$) si su función de densidad es

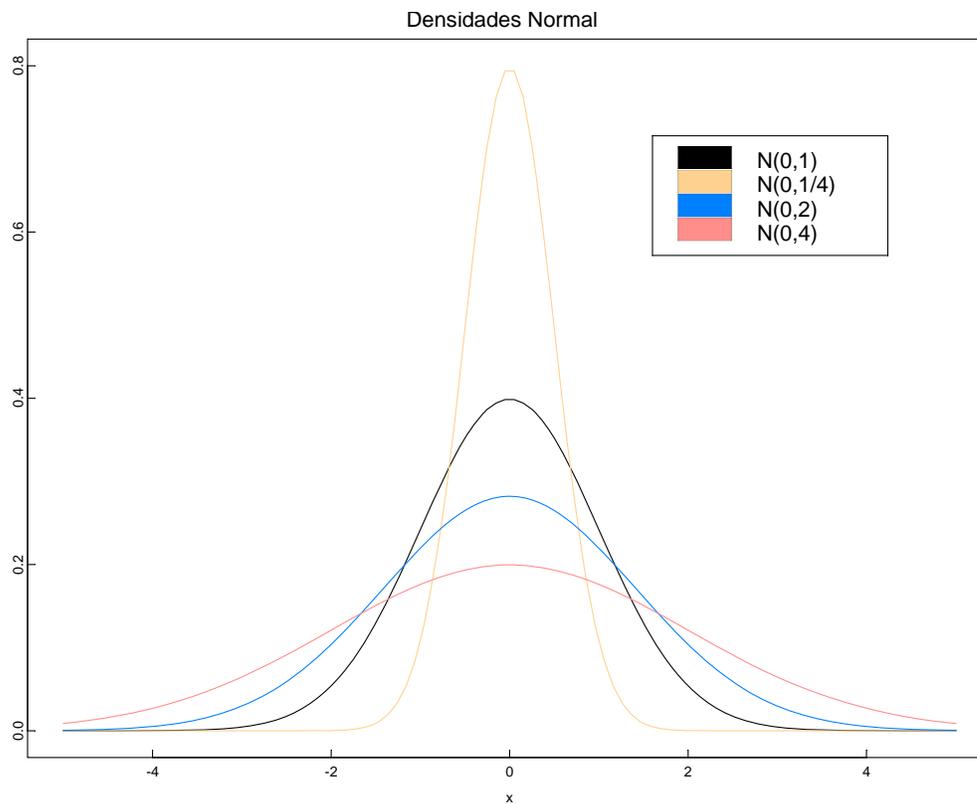
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (1)$$

Notación: $X \sim N(\mu, \sigma^2)$.

El gráfico de la función de densidad normal tiene forma de campana con eje de simetría en $x = \mu$ y puntos de inflexión en $x = \mu + \sigma$ y $x = \mu - \sigma$.



En esta distribución, μ indica la posición de la curva y σ es el parámetro de dispersión. En el siguiente gráfico se muestran densidades correspondientes a $\mu=0$ y distintos valores de σ .



La importancia de la distribución normal radica no sólo en que frecuentemente en la práctica se hallan variables que tienen esta distribución (por ejemplo, los errores de medición) sino porque, bajo ciertas condiciones, suele ser una buena aproximación a la distribución de otras variables aleatorias.

Se puede verificar que en efecto la función (1) es una función de densidad, es decir que la integral sobre toda la recta es 1. No lo haremos, pero sí verificaremos que su gráfico es simétrico respecto de μ , punto en el cual alcanza su único máximo y que tiene puntos de inflexión en $x = \mu + \sigma$ y $x = \mu - \sigma$.

Probemos en primer lugar que la densidad es simétrica respecto de μ , o sea que

$$f(\mu - x) = f(\mu + x) \quad \forall x$$

En efecto,

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

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

y, por lo tanto, se verifica la igualdad.

Observemos ahora que la densidad alcanza un único máximo en $x = \mu$.

$$\frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \right] = 0 \Leftrightarrow -\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \frac{1}{\sigma^2}(x-\mu) = 0$$

$$\Leftrightarrow (x - \mu) = 0 \Leftrightarrow x = \mu.$$

Ejercicio: Verificar que la derivada segunda en $x = \mu$ es menor que 0 y por lo tanto se trata de un máximo y que la densidad tiene dos puntos de inflexión en $x = \mu + \sigma$ y $x = \mu - \sigma$.

Distribución Normal Standard: Se dice que Z tiene distribución normal standard si sus parámetros son $\mu = 0$ y $\sigma^2 = 1$, es decir $Z \sim N(0,1)$. Su función de densidad estará dada por

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

Su función de distribución, que se notará $\Phi(z)$, es:

$$\Phi(z) = F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Esta función está tabulada, ya que su integral no tiene una expresión analítica conocida.

Ejemplo: $Z \sim N(0,1)$,

$$P(Z \leq 1.25) = \Phi(1.25) = 0.8944$$

$$P(Z > 1.25) = 1 - P(Z \leq 1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$$

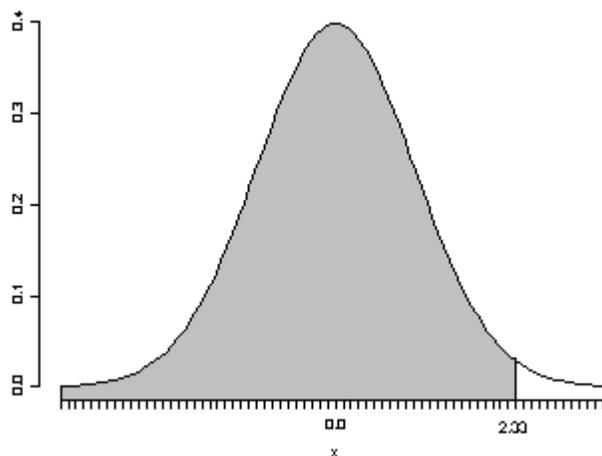
$$P(-0.38 \leq Z \leq 1.25) = \Phi(1.25) - \Phi(-0.38) = 0.5424$$

Percentiles de la distribución Normal Standard: Sea $0 < p < 1$, el percentil (100 p)-ésimo de la distribución normal standard es el valor z tal que

$$\Phi(z) = p,$$

es decir, es el valor que deja a su izquierda un área igual a p .

Ejemplo: $Z \sim N(0,1)$, el percentil 99 de la distribución es 2.33 ya que $\Phi(2.33) = 0.99$.



Propiedades de la distribución Normal:

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

Dem:

$$F_Z(z) = P(Z \leq z) = P\left(\frac{X - \mu}{\sigma} \leq z\right) = P(X \leq \sigma z + \mu) = F_X(\sigma z + \mu)$$

Como F_Z es derivable en todo punto,

$$\begin{aligned} f_Z(z) &= \frac{\partial}{\partial z} F_Z(z) = \frac{\partial}{\partial z} F_X(\sigma z + \mu) = f_X(\sigma z + \mu) \sigma = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}} \sigma = \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \end{aligned}$$

y, por lo tanto $Z \sim N(0,1)$ como queríamos demostrar.

2) Si $Z \sim N(0,1)$ y $\sigma > 0 \Rightarrow X = \sigma Z + \mu \sim N(\mu, \sigma^2)$.

Dem: Ejercicio.

3) Sean $X \sim N(\mu, \sigma^2)$ y $Z \sim N(0,1)$. Si denotamos x_p y z_p a los 100 p-ésimos percentiles de X y Z respectivamente,

$$x_p = \sigma z_p + \mu$$

Dem: El 100 p-ésimo percentil de X es el valor x_p tal que $F(x_p) = p$.

$$\begin{aligned} F(x_p) = p &\Leftrightarrow P(X \leq x_p) = p \Leftrightarrow P\left(\frac{X - \mu}{\sigma} \leq \frac{x_p - \mu}{\sigma}\right) = p \Leftrightarrow \Phi\left(\frac{x_p - \mu}{\sigma}\right) = p \\ &\Leftrightarrow \frac{x_p - \mu}{\sigma} = z_p \Leftrightarrow x_p = \sigma z_p + \mu . \end{aligned}$$

Esperanza y varianza de una variable aleatoria normal: Hallaremos inicialmente la esperanza y la varianza de la distribución normal standard y luego, utilizando propiedades ya demostradas, hallaremos la esperanza y la varianza de una distribución normal en general.

Proposición: Sea $Z \sim N(0, 1)$, entonces $E(Z) = 0$ y $V(Z) = 1$.

Dem:

$$E(Z) = \int_{-\infty}^{\infty} z f(z) dz = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0$$

pues el integrando es una función integrable e impar.

$$V(Z) = E(Z^2) - (E(Z))^2 = E(Z^2) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} dz$$

Aplicando integración por partes, con

$$u = z \quad dv = \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} dz \quad du = dz \quad v = -\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

se obtiene

$$V(Z) = -\frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \lim_{M \rightarrow \infty} \left(-\frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} \Big|_{-M}^M \right) + 1.$$

Aplicando la regla de L'Hospital,

$$\lim_{M \rightarrow \infty} \left(-\frac{1}{\sqrt{2\pi}} \frac{M}{e^{\frac{M^2}{2}}} \right) = \lim_{M \rightarrow \infty} \left(-\frac{1}{\sqrt{2\pi}} \frac{1}{M e^{\frac{M^2}{2}}} \right) = 0$$

y

$$\lim_{M \rightarrow -\infty} \left(\frac{1}{\sqrt{2\pi}} \frac{M}{e^{\frac{M^2}{2}}} \right) = \lim_{M \rightarrow -\infty} \left(\frac{1}{\sqrt{2\pi}} \frac{1}{M e^{\frac{M^2}{2}}} \right) = 0$$

y, por lo tanto, $V(Z) = 1$.

Proposición: Sea $X \sim N(\mu, \sigma^2)$, entonces $E(X) = \mu$ y $V(X) = \sigma^2$.

Dem: Recordemos que, si $X \sim N(\mu, \sigma^2)$, entonces $\frac{X - \mu}{\sigma} \sim N(0,1)$.

Como $E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = 0$, por linealidad de la esperanza,

$$\frac{1}{\sigma}(E(X) - \mu) = 0 \Rightarrow E(X) = \mu.$$

Como $V(Z) = V\left(\frac{X - \mu}{\sigma}\right) = 1$, por propiedades de la varianza,

$$\frac{1}{\sigma^2}V(X) = 1 \Rightarrow V(X) = \sigma^2.$$

Distribución Gamma: Se trata de una familia de distribuciones que provee un modelo adecuado para histogramas que presentan cierto tipo de asimetría. Antes de presentar a las v.a. con distribución Gamma, es necesario recordar cómo se define la función Gamma o factorial, la cual cumple un rol importante en muchas ramas de la Matemática..

Definición: Dado $\alpha > 0$, se define la función Gamma o función factorial como

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

Propiedades:

1) Si $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

2) Si $\alpha \in \mathbb{N}$, $\Gamma(\alpha) = (\alpha - 1)!$

3) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Dem: 1) Sea $\alpha > 1$. Aplicando integración por partes con $u = x^{\alpha-1}$ y $dv = e^{-x} dx$,

$$\begin{aligned} \Gamma(\alpha) &= \int_0^{\infty} x^{\alpha-1} e^{-x} dx = -x^{\alpha-1} e^{-x} \Big|_0^{\infty} + \int_0^{\infty} (\alpha-1)x^{\alpha-2} e^{-x} dx = \\ &= -\lim_{M \rightarrow \infty} \frac{M^{\alpha-1}}{e^M} + 0 + (\alpha-1) \int_0^{\infty} x^{(\alpha-1)-1} e^{-x} dx = 0 + 0 + (\alpha-1)\Gamma(\alpha-1) = (\alpha-1)\Gamma(\alpha-1). \end{aligned}$$

2) Ejercicio.

3)

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} x^{\frac{1}{2}-1} e^{-x} dx = \int_0^{\infty} x^{-\frac{1}{2}} e^{-x} dx$$

Aplicaremos el siguiente cambio de variable: $u = \sqrt{2x}$, con lo cual $du = \frac{2}{\sqrt{2x}} dx$.

Entonces,

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} \sqrt{2} e^{-\frac{u^2}{2}} du = \frac{1}{2} \int_{-\infty}^{\infty} \sqrt{2} e^{-\frac{u^2}{2}} du = \sqrt{\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \sqrt{\pi},$$

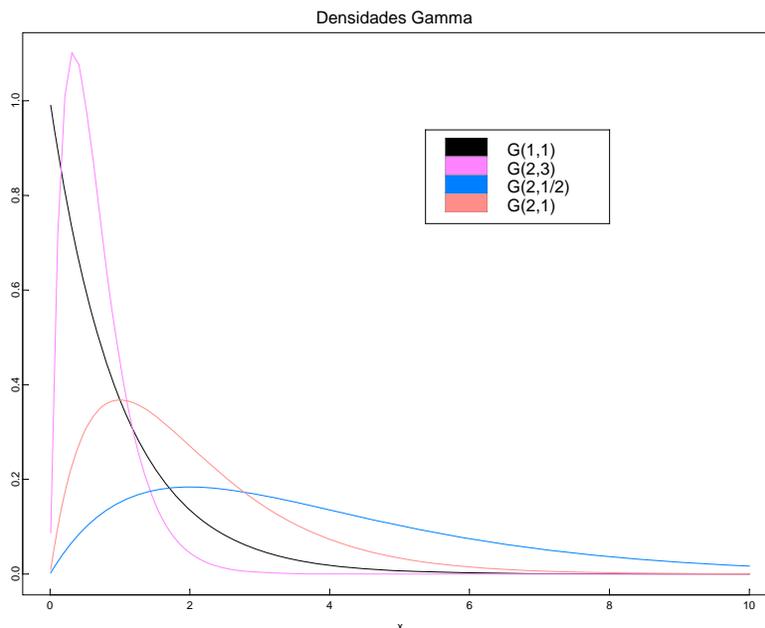
ya que la integral de la última igualdad es la integral de la densidad normal standard y por lo tanto es igual a 1.

Definición: Se dice que X tiene distribución Gamma de parámetros α y λ ($\alpha > 0$, $\lambda > 0$) si su función de densidad está dada por

$$f(x) = \frac{e^{-\lambda x} x^{\alpha-1} \lambda^{\alpha}}{\Gamma(\alpha)} I_{(0, \infty)}(x)$$

Notación: $X \sim \Gamma(\alpha, \lambda)$ o bien $X \sim G(\alpha, \lambda)$.

En el siguiente gráfico se muestra la densidad correspondiente a $X \sim G(\alpha, \lambda)$ para distintos valores de los parámetros.



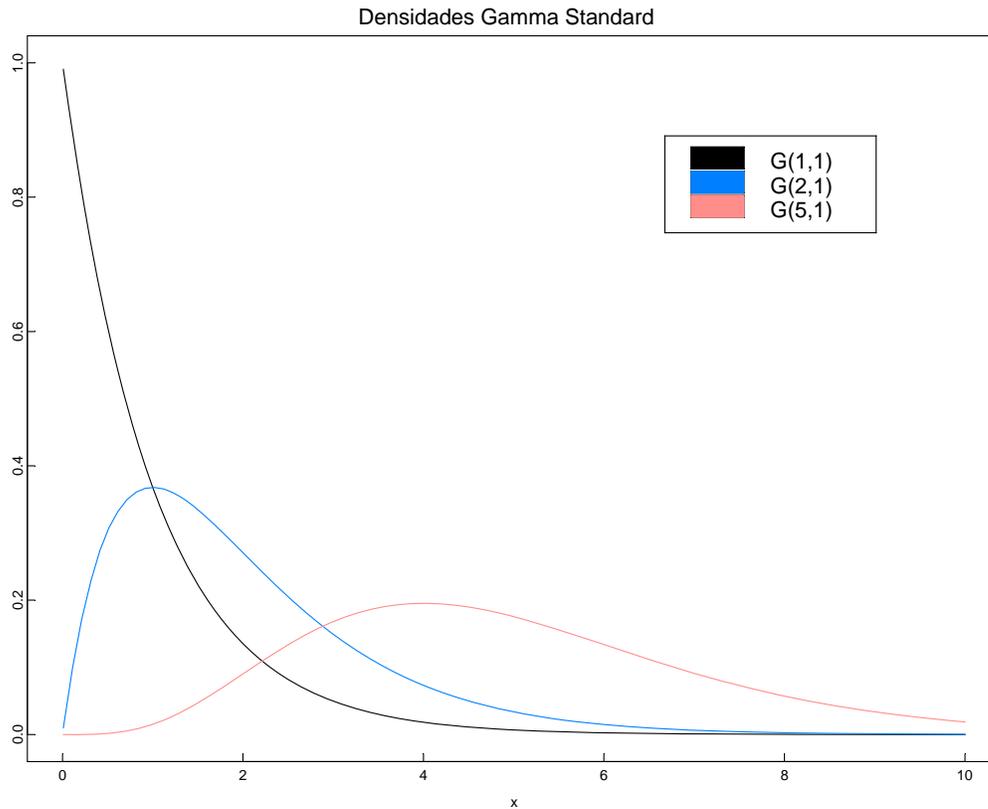
Definición: Si $\lambda = 1$, la distribución se denomina Gamma standard. Es decir, X tiene distribución Gamma standard de parámetro α ($X \sim \Gamma(\alpha, 1)$) si su densidad está dada por:

$$f(x) = \frac{e^{-x} x^{\alpha-1}}{\Gamma(\alpha)} I_{(0,\infty)}(x)$$

Esta función de densidad es estrictamente decreciente si $\alpha \leq 1$, y si $\alpha > 1$ alcanza un máximo y después decrece.

La distribución Gamma standard está tabulada para diferentes valores de α .

Volviendo a la densidad Gamma general, λ es un parámetro de escala ya que valores de λ distintos de 1, comprimen o expanden la curva.



Esperanza y varianza de una variable aleatoria Gamma:

Proposición: $X \sim \Gamma(\alpha, \lambda)$, entonces $E(X) = \frac{\alpha}{\lambda}$ y $V(X) = \frac{\alpha}{\lambda^2}$.

Dem:

$$\begin{aligned}
 E(X) &= \int_0^{\infty} x \frac{e^{-\lambda x} x^{\alpha-1} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \int_0^{\infty} \frac{e^{-\lambda x} x^{\alpha} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \int_0^{\infty} \frac{e^{-\lambda x} x^{(\alpha+1)-1} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \\
 &= \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)} \int_0^{\infty} \frac{e^{-\lambda x} x^{(\alpha+1)-1} \lambda^{\alpha+1}}{\Gamma(\alpha+1)} dx = \frac{\alpha \Gamma(\alpha)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda}.
 \end{aligned}$$

Observemos que la última integral es la integral, sobre todo su rango, de la densidad de una v.a. con distribución $\Gamma(\alpha+1, \lambda)$ y por lo tanto es igual a 1.

Calculemos ahora $E(X^2)$.

$$\begin{aligned}
 E(X^2) &= \int_0^{\infty} x^2 \frac{e^{-\lambda x} x^{\alpha-1} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \int_0^{\infty} \frac{e^{-\lambda x} x^{\alpha+1} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \int_0^{\infty} \frac{e^{-\lambda x} x^{\alpha+2-1} \lambda^{\alpha}}{\Gamma(\alpha)} dx = \\
 &= \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} \int_0^{\infty} \frac{e^{-\lambda x} x^{\alpha+2-1} \lambda^{\alpha+2}}{\Gamma(\alpha+2)} dx = \frac{(\alpha+1)\alpha \Gamma(\alpha)}{\lambda^2 \Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{\lambda^2}.
 \end{aligned}$$

Observemos que la última integral es la integral, sobre todo su rango, de la densidad de una v.a. con distribución $\Gamma(\alpha+2, \lambda)$ y por lo tanto es igual a 1.

$$\text{Finalmente, } V(X) = \frac{(\alpha+1)\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha^2}{\lambda^2} + \frac{\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}, \text{ como queríamos demostrar.}$$

Propiedad: Si $X \sim \Gamma(\alpha, \lambda)$ y $a > 0$, $aX \sim \Gamma(\alpha, \lambda/a)$.

Dem: Ejercicio.

Nota: Esta última propiedad permite obtener probabilidades para una v.a. con distribución Gamma a partir de una distribución Gamma standard. En efecto, supongamos que $X \sim \Gamma(\alpha, \lambda)$, entonces $\lambda X \sim \Gamma(\alpha, 1)$ y, por ejemplo

$$P(X \leq x) = P(\lambda X \leq \lambda x) = F_{\lambda X}(\lambda x)$$

Observación: Algunos autores, por ejemplo J. Devore, utilizan otra parametrización de la distribución Gamma, definiendo como segundo parámetro de la distribución a $1/\lambda$. es decir: $X \sim \Gamma(\alpha, \beta)$ si su función de densidad está dada por

$$f(x) = \frac{e^{-\frac{x}{\beta}} x^{\alpha-1}}{\beta^{\alpha} \Gamma(\alpha)} I_{(0, \infty)}(x)$$

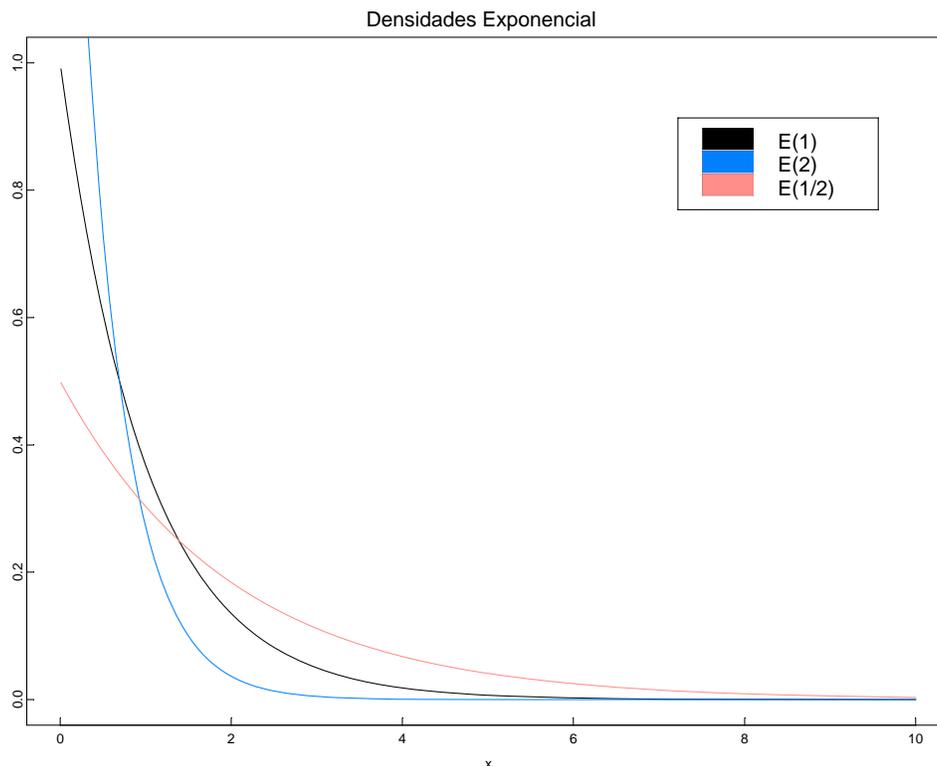
En este caso, $E(X) = \alpha \beta$ y $V(X) = \alpha \beta^2$.

Distribución Exponencial: Se trata de un caso particular de la distribución Gamma, ya que una v.a. exponencial es una v.a. Gamma con parámetro $\alpha = 1$.

Definición: X tiene distribución exponencial de parámetro λ ($\lambda > 0$) si su función de densidad está dada por:

$$f(x) = \lambda e^{-\lambda x} I_{(0, \infty)}(x)$$

Notación: $X \sim \varepsilon(\lambda)$.



Función de distribución de una v.a. exponencial: Si $X \sim \varepsilon(\lambda)$, su función de distribución acumulada está dada por

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

En efecto, si $x > 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = -e^{-\lambda x} + 1,$$

como queríamos demostrar.

Proposición: Si $X \sim \varepsilon(\lambda)$, entonces $E(X) = \frac{1}{\lambda}$ y $V(X) = \frac{1}{\lambda^2}$.

Dem: Se deduce inmediatamente de la esperanza y la varianza de una v.a. Gamma con parámetro $\alpha = 1$.

Ejemplo: Supongamos que el tiempo de respuesta de una terminal conectada en línea es una v.a. X con distribución exponencial con esperanza igual a 5 segundos.

a) ¿Cuál es la probabilidad de que el tiempo de respuesta sea mayor de 10 segundos?

Observemos que, dado que $E(X)=5$, $X \sim \varepsilon(1/5)$, entonces

$$P(X > 10) = 1 - F(10) = 1 - \left(1 - e^{-\frac{1}{5}10}\right) = e^{-2} = 0.135.$$

b) ¿Cuál es la probabilidad de que el tiempo de respuesta esté entre 5 y 10 segundos?

$$P(5 \leq X \leq 10) = F(10) - F(5) = \left(1 - e^{-\frac{10}{5}}\right) - \left(1 - e^{-\frac{5}{5}}\right) = e^{-1} - e^{-2} = 0.233.$$

Proposición (Propiedad de Falta de Memoria): Sea $X \sim \varepsilon(\lambda)$, y sean s y t números reales positivos cualesquiera,

$$P(X > s + t | X > s) = P(X > t)$$

Dem: Ejercicio. (Sugerencia: Usar que si $X \sim \varepsilon(\lambda)$, $P(X > s) = e^{-\lambda s}$).

Relación de la distribución exponencial con los procesos de Poisson: Supongamos que la ocurrencia de cierto tipo de eventos sigue un proceso de Poisson de intensidad o tasa media de ocurrencia ν , y por lo tanto la v.a. X_t : “número de ocurrencias en un intervalo de longitud t ” tiene distribución $P(\nu t)$.

Se puede demostrar que la v.a. T : “tiempo hasta la ocurrencia del primer evento” (o equivalentemente, tiempo entre la ocurrencia de dos eventos sucesivos), tiene distribución exponencial.

Proposición: Dado un proceso de Poisson de intensidad ν , si se define la v.a. T : “tiempo hasta la ocurrencia del primer evento”, entonces $T \sim \varepsilon(\nu)$.

Dem: Si $t \leq 0$, $F_T(t) = 0$. Sea $t > 0$,

$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - P(X_t = 0).$$

En efecto, si el tiempo hasta la primera ocurrencia es mayor que t , no ha ocurrido ningún evento en el intervalo $(0,t)$ y recíprocamente. Entonces,

$$F_T(t) = 1 - P(X_t = 0) = 1 - \frac{e^{-\nu} (\nu t)^0}{0!} = 1 - e^{-\nu t},$$

y por lo tanto

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\nu x} & \text{si } x > 0 \end{cases}$$

es decir, $T \sim \varepsilon(\nu)$.

Ejercicio: Demostrar que el tiempo de espera hasta la segunda ocurrencia del evento tiene distribución $\Gamma(2, \nu)$.

Función generadora de momentos:

Definición: Si X es una variable aleatoria, el momento de orden k de X se define como

$$E(X^k)$$

siempre que la esperanza exista.

Notemos que

$E(X) = \mu$	1er momento: posición
$E(X^2) = \sigma^2 + \mu^2$	2do momento: relacionado con una medida de dispersión
$E(X^3)$	3er momento: relacionado con una medida de asimetría
$E(X^4)$	4to momento: relacionado con la kurtosis

Definición: La función generadora de momentos de una v.a. X es una función a valores reales $M_X(t)$, definida como

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_{x \in R_X} e^{tx} p_X(x) & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{si } X \text{ es continua} \end{cases}$$

siempre que el valor esperado exista para todo $t \in (-h, h)$, $h > 0$. Esta última es una condición técnica necesaria para que $M_X(t)$ sea diferenciable en 0.

Se denomina función generadora de momentos porque los momentos de X ($E(X^n)$) pueden ser obtenidos derivando esta función y evaluando la derivada en $t = 0$, tal como lo establece el siguiente teorema.

Teorema: Sea X una v.a. para la cual existe la función generadora de momentos $M_X(t)$, entonces

$$E(X^n) = \left. \frac{\partial^n}{\partial t^n} M_X(t) \right|_{t=0}$$

La demostración se basa en el siguiente lema de Cálculo avanzado (ver por ejemplo, Advanced Calculus, D. Widder (1961)):

Lema: Si la función $g(t)$ definida por

$$g(t) = \sum_x e^{tx} p(x) \quad \text{ó} \quad g(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

converge para todo $t \in (-h, h)$ para algún $h > 0$, entonces existen las derivadas de orden n de $g(t)$ para todo $t \in (-h, h)$ y para todo n entero positivo y se obtienen como

$$\frac{\partial^n g(t)}{\partial t^n} = \sum_x \frac{\partial^n e^{tx}}{\partial t^n} p(x) \quad \text{ó} \quad \frac{\partial^n g(t)}{\partial t^n} = \int_{-\infty}^{\infty} \frac{\partial^n e^{tx}}{\partial t^n} f(x) dx$$

Demostración del Teorema: Si la función generadora de momentos existe para todo $t \in (-h, h)$ para algún $h > 0$, aplicando el lema,

$$\frac{\partial^n M_X(t)}{\partial t^n} = \sum_x \frac{\partial^n e^{tx}}{\partial t^n} p(x) \quad \text{ó} \quad \frac{\partial^n M_X(t)}{\partial t^n} = \int_{-\infty}^{\infty} \frac{\partial^n e^{tx}}{\partial t^n} f(x) dx$$

$$\frac{\partial^n M_X(t)}{\partial t^n} = \sum_x x^n e^{tx} p(x) \quad \text{ó} \quad \frac{\partial^n M_X(t)}{\partial t^n} = \int_{-\infty}^{\infty} x^n e^{tx} f(x) dx$$

Evaluando estas derivadas en 0 ,

$$\left. \frac{\partial^n M_X(t)}{\partial t^n} \right|_{t=0} = \sum_x x^n p(x) = E(X^n) \quad \text{ó} \quad \left. \frac{\partial^n M_X(t)}{\partial t^n} \right|_{t=0} = \int_{-\infty}^{\infty} x^n f(x) dx = E(X^n)$$

Ejemplos: 1) Sea X una v.a. con distribución exponencial de parámetro λ , o sea con densidad

$$f_X(x) = \lambda e^{-\lambda x} I_{(0, \infty)}(x)$$

$$M_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda-t)x} dx = \frac{\lambda}{\lambda-t} \int_0^{\infty} (\lambda-t) e^{-(\lambda-t)x} dx = \frac{\lambda}{\lambda-t}$$

siempre que $t < \lambda$.

Calculemos ahora $E(X)$ y $V(X)$.

$$E(X) = \left. \frac{\partial M_X(t)}{\partial t} \right|_{t=0} = \left. \frac{\partial}{\partial t} \left(\frac{\lambda}{\lambda - t} \right) \right|_{t=0} = \left. \frac{\lambda}{(\lambda - t)^2} \right|_{t=0} = \frac{1}{\lambda}.$$

Como $V(X) = E(X^2) - (E(X))^2$, calculemos $E(X^2)$.

$$E(X^2) = \left. \frac{\partial^2 M_X(t)}{\partial t^2} \right|_{t=0} = \left. \frac{\partial}{\partial t} \left(\frac{\lambda}{(\lambda - t)^2} \right) \right|_{t=0} = \left. \frac{2\lambda(\lambda - t)}{(\lambda - t)^4} \right|_{t=0} = \frac{2}{\lambda^2}$$

entonces, $V(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

2) Sea X una v.a. con distribución Binomial de parámetros, n y p , o sea $X \sim \text{Bi}(n, p)$. Su función de probabilidad puntual es

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{si } 0 \leq k \leq n$$

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^n e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^n \binom{n}{k} (e^t p)^k (1-p)^{n-k} = (e^t p + 1 - p)^n.$$

Calculemos ahora $E(X)$ y $V(X)$.

$$E(X) = \left. \frac{\partial M_X(t)}{\partial t} \right|_{t=0} = \left. \frac{\partial (e^t p + 1 - p)^n}{\partial t} \right|_{t=0} = n(e^t p + 1 - p)^{n-1} p e^t \Big|_{t=0} = np.$$

$$E(X^2) = \left. \frac{\partial^2 M_X(t)}{\partial t^2} \right|_{t=0} = \left. \frac{\partial}{\partial t} \left(n(e^t p + 1 - p)^{n-1} p e^t \right) \right|_{t=0} =$$

$$= \left. \left(n(n-1)(e^t p + 1 - p)^{n-2} (p e^t)^2 + n(e^t p + 1 - p)^{n-1} p e^t \right) \right|_{t=0} = n(n-1)p^2 + np.$$

Entonces, $V(X) = E(X^2) - (E(X))^2 = n(n-1)p^2 + np - (np)^2 = -np^2 + np = np(1-p)$.

Propiedad: Sea X una v.a. con función generadora de momentos $M_X(t)$, entonces si $Y = aX + b$, entonces $M_Y(t) = e^{bt} M_X(at)$.

Dem: Ejercicio.

Unicidad de $M_X(t)$: Además de permitir calcular momentos de una v.a., la función generadora de momentos permite identificar la función de densidad o de probabilidad de una v.a. debido a la propiedad de **unicidad**, la cual establece que hay una correspondencia uno a uno entre funciones de densidad o probabilidad y funciones generadoras de momentos.

Teorema de Unicidad: Si existe la función generadora de momentos de una variable aleatoria, es única. Además la función generadora de momentos determina a la función de densidad o probabilidad de la v.a. salvo a lo sumo en un conjunto de probabilidad 0.

A continuación, presentamos una tabla con la función generadora de momentos de algunas de las distribuciones que hemos estudiado.

Distribución	$M_X(t)$
Bi(n,p)	$(e^t p + 1 - p)^n$
P(λ)	$e^{-\lambda(e^t - 1)}$
N(μ, σ^2)	$e^{\frac{\sigma^2 t^2}{2} + \mu t}$
E(λ)	$\frac{\lambda}{\lambda - t}$
G(α, λ)	$\left(\frac{\lambda}{\lambda - t}\right)^\alpha$
U(a,b)	$\frac{e^{tb} - e^{ta}}{t(b - a)}$
G(p)	$\frac{p e^t}{1 - (1 - p) e^t}$
BN(r,p)	$\left(\frac{p e^t}{1 - (1 - p) e^t}\right)^r$

Ejercicio: ¿Para qué valores de t existe cada una de las funciones generadoras de momentos de la tabla anterior?

Generación de Números Aleatorios

Números “*elegidos al azar*” son útiles en diversas aplicaciones, entre las cuáles podemos mencionar:

- **Simulación o métodos de Monte Carlo:** se simula un proceso natural en forma computacional. Estas aplicaciones se realizan en muy variados campos con el fin de emular distintos comportamientos: física (por ejemplo, para simular colisiones entre partículas), ingeniería (diseño de obras hidráulicas, puentes, etc.), inversiones de capital, redes, servicios a clientes, call centers, etc. La simulación a través de la computadora es una herramienta poderosa para comprender la naturaleza de sistemas complejos.
- **Muestreo:** con el fin de seleccionar una submuestra de una población.
- **Análisis Numérico:** algunas técnicas para resolver problemas de análisis numérico complejos han sido desarrolladas usando números aleatorios.
- **Programación:** la generación de valores aleatorios puede ser útil para poner a prueba la efectividad de un algoritmo. También son útiles en criptología.

A pesar de que fue en la década del 40 que las primeras computadoras modernas fueron desarrolladas, la simulación ya existía en forma embrionaria aún antes de que la computadora apareciera en escena. Así, por ejemplo, en la segunda mitad del siglo XIX, se realizaban experiencias arrojando agujas al azar sobre una superficie reglada con el fin de estimar el número π . En 1908 W. S. Gosset, bajo el seudónimo de Student, realizaba un muestreo experimental con el fin de descubrir la distribución de un estimador de la correlación en una distribución normal bivariada. En ese momento los números aleatorios se generaban mediante métodos observacionales (mecanismos físicos) tales como tirar un dado, extraer una carta de un mazo o mediante una ruleta.

Dado el esfuerzo que significaba generar números aleatorios cada vez que eran necesarios, parece razonable que se hayan construido tales números y luego tabulado. Tippett (1927) publicó una tabla con 41600 números aleatorios “tomados en forma aleatoria de informes censales”. Cada número era uno de los enteros 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y el usuario tomaba varios de ellos y agregaba un punto decimal para formar un número aleatorio entre 0 y 1. Desde ese momento fueron propuestos una serie de generadores de números aleatorios. La primera máquina fue usada en 1939 por Kendall y Babington-Smith con el fin de producir una tabla de 100000 dígitos aleatorios y en 1955 la RAND Corporation utilizó extensamente una tabla de 1000000 dígitos aleatorios que fue obtenida a partir de una ruleta electrónica especialmente diseñada. ERNIE fue una famosa máquina de números aleatorios que fue usada por la lotería británica, es decir la *British Premium Savings Bonds Lottery*.

Poco después de la aparición de las computadoras, se comenzó a buscar maneras eficientes de obtener números aleatorios, pues aún cuando se podían usar las tablas existentes éste era un recurso limitado, ya sea por el espacio de memoria necesario como

por resultar, en algunos casos, cortas. Si bien máquinas como ERNIE podrían haber trabajado junto con una computadora, una solución en la que la computadora provee todo parecía más satisfactoria. La búsqueda se orientó, entonces, a la producción de números aleatorios usando operaciones aritméticas en una computadora. John von Neumann sugirió en un principio, alrededor de 1946, usar el método del “cuadrado medio”. Su idea era calcular el cuadrado del número aleatorio anterior y tomar los dígitos del medio del número calculado. Así, por ejemplo, si queremos generar un número aleatorio de 10 dígitos y el número anterior es

5772156649 \implies 33317792380594909201

el nuevo número será 7923805949.

La primera pregunta que cabe hacer es: ¿porqué motivo un número generado por este procedimiento que es determinístico, va a resultar aleatorio?. La respuesta es que el número no es aleatorio, pero parece serlo, en el sentido en que en una aplicación la relación real entre un número y el siguiente no tiene ningún significado físico. Por lo tanto, el carácter no aleatorio no es una característica indeseable y podría ser que el “cuadrado medio” resultase ser un buen “*batido*” del número anterior. Es claro, de todas formas, que un mecanismo de esta naturaleza no podría haber reemplazado a ERNIE.

Las secuencias de números generadas en forma determinística reciben el nombre de secuencias pseudo-aleatorias o quasi-aleatorias, si bien nosotros nos referiremos a ellas como secuencias aleatorias, sobreentendiendo que sólo “parecen” aleatorias. Números aleatorios generados en forma determinística en una computadora funcionan muy bien en muchísimas aplicaciones, a condición de que el método de generación sea bueno.

Volviendo a la propuesta de von Neumann, ésta no parece ser una buena fuente de números aleatorios. Podría suceder que la secuencia caiga en un ciclo corto de repeticiones, siendo el caso extremo el del cero el cual, si aparece en la secuencia, seguirá repitiéndose siempre. A partir de los años 50 se realizaron diversas experiencias con el método propuesto por von Neumann. Trabajando con números de 4 dígitos en lugar de 10, G. E. Forsythe probó con 16 números iniciales. Con 12 de ellos terminó con el ciclo 6100, 2100, 4100, 8100, 6100, etc. Y con otras dos terminó en cero. En efecto,

$6100^{**2} = 37210000$ $2100^{**2} = 4410000$ $4100^{**2} = 16810000$ $8100^{**2} = 65610000$
--

Metropolis realizó muchas pruebas con los números del “*middle-square*”, en especial con sistemas de números binarios. Mostró que en secuencias de 20 dígitos, hay 13 ciclos diferentes en los que la secuencia puede caer, el más largo de los cuales tiene longitud

142. Estas falencias del “*middle-square*” son algunas de las consideraciones que debemos hacer ante un generador de números aleatorios.

En principio consideraremos métodos para generar números con distribución uniforme en el intervalo (0,1). Ésto podemos lograrlo generando enteros X_n entre 0 y un número natural m y luego tomando la fracción:

$$U_n = \frac{X_n}{m}$$

Usualmente m es un número muy grande. El más popular de los generadores de números aleatorios es el *Método Lineal de Congruencias*, que es un caso especial del método introducido por Lehmer en 1949.

Dados cuatro números m , a , c y X_0 , formamos la secuencia de números aleatorios X_n de la siguiente forma

$$X_{n+1} \equiv (aX_n + c) \pmod{m}, \quad n \geq 0$$

es decir que X_{n+1} es el resto entero de dividir $aX_n + c$ por m (y por lo tanto es un entero entre 0 y $m-1$). Esta es una secuencia lineal congruente. Tengamos en cuenta que

m es el módulo $m > 0$
 a es el multiplicador $0 \leq a < m$
 c es el incremento $0 \leq c < m$
 X_0 es la semilla o valor inicial

En el caso en que $c = 0$, el método recibe el nombre de multiplicativo secuencial.

Por ejemplo, si $m = 10$ y $X_0 = a = c = 7$, entonces la secuencia obtenida es

7, 6, 9, 0, 7, 6, 9, 0.....

En cambio, si $m = 8$, para la misma elección del resto de las constantes, la secuencia sería:

0, 7, 0, 7....

Ésto muestra que la elección de los números m , a y c es crucial y que siempre se caerá en un loop, es decir en un ciclo de repeticiones, que se llama *período*. Es claro que cuanto más grande sea m , mayor es la posibilidad de que el período sea largo.

En realidad, las distintas elecciones de los parámetros son sometidas a una batería de tests con los que se chequean las propiedades de los números generados.

Como ya observamos más arriba, con estos algoritmos se generan números aleatorios que se comportan como si proviniesen de una distribución $U(0,1)$. La pregunta que es razonable hacerse es "porqué ésto es suficiente". El siguiente teorema nos da una respuesta.

Teorema: Sean U una variable aleatoria con distribución $U(0,1)$ y G una función de distribución acumulada continua y estrictamente creciente. Si $X = G^{-1}(U)$, entonces la función de distribución acumulada de X es G , es decir $F_X = G$.

Dem:

Recordemos que si $U \sim U(0,1)$, entonces su función de distribución es de la forma

$$F_U(u) = \begin{cases} 0 & \text{si } u \leq 0 \\ u & \text{si } 0 < u < 1 \\ 1 & \text{si } u \geq 1 \end{cases}$$

Por lo tanto, como G es una función estrictamente creciente y su imagen pertenece al intervalo $(0,1)$, entonces

$$F_X(x) = P(X \leq x) = P(G^{-1}(U) \leq x) = P(U \leq G(x)) = F_U(G(x)) = G(x)$$

con lo que queda demostrado el teorema.

Ejemplo: En el caso de una variable $X \sim E(\lambda)$, la función de distribución acumulada es de la forma

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

Dado $y \in (0,1)$, la inversa de F_X es

$$F_X^{-1}(y) = -\frac{1}{\lambda} \ln(1 - y)$$

Luego, si $U \sim U(0,1)$,

$$-\frac{1}{\lambda} \ln(1 - U) \sim E(\lambda)$$

Si la distribución G tiene saltos o es constante de a trozos, no existirá su inversa. Sin embargo se puede demostrar que existe una H con las propiedades requeridas en el teorema anterior, de manera que, aunque sin demostración, enunciaremos el siguiente resultado.

Teorema: Sean U una variable aleatoria con distribución $U(0,1)$ y G una función de distribución acumulada. Existe una función H tal que $H(U)$ tiene distribución acumulada G .

Ejemplos: Queremos generar una variable con distribución de Bernoulli de parámetro p a partir de una v.a. uniforme. Podemos aplicar el siguiente procedimiento. Generamos $U \sim U(0,1)$ y definimos:

$$X = \begin{cases} 1 & \text{si } 0 < U \leq p \\ 0 & \text{si } p < U \leq 1 \end{cases}$$

En efecto, la nueva variable X toma sólo dos valores (0 y 1) y dado que $p \in (0,1)$

$$P(X = 1) = P(U \leq p) = p$$

y por lo tanto X tiene la distribución deseada.

Notemos que en lugar del intervalo $(0, p]$ podríamos haber tomado cualquier intervalo en $(0,1)$ de longitud p .

Vectores aleatorios

Hasta ahora hemos estudiado modelos de probabilidad para una única variable aleatoria. Sin embargo, en muchos casos interesa construir modelos que involucren a más de una variable. Consideraremos inicialmente el caso de vectores aleatorios bidimensionales y luego extenderemos las definiciones y propiedades a vectores de dimensión mayor que 2.

Definición: Sean X e Y v.a. discretas definidas sobre un espacio muestral S . La **función de probabilidad conjunta** del par (X, Y) , $p_{XY}(x, y)$ se define como

$$p_{XY}(x, y) = P(X = x, Y = y)$$

El conjunto $R_{XY} = \{(x, y) / x \in R_X, y \in R_Y\}$ es el recorrido o rango del vector aleatorio (X, Y) .

Dado cualquier conjunto $A \subseteq \mathfrak{R}^2$,

$$P((X, Y) \in A) = \sum_{(x, y) \in A} p_{XY}(x, y)$$

Una función de probabilidad conjunta satisface:

- $p_{XY}(x, y) \geq 0 \quad \forall (x, y)$
- $\sum_x \sum_y p_{XY}(x, y) = 1$

Ejemplos: 1) De una urna que contiene 6 bolillas blancas y 4 negras se extraen sin reposición 3 bolillas. Se definen

X : número de bolillas blancas extraídas

$$Y = \begin{cases} 1 & \text{si el número de bolillas negras extraídas es par ó 0} \\ 0 & \text{si el número de bolillas negras extraídas es impar} \end{cases}$$

Hallemos la función de probabilidad conjunta del vector (X, Y) . Observemos que los posibles valores de X son 0, 1, 2 y 3, y los posibles valores de Y son 1 y 0. Podemos resumir la información en una tabla de la forma siguiente:

		X			
		0	1	2	3
Y	0	1/30	0	15/30	0
	1	0	9/30	0	5/30

En efecto,

$p_{XY}(0,0) = P(X = 0, Y = 0)$ equivale al suceso “se extraen 3 bolillas negras” y por lo tanto tiene probabilidad $1/30$.

$p_{XY}(0,1) = P(X = 0, Y = 1)$ equivale al suceso “se extraen 3 bolillas negras y el número de bolillas negras es par” y por lo tanto tiene probabilidad 0.

De esta forma, se completa la tabla de probabilidades conjuntas.

2) Repetir el Ejemplo 1, suponiendo que las extracciones se realizan con reposición.

Definición: Sea (X, Y) un vector aleatorio discreto con función de probabilidad conjunta $p_{XY}(x, y)$, las **funciones de probabilidad marginal** de X e Y están dadas por

$$p_X(x) = \sum_y p_{XY}(x, y)$$

$$p_Y(y) = \sum_x p_{XY}(x, y)$$

Ejemplos: 1) En el ejemplo presentado antes, hallemos las funciones de probabilidad marginal. En primer lugar, hallemos $p_X(x)$.

$$p_X(0) = p_{XY}(0,0) + p_{XY}(0,1) = \frac{1}{30} + 0 = \frac{1}{30}$$

$$p_X(1) = p_{XY}(1,0) + p_{XY}(1,1) = 0 + \frac{9}{30} = \frac{9}{30}$$

$$p_X(2) = p_{XY}(2,0) + p_{XY}(2,1) = \frac{15}{30} + 0 = \frac{15}{30}$$

$$p_X(3) = p_{XY}(3,0) + p_{XY}(3,1) = 0 + \frac{5}{30} = \frac{5}{30}$$

Respecto a $p_Y(y)$,

$$p_Y(0) = p_{XY}(0,0) + p_{XY}(1,0) + p_{XY}(2,0) + p_{XY}(3,0) = \frac{1}{30} + 0 + \frac{15}{30} + 0 = \frac{16}{30}$$

$$p_Y(1) = p_{XY}(0,1) + p_{XY}(1,1) + p_{XY}(2,1) + p_{XY}(3,1) = 0 + \frac{9}{30} + 0 + \frac{5}{30} = \frac{14}{30}$$

Observemos que las funciones de probabilidad marginal se obtienen sumando sobre filas o columnas las funciones de probabilidad conjunta contenidas en la tabla, de ahí su nombre.

		X				$p_Y(y)$
		0	1	2	3	
Y	0	1/30	0	15/30	0	16/30
	1	0	9/30	0	5/30	14/30
$p_X(x)$		1/30	9/30	15/30	5/30	1

Definición: Sea (X, Y) un vector aleatorio discreto con función de probabilidad conjunta $p_{XY}(x, y)$, la **función de distribución acumulada conjunta** de (X, Y) está dada por

$$F_{XY}(x, y) = \sum_{s \leq x} \sum_{t \leq y} p_{XY}(s, t) \quad \forall (x, y) \in \mathbb{R}^2$$

Definición: Sean X e Y v.a. continuas definidas sobre un espacio muestral S . El vector aleatorio (X, Y) es continuo si existe una función, denominada **función de densidad conjunta**, $f_{XY}(x, y): \mathbb{R}^2 \rightarrow \mathbb{R}^{\geq 0}$, tal que

$$P((X, Y) \in A) = \iint_A f_{XY}(x, y) dx dy \quad \forall A \subseteq \mathbb{R}^2$$

En particular, si $A = [a, b] \times [c, d]$,

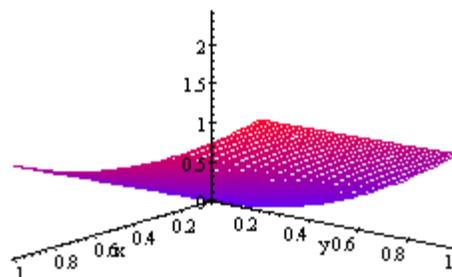
$$P((X, Y) \in A) = \int_a^b \int_c^d f_{XY}(x, y) dy dx.$$

Una función de densidad conjunta satisface:

- $f_{XY}(x, y) \geq 0 \quad \forall (x, y)$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$

Ejemplo: 1) Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta

$$f_{XY}(x, y) = \begin{cases} k(x + y^2) & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$



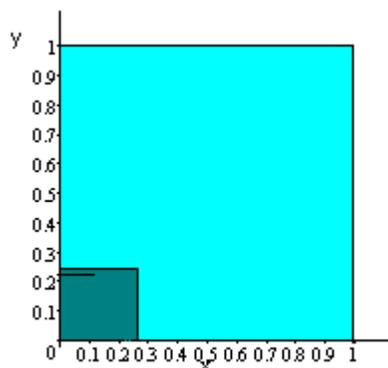
a) Hallar el valor de la constante k .

$$\begin{aligned}
 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = \int_0^1 \int_0^1 k(x + y^2) dx dy = k \int_0^1 \left(\int_0^1 (x + y^2) dx \right) dy = \\
 &= k \int_0^1 \left(\frac{x^2}{2} + xy^2 \right) \Big|_0^1 dy = k \int_0^1 \left(\frac{1}{2} + y^2 \right) dy = k \left(\frac{y}{2} + \frac{y^3}{3} \right) \Big|_0^1 = k \left(\frac{1}{2} + \frac{1}{3} \right) = k \frac{5}{6}
 \end{aligned}$$

y, por lo tanto, $k = \frac{6}{5}$.

b) Calcular $P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right)$.

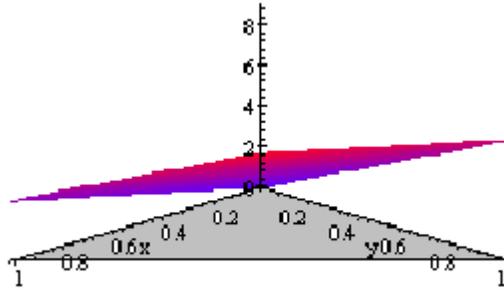
$$\begin{aligned}
 P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right) &= \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x + y^2) dx dy = \frac{6}{5} \int_0^{1/4} \left(\frac{x^2}{2} + xy^2 \right) \Big|_0^{1/4} dy = \\
 &= \frac{6}{5} \int_0^{1/4} \left(\frac{1}{16 \cdot 2} + \frac{1}{4} y^2 \right) dy = \frac{6}{5} \left(\frac{1}{32} y + \frac{1}{4} \frac{y^3}{3} \right) \Big|_0^{1/4} = \frac{6}{5} \left(\frac{1}{32 \cdot 4} + \frac{1}{4} \cdot \frac{1}{64 \cdot 3} \right) = \frac{6}{5} \left(\frac{1}{128} + \frac{1}{768} \right) = \\
 &= \frac{6}{5} \cdot \frac{7}{768} = \frac{7}{640}
 \end{aligned}$$



2) Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta

$$f_{XY}(x, y) = k(x + 2y)I_T(x, y),$$

siendo $T = \{(x, y) / 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}$.



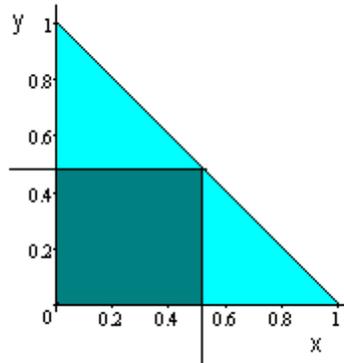
a) Hallar el valor de la constante k .

b) Hallar $P\left(X \leq \frac{1}{2}, Y \leq \frac{1}{2}\right)$.

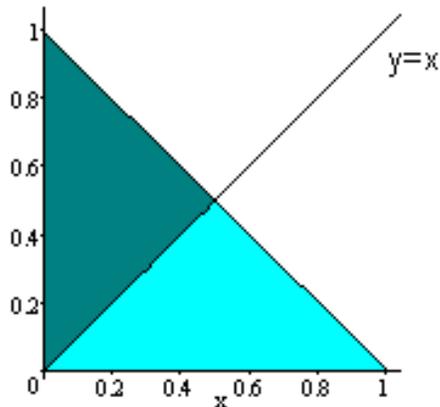
c) Hallar $P(X \leq Y)$.

$$\begin{aligned} \text{a) } 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = \int_0^1 \left(\int_0^{1-x} k(x + 2y) dy \right) dx = k \int_0^1 (xy + y^2) \Big|_0^{1-x} dx = \\ &= k \int_0^1 (x(1-x) + (1-x)^2) dx = k \int_0^1 (1-x) dx = k \left(x - \frac{x^2}{2} \right) \Big|_0^1 = k \frac{1}{2} \Rightarrow k = 2 \end{aligned}$$

$$\begin{aligned} \text{b) } P\left(X \leq \frac{1}{2}, Y \leq \frac{1}{2}\right) &= \int_0^{1/2} \int_0^{1/2} 2(x + 2y) dy dx = 2 \int_0^{1/2} (xy + y^2) \Big|_0^{1/2} dx = 2 \int_0^{1/2} \left(\frac{x}{2} + \frac{1}{4} \right) dx = \\ &= 2 \left(\frac{x^2}{4} + \frac{x}{4} \right) \Big|_0^{1/2} = 2 \left(\frac{1}{16} + \frac{1}{8} \right) = \frac{6}{16} = \frac{3}{8} \end{aligned}$$



$$\begin{aligned}
 \text{c) } P(X \leq Y) &= \int_0^{1/2} \left(\int_x^{1-x} 2(x+2y) dy \right) dx = 2 \int_0^{1/2} (xy + y^2) \Big|_x^{1-x} dx = \\
 &= 2 \left(\int_0^{1/2} x(1-x) + (1-x)^2 - x^2 - x^2 dx \right) = 2 \int_0^{1/2} (1-x-2x^2) dx = 2 \left(x - \frac{x^2}{2} - 2\frac{x^3}{3} \right) \Big|_0^{1/2} = \\
 &= 2 \left(\frac{1}{2} - \frac{1}{8} - \frac{1}{12} \right) = \frac{14}{24} = \frac{7}{12}
 \end{aligned}$$



3) En este ejemplo presentaremos a la distribución Uniforme sobre una región, la cual generaliza a la distribución Uniforme sobre un intervalo estudiada en el caso de variables aleatorias. Diremos que el vector aleatorio tiene **distribución Uniforme** sobre una región $A \subset \mathbb{R}^2$ si su densidad es constante sobre la región y 0 fuera de ella, es decir

$$(X, Y) \sim U(A) \Leftrightarrow f_{XY}(x, y) = \begin{cases} k & \text{si } (x, y) \in A \\ 0 & \text{si } (x, y) \notin A \end{cases}$$

Es inmediato verificar que $k = \frac{1}{\text{área}(A)}$, pues

$$1 = \iint_A k \, dx \, dy = k \iint_A dx \, dy = k \text{área}(A).$$

También es inmediato verificar que

$$P((X, Y) \in B) = \frac{\text{área}(A \cap B)}{\text{área}(A)} \quad \forall B \subset \mathfrak{R}^2.$$

Definición: Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta $f_{XY}(x, y)$, la **función de distribución acumulada conjunta** de (X, Y) está dada por

$$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(s, t) \, dt \, ds \quad \forall (x, y) \in \mathfrak{R}^2$$

Definición: Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta $f_{XY}(x, y)$, las **funciones de densidad marginal** de X e Y están dadas por

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx$$

Ejemplos: 1) Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta

$$f_{XY}(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Hallemos las funciones de densidad marginal.

Si $x \notin [0, 1]$, $f_X(x) = 0$ pues para esos valores de x la densidad conjunta $f_{XY}(x, y) = 0$.

Sea $x \in [0,1]$,

$$f_X(x) = \int_0^1 \frac{6}{5}(x + y^2) dy = \frac{6}{5} \left(xy + \frac{y^3}{3} \right) \Big|_0^1 = \frac{6}{5} \left(x + \frac{1}{3} \right).$$

Entonces, $f_X(x) = \frac{6}{5} \left(x + \frac{1}{3} \right) I_{[0,1]}(x)$.

Si $y \notin [0,1]$, $f_Y(y) = 0$ pues para esos valores de y la densidad conjunta $f_{XY}(x, y) = 0$.

Sea $y \in [0,1]$,

$$f_Y(y) = \int_0^1 \frac{6}{5}(x + y^2) dx = \frac{6}{5} \left(\frac{x^2}{2} + xy^2 \right) \Big|_0^1 = \frac{6}{5} \left(\frac{1}{2} + y^2 \right).$$

Entonces, $f_Y(y) = \frac{6}{5} \left(\frac{1}{2} + y^2 \right) I_{[0,1]}(y)$.

2) Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta

$$f_{XY}(x, y) = 2(x + 2y) I_T(x, y),$$

siendo $T = \{(x, y) / 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}$.

Si $x \notin [0,1]$, $f_X(x) = 0$ pues para esos valores de x la densidad conjunta $f_{XY}(x, y) = 0$.

Sea $x \in [0,1]$,

$$f_X(x) = \int_0^{1-x} 2(x + 2y) dy = 2 \left(xy + y^2 \right) \Big|_0^{1-x} = 2 \left(x(1-x) + (1-x)^2 \right) = 2(1-x).$$

Entonces, $f_X(x) = 2(1-x) I_{[0,1]}(x)$.

Si $y \notin [0,1]$, $f_Y(y) = 0$ pues para esos valores de y la densidad conjunta $f_{XY}(x, y) = 0$.

Sea $y \in [0,1]$,

$$f_Y(y) = \int_0^{1-y} 2(x+2y) dx = 2 \left(\frac{x^2}{2} + 2xy \right) \Big|_0^{1-y} = 2 \left(\frac{(1-y)^2}{2} + 2(1-y)y \right) = 1 + 2y - 3y^2.$$

Entonces, $f_Y(y) = (1 + 2y - 3y^2) I_{[0,1]}(y)$.

Definición: Sea (X, Y) un vector aleatorio discreto con función de probabilidad conjunta $p_{XY}(x, y)$ y marginales $p_X(x)$ y $p_Y(y)$, y sea x tal que $p_X(x) > 0$, la **función de probabilidad condicional** de Y dado $X = x$ está dada por

$$p_{Y|X=x}(y) = \frac{p_{XY}(x, y)}{p_X(x)}.$$

Del mismo modo, sea y tal que $p_Y(y) > 0$, la **función de probabilidad condicional** de X dado $Y = y$ está dada por

$$p_{X|Y=y}(x) = \frac{p_{XY}(x, y)}{p_Y(y)}.$$

Se puede verificar que, en efecto estas funciones son funciones de probabilidad ya que, por ejemplo, $p_{Y|X=x}(y)$ satisface

- $p_{Y|X=x}(y) \geq 0$ para todo y
- $\sum_y p_{Y|X=x}(y) = 1$

La primera condición se satisface ya que $p_X(x) > 0$ y $p_{XY}(x, y) \geq 0 \forall x, y$.

Respecto a la segunda,

$$\sum_y p_{Y|X=x}(y) = \sum_y \frac{p_{XY}(x, y)}{p_X(x)} = \frac{1}{p_X(x)} \sum_y p_{XY}(x, y) = \frac{1}{p_X(x)} p_X(x) = 1.$$

Ejemplo: Se arroja dos veces un tetraedro cuyas caras están numeradas 1, 2, 3 y 4. Se definen las variables aleatorias

X : "suma de los puntos obtenidos"

Y : "número de ases"

Hallemos en primer lugar la función de probabilidad conjunta de (X, Y) y las funciones de probabilidad marginal.

		X							$p_Y(y)$
		2	3	4	5	6	7	8	
Y	0	0	0	1/16	2/16	3/16	2/16	1/16	9/16
	1	0	2/16	2/16	2/16	0	0	0	6/16
	2	1/16	0	0	0	0	0	0	1/16
$p_X(x)$		1/16	2/16	3/16	4/16	3/16	2/16	1/16	1

Obtengamos, por ejemplo, la función de probabilidad condicional de Y, dado $X = 4$

$$p_{Y|X=4}(0) = \frac{p_{XY}(4,0)}{p_X(4)} = \frac{1/16}{3/16} = \frac{1}{3}$$

$$p_{Y|X=4}(1) = \frac{p_{XY}(4,1)}{p_X(4)} = \frac{2/16}{3/16} = \frac{2}{3}$$

$$p_{Y|X=4}(2) = \frac{p_{XY}(4,2)}{p_X(4)} = \frac{0}{3/16} = 0$$

que, podemos resumir en la siguiente tabla:

y	0	1	2
$p_{Y X=4}(y)$	1/3	2/3	0

En cambio, la función de probabilidad condicional de Y, dado $X = 3$, estará dada por

y	0	1	2
$p_{Y X=3}(y)$	0	1	0

De la misma forma, pueden obtenerse todas las funciones de probabilidad condicional de Y dado $X = x$, y las de X dado $Y = y$.

En cuanto al caso continuo, supongamos que en el Ejemplo 2) en el cual la densidad conjunta estaba dada por

$$f_{XY}(x, y) = 2(x + 2y) I_T(x, y),$$

siendo $T = \{(x, y) / 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}$, deseamos hallar $P\left(X \leq \frac{1}{2} \mid Y \leq \frac{1}{4}\right)$.

$$P\left(X \leq \frac{1}{2} \mid Y \leq \frac{1}{4}\right) = \frac{P\left(X \leq \frac{1}{2}, Y \leq \frac{1}{4}\right)}{P\left(Y \leq \frac{1}{4}\right)}$$

Por un lado,

$$\begin{aligned} P\left(X \leq \frac{1}{2}, Y \leq \frac{1}{4}\right) &= \int_0^{1/4} \int_0^{1/2} 2(x+2y) dx dy = 2 \int_0^{1/4} \left(\frac{x^2}{2} + 2xy\right) \Big|_0^{1/2} dy = \\ &= 2 \int_0^{1/4} \left(\frac{1}{8} + y\right) dy = 2 \left(\frac{1}{8}y + \frac{y^2}{2}\right) \Big|_0^{1/4} = 2 \left(\frac{1}{32} + \frac{1}{32}\right) = \frac{1}{8} \end{aligned}$$

y, por otro

$$P\left(Y \leq \frac{1}{4}\right) = \int_0^{1/4} (1+2y-3y^2) dy = (y+y^2-y^3) \Big|_0^{1/4} = \frac{1}{4} + \frac{1}{16} - \frac{1}{64} = \frac{19}{64}.$$

Entonces,

$$P\left(X \leq \frac{1}{2} \mid Y \leq \frac{1}{4}\right) = \frac{1/8}{19/64} = \frac{8}{19}.$$

¿Cómo calcularíamos $P\left(X \leq \frac{1}{2} \mid Y = \frac{1}{4}\right)$? Ahora no es aplicable directamente la definición de probabilidad condicional porque $P(Y=y)=0 \forall y$. Se requiere la siguiente definición.

Definición: Sea (X, Y) un vector aleatorio continuo con función de densidad conjunta $f_{XY}(x, y)$ y marginales $f_X(x)$ y $f_Y(y)$, y sea x tal que $f_X(x) > 0$, la **función de densidad condicional** de Y dado $X = x$ está dada por

$$f_{Y|X=x}(y) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

Del mismo modo, sea y tal que $f_Y(y) > 0$, la **función de densidad condicional** de X dado $Y = y$ está dada por

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)}.$$

Se puede verificar que, en efecto estas funciones son funciones de densidad ya que, por ejemplo, $f_{Y|X=x}(y)$ satisface

- $f_{Y|X=x}(y) \geq 0$ para todo y

- $\int_{-\infty}^{\infty} f_{Y|X=x}(y) dy = 1$

La primera condición se satisface ya que $f_X(x) > 0$ y $f_{XY}(x, y) \geq 0 \forall x, y$.

Respecto a la segunda,

$$\int_{-\infty}^{\infty} f_{Y|X=x}(y) dy = \int_{-\infty}^{\infty} \frac{f_{XY}(x, y)}{f_X(x)} dy = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{XY}(x, y) dy = \frac{1}{f_X(x)} f_X(x) = 1.$$

Ejemplo: Volviendo al ejemplo 2 y a la pregunta que motivó esta definición,

$$P\left(X \leq \frac{1}{2} \mid Y = \frac{1}{4}\right) = \int_0^{1/2} f_{X|Y=1/4}(x) dx$$

Hallemos la densidad condicional de X , dado $Y=1/4$.

$$f_{X|Y=1/4}(x) = \frac{f_{XY}(x, 1/4)}{f_Y(1/4)} = \frac{2(x + 2/4) I_{(0, 3/4)}(x)}{1 + \frac{1}{2} - \frac{3}{16}} = \frac{32}{21} \left(x + \frac{1}{2}\right) I_{(0, 3/4)}(x).$$

Notemos que, dado $Y = y$, X toma valores en el intervalo $(0, 1-y)$. De ahí que, como $Y = 1/4$, X toma valores en el intervalo $(0, 3/4)$. Finalmente,

$$P\left(X \leq \frac{1}{2} \mid Y = \frac{1}{4}\right) = \int_0^{1/2} \frac{32}{21} \left(x + \frac{1}{2}\right) dx = \frac{32}{21} \left(\frac{x^2}{2} + \frac{x}{2}\right) \Big|_0^{1/2} = \frac{32}{21} \left(\frac{1}{8} + \frac{1}{4}\right) = \frac{4}{7}.$$

Independencia de variables aleatorias

Definición: Las variables aleatorias X e Y son independientes si y sólo si para todo $a < b$ y $c < d$ se satisface

$$P(\{a < X < b\} \cap \{c < Y < d\}) = P(a < X < b) P(c < Y < d)$$

Si esta condición no se satisface, diremos que X e Y son dependientes.

Caso 1: Si el vector (X, Y) es **discreto**, la condición de independencia es equivalente a la siguiente: X e Y son independientes si y sólo si

$$p_{XY}(x, y) = p_X(x) p_Y(y) \quad \forall (x, y) \in \mathbb{R}^2$$

Luego, para probar que dos variables discretas no son independientes, es suficiente con exhibir un punto (x_0, y_0) en el que $p_{XY}(x_0, y_0) \neq p_X(x_0) p_Y(y_0)$.

Caso 2: Si el vector (X, Y) es **continuo** y

$$f_{XY}(x, y) = f_X(x) f_Y(y) \quad \forall (x, y) \in \mathbb{R}^2$$

entonces, claramente, X e Y son independientes.

Para probar que dos variables continuas no son independientes deberíamos exhibir un conjunto $(a, b) \times (c, d) \subseteq \mathbb{R}^2$ (es decir un conjunto de medida no nula) en el que no se satisfaga la condición $f_{XY}(x, y) = f_X(x) f_Y(y)$.

Se denomina soporte de una densidad al conjunto de valores en los cuales la densidad es positiva. Si el soporte de la densidad conjunta no es igual al producto cartesiano de los soportes de las densidades de X e Y es inmediato encontrar un conjunto así: bastaría con exhibir un rectángulo $(a_1, b_1) \times (c_1, d_1)$ tal que el intervalo (a_1, b_1) esté contenido en el soporte de X , el intervalo (c_1, d_1) en el soporte de Y y el rectángulo $(a_1, b_1) \times (c_1, d_1)$ no esté contenido en el soporte de (X, Y) .

Otra forma de probar que X e Y no son independientes es encontrar un punto (x_0, y_0) en el cual $f_{XY}(x_0, y_0) \neq f_X(x_0) f_Y(y_0)$ y en el cual todas las densidades sean continuas. Por continuidad, la condición se cumplirá en un entorno rectangular del punto.

Observemos que si X e Y son independientes, las funciones de probabilidad o densidad condicional coinciden con las correspondientes marginales.

Ejemplos: 1) Consideremos el primer ejemplo presentado para el caso discreto, cuya función de probabilidad conjunta y sus funciones de probabilidad marginal están dadas por:

		X				$p_Y(y)$
		0	1	2	3	
Y	0	1/30	0	15/30	0	16/30
	1	0	9/30	0	5/30	14/30
$p_X(x)$		1/30	9/30	15/30	5/30	1

Claramente X e Y no son independientes ya que, por ejemplo,

$$p_{XY}(0,1) = 0 \neq \frac{1}{30} \cdot \frac{14}{30} = p_X(0)p_Y(1).$$

2) Sean X e Y v.a. independientes con distribución exponencial de parámetro λ , entonces la función de densidad conjunta del vector (X, Y) estará dada por

$$\begin{aligned} f_{XY}(x, y) &= f_X(x) f_Y(y) = \lambda e^{-\lambda x} \lambda e^{-\lambda y} I_{(0,\infty)}(x) I_{(0,\infty)}(y) = \\ &= \lambda^2 e^{-\lambda(x+y)} I_{(0,\infty)}(x) I_{(0,\infty)}(y). \end{aligned}$$

Esperanza de una función de dos variables aleatorias

Hemos visto que, dada una v.a. X y una función real h , $h(X)$ también es una v.a. y que para calcular su esperanza no necesitamos hallar la distribución de $h(X)$ ya que se obtiene a partir de la función de probabilidad puntual o de densidad de la v.a. X , según sea ésta discreta o continua, en la forma

$$E(h(X)) = \sum_x h(x) p_X(x) \quad \text{ó} \quad E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

Un resultado similar se obtiene en el caso de una función real de un vector aleatorio y está dado por las dos proposiciones siguientes, cuya demostración no haremos.

Proposición: Sean X e Y dos variables aleatorias discretas con función de probabilidad conjunta $p_{XY}(x, y)$ y sea $h(x, y): \mathfrak{R}^2 \rightarrow \mathfrak{R}$, entonces $h(X, Y)$ es una variable aleatoria y

$$E(h(X, Y)) = \sum_x \sum_y h(x, y) p_{XY}(x, y)$$

siempre que esta esperanza exista.

Proposición: Sean X e Y dos variables aleatorias continuas con función de densidad conjunta $f_{XY}(x, y)$ y sea $h(x, y): \mathfrak{R}^2 \rightarrow \mathfrak{R}$, entonces $h(X, Y)$ es una variable aleatoria y

$$E(h(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) dx dy$$

siempre que esta esperanza exista.

Proposición: Sean X e Y dos v.a. discretas o continuas con función de probabilidad conjunta o de densidad $p_{XY}(x, y)$ ó $f_{XY}(x, y)$ respectivamente y sean a y b números reales, entonces

$$E(aX + bY) = aE(X) + bE(Y)$$

Dem: Haremos la demostración para el caso continuo. La demostración para el caso discreto es similar.

Sea $h(X, Y) = aX + bY$, entonces

$$\begin{aligned} E(h(X, Y)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{XY}(x, y) dx dy = \\ &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy = \\ &= a \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{XY}(x, y) dy \right) dx + b \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{XY}(x, y) dx \right) dy = \\ &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} y f_Y(y) dy = aE(X) + bE(Y) \end{aligned}$$

como queríamos demostrar.

Proposición: Si X e Y son v.a. independientes, $E(XY) = E(X)E(Y)$.

Dem: Ejercicio.

Covarianza y correlación

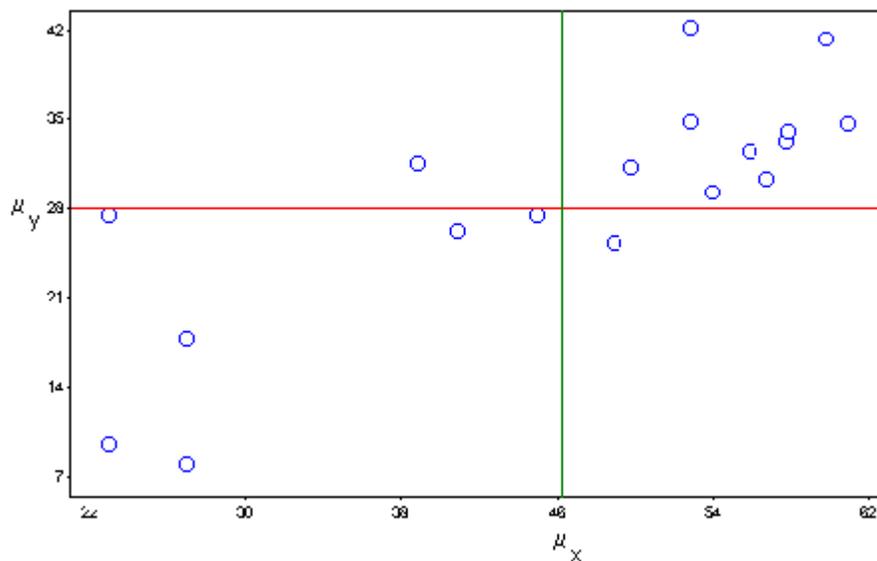
Definición: Sean X e Y dos v.a. con esperanzas μ_X y μ_Y respectivamente, la **covarianza** entre X e Y se define como

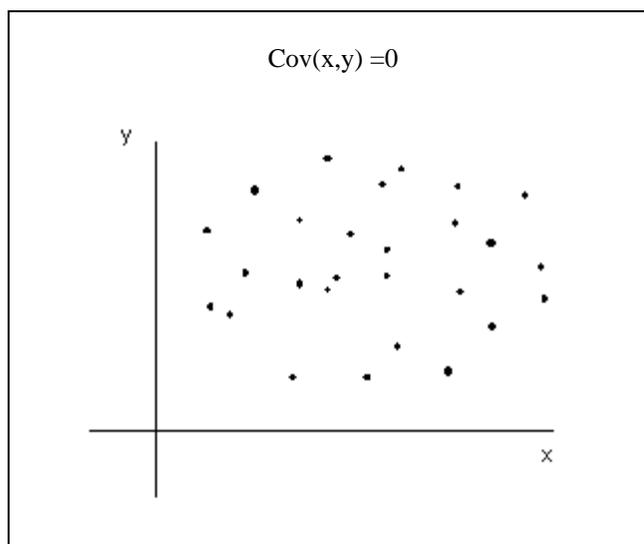
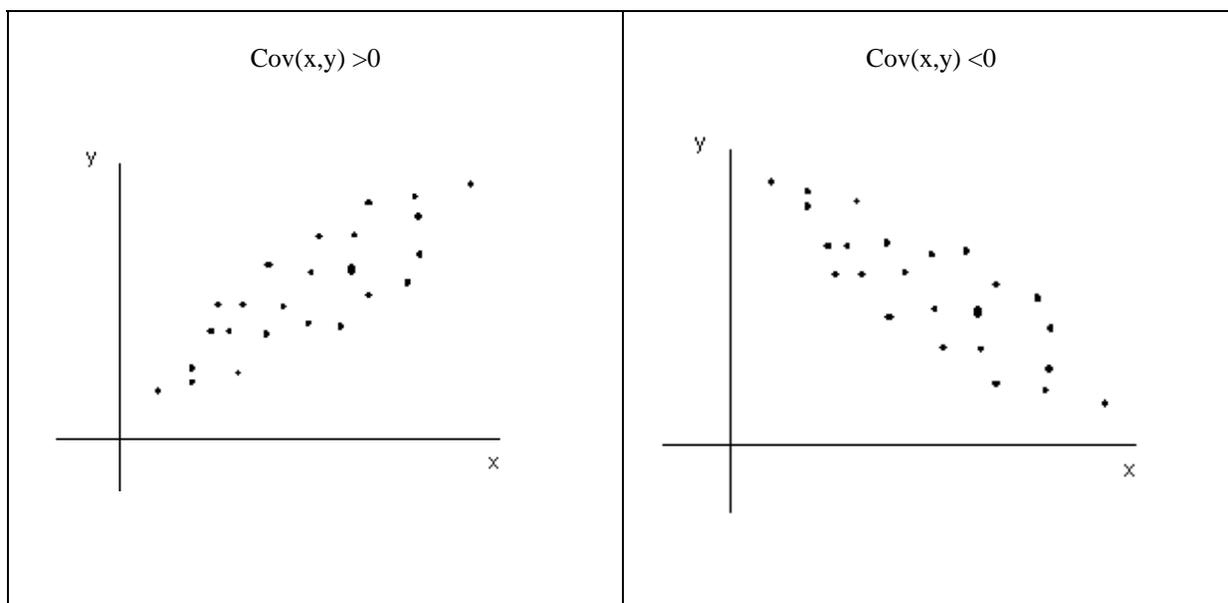
$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p_{XY}(x, y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x, y) dx dy \end{cases}$$

según sean X e Y discretas o continuas.

Observación: $\text{Cov}(X, X) = V(X)$.

Idea intuitiva: Si X e Y tienen una fuerte relación positiva, en el sentido que valores grandes de X aparecen asociados con valores grandes de Y y valores pequeños de X aparecen asociados con valores pequeños de Y , entonces la mayoría de los productos $(x - \mu_X)(y - \mu_Y)$ serán positivos y por lo tanto la covarianza será positiva. Por otra parte, si X e Y tienen una fuerte relación negativa, en el sentido que valores grandes de X aparecen asociados con valores pequeños de Y y valores pequeños de X aparecen asociados con valores grandes de Y , entonces la mayoría de los productos $(x - \mu_X)(y - \mu_Y)$ serán negativos y por lo tanto la covarianza será negativa.





Proposición: $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

Dem: Lo haremos sólo para el caso discreto. Para el caso continuo se demuestra en forma similar. Denotemos $E(X) = \mu_x$ y $E(Y) = \mu_y$,

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = \sum_x \sum_y (x - \mu_x)(y - \mu_y) p_{XY}(x, y) =$$

$$\begin{aligned}
 &= \sum_x \sum_y (xy - x\mu_Y - y\mu_X + \mu_X \mu_Y) p_{XY}(x, y) = \\
 &= \sum_x \sum_y xy p_{XY}(x, y) - \mu_Y \sum_x \sum_y x p_{XY}(x, y) - \mu_X \sum_x \sum_y y p_{XY}(x, y) + \mu_X \mu_Y \sum_x \sum_y p_{XY}(x, y) = \\
 &= E(XY) - \mu_Y \sum_x x \sum_y p_{XY}(x, y) - \mu_X \sum_y y \sum_x p_{XY}(x, y) + \mu_X \mu_Y = \\
 &= E(XY) - \mu_Y \sum_x x p_X(x) - \mu_X \sum_y y p_Y(y) + \mu_X \mu_Y = \\
 &= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y
 \end{aligned}$$

como queríamos demostrar.

Ejemplos: 1) Consideremos nuevamente el primer ejemplo presentado para el caso discreto, cuya función de probabilidad conjunta y sus funciones de probabilidad marginal están dadas por:

		X				$p_Y(y)$
		0	1	2	3	
Y	0	1/30	0	15/30	0	16/30
	1	0	9/30	0	5/30	14/30
$p_X(x)$		1/30	9/30	15/30	5/30	1

y calculemos $Cov(X, Y)$.

$$\begin{aligned}
 Cov(X, Y) &= E(XY) - E(X)E(Y) = \sum_{k=0}^3 \sum_{j=0}^1 k j p_{XY}(k, j) - \left(\sum_{k=0}^3 k p_X(k) \right) \left(\sum_{i=0}^1 i p_Y(i) \right) \\
 &= 1 \cdot \frac{9}{30} + 3 \cdot \frac{5}{30} - \left(1 \cdot \frac{9}{30} + 2 \cdot \frac{15}{30} + 3 \cdot \frac{5}{30} \right) \left(1 \cdot \frac{14}{30} \right) = \frac{24}{30} - \frac{54}{30} \cdot \frac{14}{30} = -\frac{4}{100}
 \end{aligned}$$

2) Consideremos nuevamente el primer ejemplo presentado para el caso continuo, es decir un vector aleatorio (X, Y) con función de densidad conjunta

$$f_{XY}(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

y marginales $f_X(x) = \frac{6}{5} \left(x + \frac{1}{3}\right) I_{[0,1]}(x)$ y $f_Y(y) = \frac{6}{5} \left(\frac{1}{2} + y^2\right) I_{[0,1]}(y)$.

Calculemos $\text{Cov}(X, Y)$. En primer lugar,

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^1 xy \frac{6}{5} (x + y^2) dx dy = \frac{6}{5} \int_0^1 \int_0^1 (x^2 y + xy^3) dx dy = \\ &= \frac{6}{5} \int_0^1 \left(\frac{x^3 y}{3} + \frac{x^2 y^3}{2} \right) \Big|_0^1 dy = \frac{6}{5} \int_0^1 \left(\frac{y}{3} + \frac{y^3}{2} \right) dy = \frac{6}{5} \left(\frac{y^2}{6} + \frac{y^4}{8} \right) \Big|_0^1 = \\ &= \frac{6}{5} \left(\frac{1}{6} + \frac{1}{8} \right) = \frac{6}{5} \cdot \frac{7}{24} = \frac{7}{20} \end{aligned}$$

Por otra parte,

$$E(X) = \int_0^1 x \frac{6}{5} \left(x + \frac{1}{3}\right) dx = \frac{6}{5} \int_0^1 \left(x^2 + \frac{x}{3}\right) dx = \frac{6}{5} \left(\frac{x^3}{3} + \frac{x^2}{6} \right) \Big|_0^1 = \frac{6}{5} \cdot \frac{1}{2} = \frac{3}{5}$$

$$E(Y) = \int_0^1 y \frac{6}{5} \left(\frac{1}{2} + y^2\right) dy = \frac{6}{5} \int_0^1 \left(\frac{y}{2} + y^3\right) dy = \frac{6}{5} \left(\frac{y^2}{4} + \frac{y^4}{4} \right) \Big|_0^1 = \frac{6}{5} \cdot \frac{1}{2} = \frac{3}{5}$$

Entonces,

$$\text{Cov}(X, Y) = \frac{7}{20} - \frac{3}{5} \cdot \frac{3}{5} = -\frac{1}{100}.$$

Propiedad: Si X e Y son v.a. independientes, $\text{Cov}(X, Y) = 0$. La recíproca no es cierta en general.

Dem: Hemos visto que si X e Y son independientes, $E(XY) = E(X)E(Y)$ y por lo tanto es inmediato que $\text{Cov}(X, Y) = 0$.

Para ejemplificar que la recíproca no es en general cierta, consideremos un vector aleatorio discreto con la siguiente función de probabilidad conjunta

		X					$p_Y(y)$
		0	1	2	3	4	
Y	0	1/5	0	0	0	1/5	2/5
	3	0	1/5	0	1/5	0	2/5
	4	0	0	1/5	0	0	1/5

$p_X(x)$	1/5	1/5	1/5	1/5	1/5	1
----------	-----	-----	-----	-----	-----	---

Se observa que X e Y no son independientes ya que, por ejemplo,

$$p_{XY}(2,3) = 0 \neq p_X(2) p_Y(3) = \frac{1}{5} \cdot \frac{2}{5}$$

Sin embargo, se puede verificar que $\text{Cov}(X, Y) = 0$. En efecto,

$$E(XY) = 1 \cdot 3 \cdot \frac{1}{5} + 2 \cdot 4 \cdot \frac{1}{5} + 3 \cdot 3 \cdot \frac{1}{5} = 4$$

$$E(X) = 0 \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} = 2$$

$$E(Y) = 0 \cdot \frac{2}{5} + 3 \cdot \frac{2}{5} + 4 \cdot \frac{1}{5} = 2$$

Entonces, $\text{Cov}(X, Y) = 4 - 2 \cdot 2 = 0$.

Observación: La covarianza depende de las unidades en que se expresan las variables aleatorias. Este inconveniente puede salvarse standarizándolas. De este modo se obtiene una medida de la fuerza de la relación entre las v.a. que no depende de sus unidades.

Definición: Sean X e Y dos v.a. con esperanzas μ_X y μ_Y respectivamente y varianza positiva, el **coeficiente de correlación** entre X e Y se define como

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

siendo σ_X y σ_Y los desvíos standard de X e Y respectivamente.

Proposición: 1) Sean a, b, c y d números reales, $a \neq 0$, $c \neq 0$ y X e Y dos v.a. cualesquiera con varianza positiva, entonces

$$\rho(aX + b, cY + d) = \text{sg}(ac) \rho(X, Y)$$

donde sg denota la función signo.

2) $-1 \leq \rho(X, Y) \leq 1$

3) $|\rho(X, Y)| = 1 \Leftrightarrow Y = aX + b$ con probabilidad 1, para ciertos valores reales a y b , $a \neq 0$. Observemos que el coeficiente de correlación mide *relación lineal* entre las v.a.

Dem: 1)

$$\begin{aligned}\text{Cov}(aX + b, cY + d) &= E[(aX + b)(cY + d)] - E(aX + b) E(cY + d) = \\ &= E[acXY + adX + bcY + bd] - (aE(X) + b)(cE(Y) + d) = \\ &= acE(XY) + adE(X) + bcE(Y) + bd - [acE(X)E(Y) + adE(X) + bcE(Y) + bd] = \\ &= ac[E(XY) - E(X)E(Y)] = ac \text{Cov}(X, Y).\end{aligned}$$

Por otra parte, $\sigma_{aX+b} = |a| \sigma_X$ y $\sigma_{cY+d} = |c| \sigma_Y$ y, por lo tanto

$$\rho(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{\sigma_{aX+b} \sigma_{cY+d}} = \frac{ac \text{Cov}(X, Y)}{|a||c| \sigma_X \sigma_Y} = \text{sg}(ac) \rho(X, Y)$$

como queríamos demostrar.

2) Consideremos la siguiente función real,

$$q(t) = E[(Y - \mu_Y) - t(X - \mu_X)]^2 = E[V - tW]^2$$

siendo $V = Y - \mu_Y$ y $W = X - \mu_X$.

Observemos que $q(t) \geq 0 \forall t$.

Como

$$q(t) = E[V - tW]^2 = E(V^2) - 2t E(VW) + t^2 E(W^2)$$

es una función cuadrática en t que toma valores mayores o iguales que 0, su gráfico, o no corta al eje t o lo corta en un solo punto. Es decir que la ecuación $q(t) = 0$ tiene a lo sumo una raíz y por lo tanto su discriminante es menor o igual que 0. (Recordemos que el discriminante de una ecuación de segundo grado $ax^2 + bx + c = 0$ es $b^2 - 4ac$). En nuestro caso, el discriminante es

$$4[E(VW)]^2 - 4E(V^2)E(W^2)$$

y, por lo tanto,

$$4[E(VW)]^2 - 4E(V^2)E(W^2) \leq 0 \Leftrightarrow \frac{[E(VW)]^2}{E(V^2)E(W^2)} \leq 1 \Leftrightarrow \frac{[E((X - \mu_X)(Y - \mu_Y))]^2}{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]} \leq 1$$

$$\Leftrightarrow [\rho(X, Y)]^2 \leq 1 \Leftrightarrow -1 \leq \rho(X, Y) \leq 1.$$

3) Demostraremos las dos implicaciones.

(\Rightarrow) Si $\rho^2(X, Y) = 1$, y volviendo a la demostración de la propiedad anterior, existe t_o tal que $q(t_o) = 0$, o sea tal que

$$E[V - t_o W]^2 = 0,$$

Pero además $E(V - t_o W) = 0$, pues V y W tienen esperanza igual a 0. Entonces la v.a. $V - t_o W$ tiene varianza cero y por lo tanto es constante con probabilidad 1, es decir

$$P(V - t_o W = E(V - t_o W)) = P(V - t_o W = 0) = 1$$

o sea,

$$P((Y - \mu_Y) - t_o (X - \mu_X) = 0) = 1 \Leftrightarrow P(Y = t_o X + \mu_Y - t_o \mu_X) = 1.$$

Entonces, $Y = aX + b$ con probabilidad 1, siendo $a = t_o$ y $b = \mu_Y - t_o \mu_X$. Falta verificar que $a = t_o \neq 0$.

En efecto, si t_o fuese igual a 0, ésto implicaría que $E(V^2) = \text{Var}(Y) = 0$.

(\Leftarrow) Sea $Y = aX + b$ para ciertos valores $a \neq 0$ y b . Entonces

$$\begin{aligned} \rho(X, Y) = \rho(X, aX + b) &= \frac{\text{Cov}(X, aX + b)}{\sigma_X \sigma_{aX + b}} = \frac{E(X(aX + b)) - E(X)E(aX + b)}{\sigma_X |a| \sigma_X} = \\ &= \frac{aE(X^2) + bE(X) - a[E(X)]^2 - bE(X)}{|a| \sigma_X^2} = \frac{a(E(X^2) - E^2(X))}{|a| \sigma_X^2} = \frac{a \sigma_X^2}{|a| \sigma_X^2} = \pm 1 \end{aligned}$$

como queríamos demostrar.

Vectores aleatorios.

Extensión a más de dos dimensiones

Definición: Sean X_1, \dots, X_k variables aleatorias discretas, la **función de probabilidad conjunta** del vector aleatorio (X_1, \dots, X_k) se define como:

$$p_{X_1, \dots, X_k}(x_1, \dots, x_k) = P(X_1 = x_1, \dots, X_k = x_k)$$

y, dado cualquier conjunto $A \subseteq \mathfrak{R}^k$,

$$P((X_1, \dots, X_k) \in A) = \sum_{(x_1, \dots, x_k) \in A} \dots \sum p_{X_1, \dots, X_k}(x_1, \dots, x_k)$$

Esta función satisface las siguientes propiedades:

- $p_{X_1, \dots, X_k}(x_1, \dots, x_k) \geq 0 \quad \forall (x_1, \dots, x_k)$
- $\sum_{x_1} \dots \sum_{x_k} p_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k) = 1$

En forma similar a lo hecho para el caso bidimensional se pueden definir las **funciones de probabilidad marginal**. Por ejemplo, la función de probabilidad marginal de X_1 está dada por:

$$p_{X_1}(x_1) = \sum_{x_2} \dots \sum_{x_k} p_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k)$$

y la función de probabilidad marginal de (X_1, X_2) está dada por:

$$p_{X_1, X_2}(x_1, x_2) = \sum_{x_3} \dots \sum_{x_k} p_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k).$$

Distribución multinomial: Es una generalización de la distribución Binomial. Supongamos que se repite n veces en forma independiente una experiencia, que en cada repetición hay k resultados posibles ($k \geq 2$), cada uno de los cuales ocurre con probabilidad p_i ($1 \leq i \leq k$) y que estas probabilidades se mantienen constantes en todas las repeticiones. Este experimento se denomina experimento multinomial. Si definimos

X_i : número de veces que ocurre el resultado i ($1 \leq i \leq k$)

la distribución conjunta de (X_1, \dots, X_k) se denomina distribución multinomial de parámetros n, p_1, \dots, p_k .

Notación: $(X_1, \dots, X_k) \sim M(n, p_1, \dots, p_k)$

La correspondiente función de probabilidad conjunta está dada por

$$p_{X_1, \dots, X_k}(x_1, \dots, x_k) = \begin{cases} \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} & \text{si } 0 \leq x_i \leq n \forall i, \sum_{i=1}^k x_i = n \\ 0 & \text{en otro caso} \end{cases} \quad (1)$$

En efecto, en primer lugar hay que notar que si $x_1 + x_2 + \dots + x_k \neq n$, la función de probabilidad puntual es cero. Sean ahora $0 \leq x_i \leq n$, tales que $x_1 + x_2 + \dots + x_k = n$. Indicando por R_i ($1 \leq i \leq k$) cada uno de los k resultados posibles, una de las posibles configuraciones que producen x_i resultados R_i ($1 \leq i \leq k$), es

$$\underbrace{R_1 \dots R_1}_{x_1} \underbrace{R_2 \dots R_2}_{x_2} \dots \underbrace{R_k \dots R_k}_{x_k}$$

(alguno de los x_i 's podría ser 0, en cuyo caso no aparecería ninguno de los correspondientes R_i).

Como hemos supuesto independencia entre las repeticiones, esa configuración tiene probabilidad $p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$, pero es sólo una de las configuraciones posibles que producen x_i resultados R_i para $1 \leq i \leq k$.

¿Cuántas configuraciones diferentes hay?

$$\begin{aligned} \binom{n}{x_1} \cdot \binom{n-x_1}{x_2} \cdot \binom{n-x_1-x_2}{x_3} \dots \binom{x_k}{x_k} &= \frac{n!}{x_1!(n-x_1)!} \cdot \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \dots \frac{x_k!}{x_k!0!} \\ &= \frac{n!}{x_1! x_2! \dots x_k!} \end{aligned}$$

y se obtiene la función de probabilidad dada en (1).

Observación: La distribución marginal de X_i es binomial de parámetros n y p_i para todo $1 \leq i \leq k$. En general, las marginales de una distribución multinomial son binomiales o multinomiales.

Ejemplo: De una urna que contiene 3 bolillas rojas, 2 negras, 4 azules y 1 blanca se extraen 12 bolillas *con reposición*. Definiendo

X_1 : número de bolillas rojas
 X_2 : número de bolillas negras
 X_3 : número de bolillas azules
 X_4 : número de bolillas blancas

el vector (X_1, X_2, X_3, X_4) tiene distribución multinomial, es decir

$$(X_1, X_2, X_3, X_4) \sim M\left(12, \frac{3}{10}, \frac{2}{10}, \frac{4}{10}, \frac{1}{10}\right)$$

a) ¿Cuál es la probabilidad de que se obtengan 3 bolillas rojas, 5 negras, 4 azules y ninguna blanca?

$$p_{X_1, X_2, X_3, X_4}(3, 5, 4, 0) = \frac{12!}{3!5!4!0!} \left(\frac{3}{10}\right)^3 \left(\frac{2}{10}\right)^5 \left(\frac{4}{10}\right)^4 \left(\frac{1}{10}\right)^0 = 0.006$$

b) Calcular la probabilidad de obtener a lo sumo dos bolillas rojas.

Como $X_1 \sim \text{Bi}\left(12, \frac{3}{10}\right)$, entonces

$$P(X_1 \leq 2) = \sum_{i=0}^2 p_{X_1}(i) = \sum_{i=0}^2 \binom{12}{i} \left(\frac{3}{10}\right)^i \left(\frac{7}{10}\right)^{12-i} = 0.25$$

c) Calcular la probabilidad de obtener 3 bolillas rojas y 2 blancas.

Como las v.a. que nos interesan son X_1 y X_4 , defino una nueva v.a. $Y = X_2 + X_3$. El vector aleatorio (X_1, X_4, Y) también tendrá distribución multinomial.

$$(X_1, X_4, Y) \sim M\left(12, \frac{3}{10}, \frac{1}{10}, \frac{6}{10}\right)$$

y, por lo tanto, la probabilidad pedida será

$$p_{X_1, X_4, Y}(3, 2, 7) = \frac{12!}{3!2!7!} \left(\frac{3}{10}\right)^3 \left(\frac{1}{10}\right)^2 \left(\frac{6}{10}\right)^7 = 0.06$$

Definición: El vector aleatorio (X_1, \dots, X_k) es continuo si existe una función $f_{X_1, \dots, X_k} : \mathfrak{R}^k \rightarrow \mathfrak{R}^{\geq 0}$, denominada **función de densidad conjunta**, tal que

$$P((X_1, \dots, X_k) \in A) = \int_A \dots \int f_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k) dx_1 \dots dx_k \quad \forall A \subseteq \mathfrak{R}^k$$

Esta función satisface las siguientes propiedades:

- $f_{X_1, \dots, X_k}(x_1, \dots, x_k) \geq 0 \quad \forall (x_1, \dots, x_k)$
- $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, \dots, X_k}(x_1, \dots, x_k) dx_1 \dots dx_k = 1$

En forma similar a lo hecho para el caso bidimensional se pueden definir las **funciones de densidad marginal**. Por ejemplo, la función de densidad marginal de X_1 está dada por:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k) dx_2 \dots dx_k$$

y la función de densidad marginal de (X_1, X_2) , está dada por:

$$f_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, \dots, X_k}(x_1, x_2, \dots, x_k) dx_3 \dots dx_k$$

Definición: X_1, \dots, X_k son variables aleatorias **independientes** si y sólo si

$$p_{X_1, \dots, X_k}(x_1, \dots, x_k) = p_{X_1}(x_1) \dots p_{X_k}(x_k) \quad \forall (x_1, \dots, x_k) \quad \text{en el caso discreto}$$

$f_{X_1, \dots, X_k}(x_1, \dots, x_k) = f_{X_1}(x_1) \dots f_{X_k}(x_k) \quad \forall (x_1, \dots, x_k)$ salvo, eventualmente, en un conjunto de probabilidad cero en el caso continuo.

Ejemplos: 1) En el caso de la distribución multinomial, las componentes del vector aleatorio son v.a. con distribución binomial no independientes y ésto que es intuitivo ya que su suma es constante (es igual a n), puede verificarse aplicando la definición.

2) Sea (X_1, X_2, X_3) un vector aleatorio con distribución uniforme en el prisma de vértices $(0,0,0), (1,0,0), (0,2,0), (1,2,0), (0,0,3), (1,0,3), (0,2,3), (1,2,3)$, cuyo volumen es igual a 6. Entonces, su función de densidad conjunta dada por

$$f_{X_1, X_2, X_3}(x_1, x_2, x_3) = \begin{cases} 1/6 & \text{si } 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 2, 0 \leq x_3 \leq 3 \\ 0 & \text{en otro caso} \end{cases}$$

Es inmediato verificar que las componentes del vector son variables aleatorias independientes, ya que

$$f_{X_1}(x_1) = \begin{cases} \int_0^3 \int_0^2 \frac{1}{6} dx_2 dx_3 = \frac{1}{6} \cdot 6 = 1 & \text{si } x_1 \in [0,1] \\ 0 & \text{si } x_1 \notin [0,1] \end{cases}$$

$$f_{X_2}(x_2) = \begin{cases} \int_0^3 \int_0^1 \frac{1}{6} dx_1 dx_3 = \frac{1}{6} \cdot 3 = \frac{1}{2} & \text{si } x_2 \in [0,2] \\ 0 & \text{si } x_2 \notin [0,2] \end{cases}$$

$$f_{X_3}(x_3) = \begin{cases} \int_0^2 \int_0^1 \frac{1}{6} dx_1 dx_2 = \frac{1}{6} \cdot 2 = \frac{1}{3} & \text{si } x_3 \in [0,3] \\ 0 & \text{si } x_3 \notin [0,3] \end{cases}$$

entonces,

$$f_{X_1, X_2, X_3}(x_1, x_2, x_3) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3) \quad \forall (x_1, x_2, x_3)$$

Distribución de la suma de dos variables aleatorias

Sean X e Y dos v.a. de las cuáles se conoce la distribución conjunta. Estamos interesados en la distribución de la v.a. $V = X + Y$.

Consideraremos dos ejemplos, uno para el caso de un vector aleatorio discreto y otro para el caso continuo.

Ejemplos: 1) Sean $X \sim P(\lambda)$ e $Y \sim P(\mu)$, v.a. independientes, y sea $V = X + Y$. Claramente el recorrido de la v.a. V es el conjunto $R_V = \{0,1,2,\dots\}$. Sea $k \in R_V$,

$$P(X + Y = k) = \sum_{i=0}^{\infty} p_{XY}(i, k-i) = \sum_{i=0}^k p_X(i) p_Y(k-i)$$

por ser X e Y independientes. Entonces,

$$\begin{aligned} P(X + Y = k) &= \sum_{i=0}^k \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{e^{-\mu} \mu^{k-i}}{(k-i)!} = \frac{e^{-(\lambda+\mu)}}{k!} \sum_{i=0}^k \frac{k!}{i!(k-i)!} \lambda^i \mu^{k-i} = \\ &= \frac{e^{-(\lambda+\mu)}}{k!} (\lambda + \mu)^k. \end{aligned}$$

Entonces, $V = X + Y$ tiene distribución de Poisson de parámetro $\lambda + \mu$. O sea

$$X + Y \sim P(\lambda + \mu)$$

Este resultado se extiende por inducción al caso de n v.a. : si X_1, \dots, X_n son v.a. independientes tales que $X_i \sim P(\lambda_i)$ para $i = 1, \dots, n$, entonces $X_1 + \dots + X_n \sim P(\lambda_1 + \dots + \lambda_n)$.

2) Sean X e Y v.a. independientes con distribución exponencial de parámetro λ , o sea, sean $X \sim E(\lambda)$ e $Y \sim E(\lambda)$ independientes, y sea $V = X + Y$. La v.a. V toma valores en el intervalo $(0, \infty)$, por lo tanto, si $v \leq 0$, $F_V(v) = 0$. Sea $v > 0$,

$$F_V(v) = P(X + Y \leq v) = \iint_{\{(x,y)/x+y \leq v\}} f_{XY}(x,y) dx dy = \iint_{\{(x,y)/x+y \leq v\}} f_X(x) f_Y(y) dx dy$$

pues X e Y son independientes. Entonces,

$$\begin{aligned} P(X + Y \leq v) &= \int_0^v \left(\int_0^{v-y} f_X(x) f_Y(y) dx \right) dy = \int_0^v \left(\int_0^{v-y} \lambda e^{-\lambda x} \lambda e^{-\lambda y} dx \right) dy = \\ &= \int_0^v \lambda e^{-\lambda y} \left(\int_0^{v-y} \lambda e^{-\lambda x} dx \right) dy = \int_0^v \lambda e^{-\lambda y} \left(1 - e^{-\lambda(v-y)} \right) dy = \end{aligned}$$

$$= \int_0^v \lambda e^{-\lambda y} dy - \int_0^v \lambda e^{-\lambda v} dy = 1 - e^{-\lambda v} - \lambda e^{-\lambda v} v$$

Derivando respecto de v , se obtiene la densidad de $V = X + Y$, que es

$$f_V(v) = (\lambda e^{-\lambda v} + \lambda^2 e^{-\lambda v} v - \lambda e^{-\lambda v}) I_{(0,\infty)}(v) = \lambda^2 e^{-\lambda v} v I_{(0,\infty)}(v)$$

lo que demuestra que V tiene distribución Gamma de parámetros $(2, \lambda)$.

3) Se puede demostrar que, en general, si $X \sim \Gamma(\alpha, \lambda)$ e $Y \sim \Gamma(\beta, \lambda)$ son variables aleatorias independientes, entonces

$$X + Y \sim \Gamma(\alpha + \beta, \lambda)$$

Función generadora de momentos de la suma de v.a. independientes: Sean, en principio X e Y dos v.a. independientes, entonces la función generadora de la suma $X + Y$ es el producto de las funciones generadoras, es decir

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

En efecto, si por ejemplo X e Y son dos v.a. continuas e independientes,

$$\begin{aligned} M_{X+Y}(t) &= E(e^{t(X+Y)}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t(x+y)} f_{XY}(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{tx} e^{ty} f_X(x) f_Y(y) dx dy = \\ &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \int_{-\infty}^{\infty} e^{ty} f_Y(y) dy = E(e^{tX}) E(e^{tY}) = M_X(t) M_Y(t) \end{aligned}$$

como queríamos demostrar. Para el caso discreto, se demuestra en forma similar.

Es inmediato verificar que si X_1, X_2, \dots, X_n son v.a. independientes,

$$M_{X_1+X_2+\dots+X_n}(t) = \prod_{i=1}^n M_{X_i}(t)$$

Ejemplos: 1) Demostraremos, usando funciones generadoras de momentos que si $X \sim P(\lambda)$ e $Y \sim P(\mu)$ son v.a. independientes, $X + Y \sim P(\lambda + \mu)$. En efecto,

$$M_{X+Y}(t) = M_X(t) M_Y(t) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$$

y se obtiene la función generadora de momentos de una v.a. Poisson con parámetro $(\lambda + \mu)$. Recordemos que la función generadora de momentos determina la distribución de la v.a..

2) Demostraremos ahora, usando funciones generadoras de momentos que si X e Y son v.a. independientes con distribución exponencial de parámetro λ , o sea $X \sim E(\lambda)$ e $Y \sim E(\lambda)$, entonces $V = X + Y \sim \Gamma(2, \lambda)$. En efecto,

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{\lambda}{\lambda-t} \frac{\lambda}{\lambda-t} = \left(\frac{\lambda}{\lambda-t} \right)^2$$

y se obtiene la función generadora de momentos de una v.a. $\Gamma(2, \lambda)$.

Sumas y promedios de variables aleatorias

En la página 100, demostramos que

$$E(a_1 X_1 + a_2 X_2) = a_1 E(X_1) + a_2 E(X_2).$$

¿Qué ocurre con la varianza de una combinación lineal de dos variables aleatorias?

$$\begin{aligned} V(a_1 X_1 + a_2 X_2) &= E\left[\left((a_1 X_1 + a_2 X_2) - E(a_1 X_1 + a_2 X_2)\right)^2\right] = \\ &= E\left[\left((a_1 X_1 + a_2 X_2) - (a_1 \mu_1 + a_2 \mu_2)\right)^2\right] = E\left[\left((a_1 X_1 - a_1 \mu_1) + (a_2 X_2 - a_2 \mu_2)\right)^2\right] = \\ &= E\left(a_1^2 (X_1 - \mu_1)^2 + E\left(a_2^2 (X_2 - \mu_2)^2\right) + 2E\left[a_1 a_2 (X_1 - \mu_1)(X_2 - \mu_2)\right]\right) = \\ &= a_1^2 V(X_1) + a_2^2 V(X_2) + 2a_1 a_2 \text{cov}(X_1, X_2) \end{aligned}$$

La siguiente proposición generaliza estos resultados para todo $n \geq 2$.

Proposición: Sean X_1, X_2, \dots, X_n v.a. cualesquiera con $E(X_i) = \mu_i$ y $V(X_i) = \sigma_i^2$ y a_1, a_2, \dots, a_n números reales, entonces

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mu_i \tag{1}$$

$$V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \sigma_i^2 + 2 \sum_{i < j} a_i a_j \text{cov}(X_i, X_j)$$

Dem: En primer lugar, probemos la expresión para la esperanza mediante inducción en n .

Como dijimos, ya lo hemos demostrado para $n=2$, supongamos ahora que la expresión es cierta para $n = k$ y probémosla para $n = k + 1$.

$$E\left(\sum_{i=1}^{k+1} a_i X_i\right) = E\left(\sum_{i=1}^k a_i X_i + a_{k+1} X_{k+1}\right) = E(Y + a_{k+1} X_{k+1})$$

siendo $Y = \sum_{i=1}^k a_i X_i$. Como para $n = 2$ se cumple, se obtiene

$$E\left(\sum_{i=1}^{k+1} a_i X_i\right) = E(Y + a_{k+1} X_{k+1}) = E(Y) + a_{k+1} E(X_{k+1}) = E\left(\sum_{i=1}^k a_i X_i\right) + a_{k+1} \mu_{k+1}$$

y, utilizando la hipótesis inductiva

$$E\left(\sum_{i=1}^{k+1} a_i X_i\right) = \sum_{i=1}^k a_i \mu_i + a_{k+1} \mu_{k+1} = \sum_{i=1}^{k+1} a_i \mu_i$$

como queríamos demostrar.

Probemos ahora la expresión correspondiente a la varianza.

$$\begin{aligned} V\left(\sum_{i=1}^n a_i X_i\right) &= \text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{i=1}^n a_i X_i\right) = E\left(\sum_{i=1}^n a_i X_i \cdot \sum_{j=1}^n a_j X_j\right) - E\left(\sum_{i=1}^n a_i X_i\right)E\left(\sum_{j=1}^n a_j X_j\right) = \\ &= E\left(\sum_{i=1}^n \sum_{j=1}^n a_i a_j X_i X_j\right) - \left(\sum_{i=1}^n a_i \mu_i\right)\left(\sum_{j=1}^n a_j \mu_j\right) = \\ &= \left(\sum_{i=1}^n \sum_{j=1}^n a_i a_j E(X_i X_j)\right) - \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mu_i \mu_j = \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j (E(X_i X_j) - \mu_i \mu_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{cov}(X_i, X_j) \end{aligned}$$

Teniendo en cuenta que si $i = j$, $\text{cov}(X_i, X_i) = V(X_i)$ y que $\text{cov}(X_i, X_i) = \text{cov}(X_j, X_j)$, obtenemos el resultado que queríamos demostrar.

Corolario: Sean X_1, X_2, \dots, X_n v.a. **independientes** con $E(X_i) = \mu_i$ y $V(X_i) = \sigma_i^2$ y a_1, a_2, \dots, a_n números reales, entonces

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mu_i \quad V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \sigma_i^2$$

Dem: Resulta inmediatamente del hecho que, por ser las v.a. independientes,

$$\text{cov}(X_i, X_j) = 0 \quad \forall i \neq j.$$

Corolario: Sean X_1, X_2, \dots, X_n v.a. **independientes e idénticamente distribuidas** (i.i.d.) con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 \quad \forall i = 1, \dots, n$ y a_1, a_2, \dots, a_n números reales, entonces

$$E\left(\sum_{i=1}^n a_i X_i\right) = \mu \sum_{i=1}^n a_i \quad V\left(\sum_{i=1}^n a_i X_i\right) = \sigma^2 \sum_{i=1}^n a_i^2$$

Dem: Se verifica inmediatamente a partir del corolario anterior.

Propiedad: Sean X_1, X_2, \dots, X_n v.a. independientes e idénticamente distribuidas (i.i.d.) con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 \quad \forall i = 1, \dots, n$, entonces

$$\text{a) } E\left(\sum_{i=1}^n X_i\right) = n\mu \quad V\left(\sum_{i=1}^n X_i\right) = n\sigma^2$$

$$\text{b) } E(\bar{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \mu \quad V(\bar{X}) = V\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sigma^2}{n}$$

Dem: Ejercicio.

Desigualdad de Chebyshev:

Para calcular la probabilidad de un evento descrito en términos de una v.a. X es necesario conocer la distribución de la v.a. La desigualdad de Chebyshev provee una cota que no depende de la distribución sino sólo de la esperanza y la varianza de X .

Proposición: Sea X una v.a. con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$, entonces

$$\forall \varepsilon > 0, \quad P(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Dem: Lo haremos para el caso continuo. La demostración para el caso discreto es similar.

$$\begin{aligned} \sigma^2 &= E((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \\ &= \int_{\{x/|x-\mu|>\varepsilon\}} (x - \mu)^2 f(x) dx + \int_{\{x/|x-\mu|\leq\varepsilon\}} (x - \mu)^2 f(x) dx \geq \\ &\geq \int_{\{x/|x-\mu|>\varepsilon\}} (x - \mu)^2 f(x) dx \geq \int_{\{x/|x-\mu|>\varepsilon\}} \varepsilon^2 f(x) dx = \varepsilon^2 P(|X - \mu| > \varepsilon) \end{aligned}$$

Entonces,

$$\frac{\sigma^2}{\varepsilon^2} \geq P(|X - \mu| > \varepsilon)$$

como queríamos demostrar.

Observación: La cota que provee la desigualdad de Chebyshev puede ser grosera o, peor aún, no informativa, por ejemplo, si $\varepsilon^2 \leq \sigma^2$.

Ejemplo: Sea $X \sim U(0,10)$, entonces $E(X) = 5$ y $V(X) = 100/12$.

Aplicando la desigualdad de Chebyshev,

$$P(|X - 5| > 4) \leq \frac{\sigma^2}{16} = \frac{100/12}{16} \cong 0.52$$

pero, si calculamos en forma exacta esa probabilidad,

$$\begin{aligned} P(|X - 5| > 4) &= 1 - P(|X - 5| \leq 4) = 1 - P(-4 \leq X - 5 \leq 4) = 1 - P(1 \leq X \leq 9) = \\ &= 1 - F_X(9) + F_X(1) = 1 - \frac{9}{10} + \frac{1}{10} = 0.20 \end{aligned}$$

Formas equivalentes de la desigualdad de Chebyshev:

a) $\forall \varepsilon > 0, \quad P(|X - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$

b) $\forall k > 1, \quad P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$

c) $\forall k > 1, \quad P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

(En realidad, b) y c) son ciertas para todo $k > 0$, pero si $k \leq 1$ la desigualdad es trivial)

Las dos últimas formas muestran como el desvío standard mide el grado de “concentración” de la distribución alrededor de $\mu = E(X)$.

Ley de los Grandes Números:

Sea X una v.a. con función de densidad $f(x)$ o función de probabilidad puntual $p(x)$ y con $E(X) = \mu$. Supongamos que se desea “estimar” μ . Como hemos visto que la esperanza de una v.a. se puede pensar como un promedio de sus valores, parece razonable estimarla mediante el promedio de valores observados de X . Por supuesto que en una situación real sólo tendremos un número finito de observaciones y nos preguntamos: usando sólo un número finito de valores de X , ¿puede hacerse inferencia confiable respecto de $E(X)$? La respuesta es SI y se demuestra a través de la Ley de los Grandes Números que nos dice que el promedio \bar{X} converge a μ cuando el número de observaciones (o tamaño de la muestra) tiende a infinito. Observemos lo que sucede en la siguiente figura.

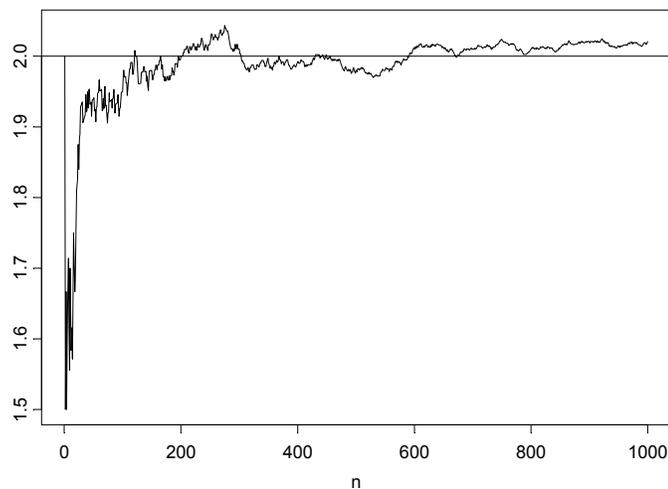


Figura 1: Comportamiento asintótico del promedio muestral. El promedio del número observado de caras, \bar{X} , cuando 4 monedas equilibradas son arrojadas se aproxima al valor medio $\mu=2$ de la distribución.

¿En qué sentido converge \bar{X} a μ ?

Sea (X_n) ($n \geq 1$) una sucesión de variables aleatorias, diremos que X_n converge en probabilidad a la v.a. X y lo notaremos $X_n \xrightarrow{p} X$, si

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$

Ley de los Grandes Números: Sean X_1, X_2, \dots v.a. independientes e idénticamente distribuidas (muestra aleatoria) con $E(X) = \mu$ y $V(X) = \sigma^2 < \infty$, entonces

$$\bar{X}_n \xrightarrow{p} \mu$$

siendo $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ el denominador promedio muestral.

Dem: Sabemos que $E(\bar{X}_n) = \mu$ y $V(\bar{X}_n) = \frac{\sigma^2}{n}$, entonces aplicando la desigualdad de Chebyshev,

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{V(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n \varepsilon^2} \quad \forall \varepsilon > 0$$

y, por lo tanto

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{n \varepsilon^2} = 0 \quad \forall \varepsilon > 0$$

Luego, $\bar{X}_n \xrightarrow{p} \mu$, como queríamos demostrar.

Versión Bernoulli de la Ley de los Grandes Números: Consideremos n repeticiones independientes de un experimento aleatorio y sea A un suceso con probabilidad $P(A) = p$, constante en las n repeticiones. Si llamamos n_A a la frecuencia absoluta de A (número de veces que ocurre A en las n repeticiones) y $f_A = n_A / n$ a la frecuencia relativa, entonces

$$f_A \xrightarrow{p} p$$

es decir,

$$P(|f_A - p| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0$$

Dem: Como $n_A \sim \text{Bi}(n,p)$ con $p = P(A)$, entonces $E(n_A) = n p$ y $V(n_A) = n p (1-p)$. Luego

$$E\left(\frac{n_A}{n}\right) = p \quad V\left(\frac{n_A}{n}\right) = \frac{p(1-p)}{n}$$

y, aplicando la desigualdad de Chebyshev,

$$P(|f_A - p| > \varepsilon) \leq \frac{V(f_A)}{\varepsilon^2} = \frac{p(1-p)}{n \varepsilon^2} \quad \forall \varepsilon > 0$$

Luego,

$$\lim_{n \rightarrow \infty} P(|f_A - p| > \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{p(1-p)}{n \varepsilon^2} = 0 \quad \forall \varepsilon > 0$$

como queríamos demostrar.

Ejemplo: ¿Cuántas repeticiones del experimento deberían hacerse para que la frecuencia relativa difiera de p en menos de 0.01 con probabilidad mayor o igual que 0.95?

En este caso, $\varepsilon = 0.01$ y queremos encontrar n tal que

$$P(|f_A - p| < 0.01) \geq 0.95$$

Pero, dado que

$$P(|f_A - p| < 0.01) \geq 1 - \frac{p(1-p)}{n(0.01)^2}$$

$$1 - \frac{p(1-p)}{n(0.01)^2} \geq 0.95 \Leftrightarrow \frac{p(1-p)}{n(0.01)^2} \leq 0.05 \Leftrightarrow n \geq \frac{p(1-p)}{(0.01)^2(0.05)}$$

El valor mínimo de n depende de p y es máximo cuando $p = 0.50$. Por ejemplo,

$$\begin{aligned} p = 0.50 &\Rightarrow n \geq 50000 \\ p = 0.10 &\Rightarrow n \geq 18500 \\ p = 0.01 &\Rightarrow n \geq 1980 \end{aligned}$$

Distribución de la suma de variables aleatorias independientes: En general es difícil calcular la distribución de la suma o de una combinación lineal de n v.a. independientes, aún cuando tengan la misma distribución. Sin embargo, en algunos casos la distribución de la suma o de combinaciones lineales es conocida. Recapitulemos algunos resultados.

a) Si X_1, X_2, \dots, X_n son v.a. independientes tales que $X_i \sim Bi(n_i, p)$, entonces

$$\sum_{i=1}^n X_i \sim Bi\left(\sum_{i=1}^n n_i, p\right).$$

En particular, si $X_i \sim Bi(1, p) \quad \forall i$, entonces $\sum_{i=1}^n X_i \sim Bi(n, p)$.

b) Si X_1, X_2, \dots, X_n son v.a. independientes tales que $X_i \sim P(\lambda_i)$, entonces

$$\sum_{i=1}^n X_i \sim P\left(\sum_{i=1}^n \lambda_i\right).$$

c) Si X_1, X_2, \dots, X_n son v.a. i.i.d. tales que $X_i \sim G(p)$, entonces $\sum_{i=1}^n X_i \sim BN(n, p)$.

d) Si X_1, X_2, \dots, X_n son v.a. i.i.d. tales que $X_i \sim \mathcal{E}(\lambda)$, entonces $\sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$.

e) Si X_1, X_2, \dots, X_n son v.a. independientes tales que $X_i \sim \Gamma(n_i, \lambda)$, entonces

$$\sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n n_i, \lambda\right).$$

f) Si X_1, X_2, \dots, X_n son v.a. independientes tales que $X_i \sim N(\mu_i, \sigma_i^2)$ y a_1, a_2, \dots, a_n son

números reales, entonces $\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$.

En particular, si X_1, X_2, \dots, X_n son v.a. i.i.d. tales que $X_i \sim N(\mu, \sigma^2)$, entonces

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{y} \quad T = \sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2).$$

Dem: Todos estos resultados pueden demostrarse fácilmente usando funciones generadoras de momentos. Como ejemplo, demostremos la propiedad e), es decir que si

X_1, X_2, \dots, X_n son v.a. independientes tales que $X_i \sim \Gamma(n_i, \lambda)$, entonces $\sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n n_i, \lambda\right)$.

Por ser las X_i v.a. independientes, la función generadora de la suma es el producto de las funciones generadoras, entonces

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left(\frac{\lambda}{\lambda - t}\right)^{n_i} = \left(\frac{\lambda}{\lambda - t}\right)^{\sum_{i=1}^n n_i} \Rightarrow \sum_{i=1}^n X_i \sim \Gamma\left[\sum_{i=1}^n n_i, \lambda\right]$$

como queríamos demostrar.

Veremos ahora que, cuando las v.a. no son normales, la distribución normal resulta una buena aproximación para la distribución de \bar{X} y T .

Teorema Central del Límite: Sean X_1, X_2, \dots v.a. i.i.d con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 < \infty$, entonces si n es suficientemente grande,

$$\frac{T - n\mu}{\sqrt{n} \sigma} \stackrel{(a)}{\sim} N(0,1) \qquad \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \stackrel{(a)}{\sim} N(0,1)$$

o, dicho de otro modo,

$$\frac{T - n\mu}{\sqrt{n} \sigma} \xrightarrow{d} Z \sim N(0,1) \qquad \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$$

donde la convergencia en distribución (\xrightarrow{d}) se interpreta en el siguiente sentido:

$$P\left(\frac{T - n\mu}{\sqrt{n} \sigma} \leq a\right) \cong \Phi(a) \qquad P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq a\right) \cong \Phi(a)$$

es decir, que las funciones de distribución convergen a la función de distribución normal standard.

Dem: Lo demostraremos bajo la hipótesis de que la función generadora de momentos de X_i , $M_{X_i}(t)$ existe y es finita.

Supongamos inicialmente que $\mu = 0$ y $\sigma^2 = 1$. En este caso, la función generadora de momentos de $\frac{T}{\sqrt{n}}$ está dada por

$$M_{T/\sqrt{n}}(t) = M_T(t/\sqrt{n}) = M_{\sum_{i=1}^n X_i}(t/\sqrt{n}) = \prod_{i=1}^n M_{X_i}(t/\sqrt{n}) = \left(M_{X_i}(t/\sqrt{n})\right)^n$$

por ser las X_i independientes. Sea $L(u) = \ln(M_{X_i}(u))$, entonces

$$L(0) = \ln(M_{X_i}(0)) = \ln(1) = 0$$

$$L'(0) = \left. \frac{\partial \ln M_{X_i}(u)}{\partial u} \right|_{u=0} = \left. \frac{M'_{X_i}(u)}{M_{X_i}(u)} \right|_{u=0} = \frac{M'_{X_i}(0)}{M_{X_i}(0)} = \frac{\mu}{1} = \mu = 0$$

$$L''(0) = \left. \frac{\partial^2 \ln M_{X_i}(u)}{\partial u^2} \right|_{u=0} = \left. \frac{M''_{X_i}(u) \cdot M_{X_i}(u) - [M'_{X_i}(u)]^2}{[M_{X_i}(u)]^2} \right|_{u=0} = \frac{M''_{X_i}(0)M_{X_i}(0) - [M'_{X_i}(0)]^2}{[M_{X_i}(0)]^2} = E(X_i^2) = 1$$

Ahora, para probar el teorema, demostraremos que $M_{T/\sqrt{n}}(t) \rightarrow e^{t^2/2}$ o equivalentemente, que $nL(t/\sqrt{n}) \rightarrow t^2/2$. Aplicando la regla de L'Hospital dos veces,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{L(t/\sqrt{n})}{1/n} &= \lim_{n \rightarrow \infty} \frac{-L'(t/\sqrt{n})t n^{-3/2}}{-2n^{-2}} = \lim_{n \rightarrow \infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} = \\ &= \lim_{n \rightarrow \infty} \frac{-L''(t/\sqrt{n})t^2 n^{-3/2}}{-2n^{-3/2}} = \lim_{n \rightarrow \infty} \frac{L''(t/\sqrt{n})t^2}{2} = \frac{t^2}{2}. \end{aligned}$$

por lo tanto hemos probado el Teorema Central del Límite para $\mu = 0$ y $\sigma^2 = 1$. El caso general resulta considerando las v.a. standarizadas $\frac{X_i - \mu}{\sigma} = X_i^*$.

Observación: ¿Qué significa n suficientemente grande? ¿Cómo sabemos si la aproximación es buena? El tamaño de muestra requerido para que la aproximación sea razonable depende de la forma de la distribución de las X_i . Mientras más simétrica y acampanada sea, más rápidamente se obtiene una buena aproximación.

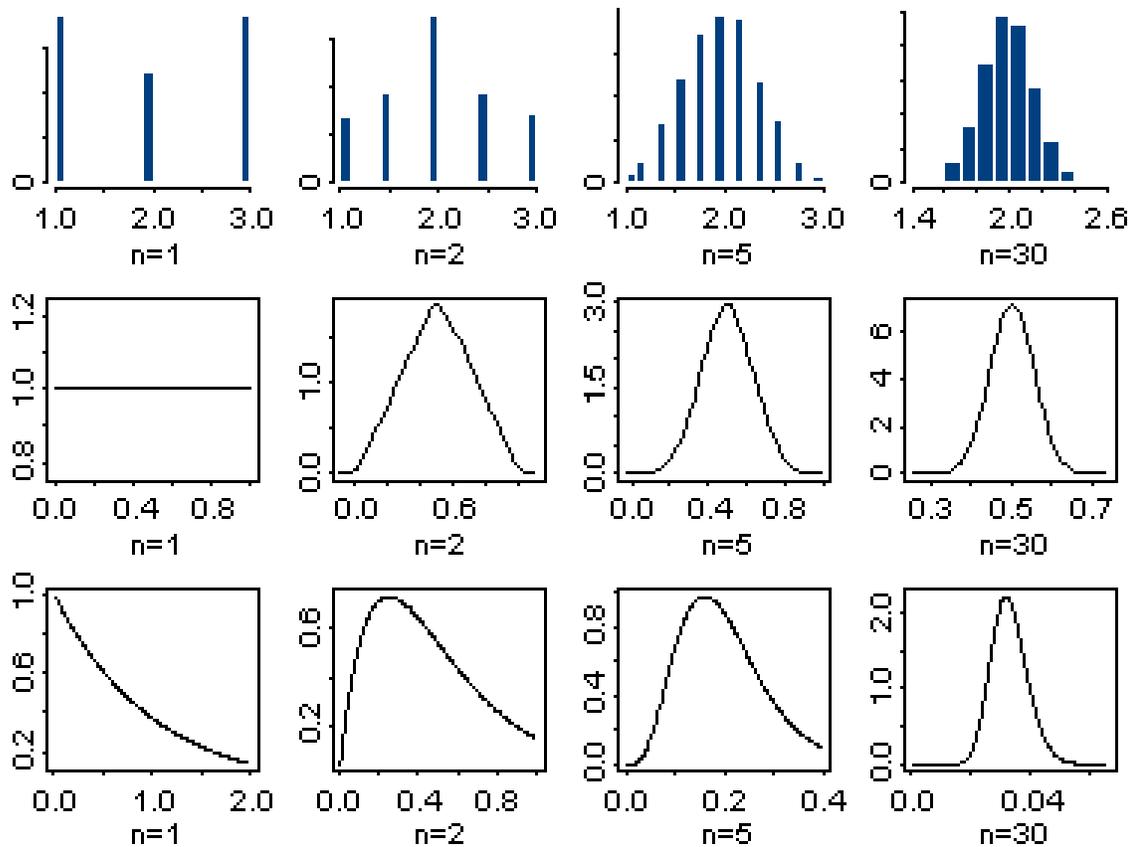


Figura 2: Distribución de \bar{x} para distintas distribuciones cuando $n=2, 5$ y 30 .
 a) Distribución discreta, b) Distribución Uniforme, c) Distribución Exponencial

Ejemplo: Al sumar números, una calculadora aproxima cada número al entero más próximo. Los errores de aproximación se suponen independientes y con distribución $U(-0.5, 0.5)$.

- a) Si se suman 1500 números, ¿cuál es la probabilidad de que el valor absoluto del error total exceda 15?

Si llamamos X_i al error correspondiente al i -ésimo sumando, el error total es $T_{1500} = \sum_{i=1}^{1500} X_i$.

Entonces,

$$P(|T_{1500}| > 15) = 1 - P(|T_{1500}| \leq 15) = 1 - P(-15 \leq T_{1500} \leq 15) =$$

$$= 1 - P\left(\frac{-15}{\sqrt{1500/12}} \leq \frac{T_{1500}}{\sqrt{1500/12}} \leq \frac{15}{\sqrt{1500/12}}\right) \cong 1 - \Phi\left(\frac{15}{\sqrt{1500/12}}\right) + \Phi\left(\frac{-15}{\sqrt{1500/12}}\right) =$$

$$= 1 - \Phi(1.34) + \Phi(-1.34) = 0.18$$

Hemos usado que $E(X_i) = 0$ y $V(X_i) = \frac{1}{12}$ y por lo tanto $E(T_{1500}) = 0$ y $V(T_{1500}) = \frac{1500}{12}$.

b) ¿Cuántos números pueden sumarse a fin de que el valor absoluto del error total sea menor o igual que 10 con probabilidad mayor o igual que 0.90?

Buscamos el valor de n tal que

$$P(|T_n| \leq 10) \geq 0.90$$

$$P(|T_n| \leq 10) \geq 0.90 \Leftrightarrow P(-10 \leq T_n \leq 10) \geq 0.90 \Leftrightarrow P\left(\frac{-10}{\sqrt{n/12}} \leq T_n \leq \frac{10}{\sqrt{n/12}}\right) \geq 0.90$$

Aplicando la aproximación normal, debemos hallar n tal que

$$\Phi\left(\frac{10}{\sqrt{n/12}}\right) - \Phi\left(\frac{-10}{\sqrt{n/12}}\right) \geq 0.90 \Leftrightarrow 2\Phi\left(\frac{10}{\sqrt{n/12}}\right) - 1 \geq 0.90 \Leftrightarrow \Phi\left(\frac{10}{\sqrt{n/12}}\right) \geq 0.95$$

$$\Leftrightarrow \frac{10}{\sqrt{n/12}} \geq 1.64 \Leftrightarrow \sqrt{n} \leq 21.12 \Leftrightarrow n \leq 446$$

es decir, que se pueden sumar a lo sumo 446 números para que el valor absoluto del error total sea menor o igual que 10 con probabilidad mayor o igual que 0.90.

Aproximación de la distribución binomial por la normal: Sea $X \sim \text{Bi}(n, p)$, entonces X es el número de éxitos en n repeticiones de un experimento binomial con probabilidad de éxito igual a p , y X/n es la proporción muestral de éxitos.

Definamos las siguientes variables aleatorias

$$X_i = \begin{cases} 1 & \text{si se obtuvo Éxito en la repetición } i \\ 0 & \text{si se obtuvo Fracaso en la repetición } i \end{cases}$$

para $i = 1, \dots, n$. Estas v.a. son independientes, $X_i \sim \text{Bi}(1, p) \forall i$ y $X = \sum_{i=1}^n X_i$.

Aplicando el Teorema Central del Límite, si n es suficientemente grande,

$$X \stackrel{(a)}{\sim} N(np, np(1-p)) \qquad \frac{X}{n} \stackrel{(a)}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Se considera que la aproximación es buena si $np \geq 5$ y $n(1-p) \geq 5$.

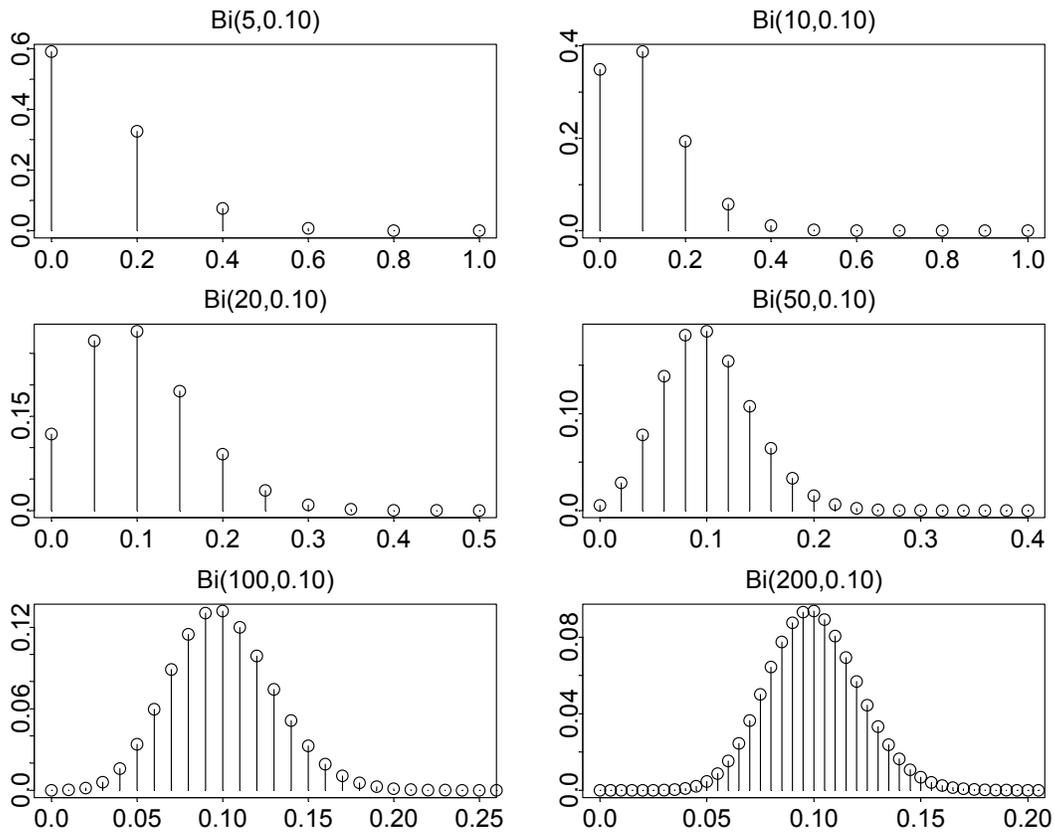


Figura 3: Distribución de $\frac{X}{n}$

Corrección por continuidad: Cuando se aproxima una distribución discreta por una continua, como es el caso de la aproximación de la distribución binomial por la normal, es necesario efectuar una corrección. Consideremos el siguiente ejemplo:

Sea $X \sim Bi(100, 0.6)$ y calculemos en forma aproximada $P(X \leq 50)$ y $P(X \geq 51)$.

Si aplicamos directamente el TCL, obtenemos:

$$P(X \leq 50) = P\left(\frac{X - 60}{\sqrt{24}} \leq \frac{50 - 60}{\sqrt{24}}\right) \cong \Phi(-2.04) = 0.021$$

$$P(X \geq 51) = P\left(\frac{X - 60}{\sqrt{24}} \geq \frac{51 - 60}{\sqrt{24}}\right) \cong 1 - \Phi(-1.84) = 0.967$$

Si bien, $P(X \leq 50) + P(X \geq 51) = 1$, los valores aproximados no satisfacen esta restricción. Para evitar este problema, se efectúa la siguiente corrección, denominada *corrección por continuidad*,

$$P(X \leq 50) = P(X \leq 50.5) = P\left(\frac{X - 60}{\sqrt{24}} \leq \frac{50.5 - 60}{\sqrt{24}}\right) \cong \Phi(-1.94) = 0.026$$

$$P(X \geq 51) = P(X \geq 50.5) = P\left(\frac{X - 60}{\sqrt{24}} \geq \frac{50.5 - 60}{\sqrt{24}}\right) \cong 1 - \Phi(-1.94) = 0.974$$

En general, cuando la v.a. es discreta y $x_i - x_{i-1} = 1$, la corrección se realiza en la forma:

$$P(X \leq a) = P(X \leq a + 0.5)$$

$$P(X \geq a) = P(X \geq a - 0.5)$$

Si la distancia entre dos valores sucesivos de X es $k > 1$, ¿cómo aplicaría la corrección por continuidad?

Ejemplo: Sea $X \sim Bi(60, 1/3)$. Calcular en forma aproximada la probabilidad de que X sea mayor o igual que 25.

$$P(X \geq 25) = P(X \geq 24.5) = P\left(\frac{X - 60 \cdot \frac{1}{3}}{\sqrt{60 \cdot \frac{1}{3} \cdot \frac{2}{3}}} \geq \frac{24.5 - 60 \cdot \frac{1}{3}}{\sqrt{60 \cdot \frac{1}{3} \cdot \frac{2}{3}}}\right) \cong 1 - \Phi(1.23) = 0.11$$

Otras aplicaciones del Teorema Central del Límite:

a) Sean X_1, X_2, \dots, X_n v.a. i.i.d. con distribución Poisson de parámetro λ , entonces

$$\sum_{i=1}^n X_i \sim P(n\lambda)$$

Por lo tanto, cualquier v.a. con distribución de Poisson con parámetro suficientemente grande puede ser aproximada por la distribución normal.

b) Sean X_1, X_2, \dots, X_n v.a. independientes con distribución Gamma de parámetros n_i y λ , o sea $X_i \sim \Gamma(n_i, \lambda)$ entonces

$$\sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n n_i, \lambda\right)$$

Por lo tanto, cualquier v.a. con distribución $\Gamma(m, \lambda)$ con parámetro m suficientemente grande puede ser aproximada por la distribución normal.

Una aplicación de suma de v.a. independientes y generación de números al azar:

Recordemos que un **proceso de Poisson** permite modelar una situación en la que los eventos ocurren a lo largo del tiempo (o espacio, volumen, etc.).

Hemos visto, que bajo ciertos supuestos, si definimos la variable

$$X_t = \text{cantidad de eventos que ocurren en el intervalo } [0,t]$$

entonces $X_t \sim P(\lambda t)$, donde λ es la tasa media de ocurrencias o intensidad del proceso.

También hemos mencionado que, si denotamos

- T_1 = tiempo que transcurre entre que empezamos a medir y el momento en que ocurre el primer evento
- T_2 = tiempo que transcurre entre el primer evento y el segundo evento.

y, en general,

- T_i = tiempo que transcurre entre el $(i-1)$ -ésimo evento y el i -ésimo evento ($i \in N$)

las T_i son variables aleatorias independientes y con distribución exponencial, todas con el mismo parámetro λ .

Es claro que, si a uno le interesara el tiempo que transcurre desde el inicio hasta la k -ésima ocurrencia, esta variable aleatoria podría expresarse como

$$\sum_{i=1}^k T_i$$

Veamos la recíproca, es decir, veamos como podemos construir un proceso de Poisson a partir de v.a. i.i.d. con distribución exponencial.

Proposición: Sean $W_1, W_2, \dots, W_k, \dots$ v.a. independientes con distribución $E(1)$. Consideremos el siguiente proceso. Comenzamos a medir el tiempo en $t = 0$ y consideramos que ocurre el primer evento en el instante W_1 , el segundo en el instante $W_1 + W_2$, y en general el k -ésimo evento en el instante $W_1 + W_2 + \dots + W_k$. Si para $t > 0$, definimos la variable aleatoria

$$X_t = \text{cantidad de eventos que ocurren en el intervalo } [0,t]$$

entonces X_t es una variable discreta y su distribución es $P(t)$.

Dem: Sea $k \in N \cup \{0\}$ y consideremos el evento $[X_t \geq k]$. Observemos que

$$[X_t \geq k] \iff \text{hubo } k \text{ ó más eventos en el intervalo } [0,t] \iff \text{hubo por lo menos } k \text{ eventos en el intervalo } [0,t] \iff \sum_{i=1}^k W_i \leq t$$

Calculemos la probabilidad de dicho evento:

$$P(X_t \geq k) = P\left(\sum_{i=1}^k W_i \leq t\right)$$

Como las $W_1, W_2, \dots, W_k, \dots$ son variables aleatorias independientes y con distribución $E(1) = \Gamma(1,1)$, entonces

$$\sum_{i=1}^k W_i \sim \Gamma(k,1)$$

y por lo tanto

$$P\left(\sum_{i=1}^k W_i \leq t\right) = \int_{-\infty}^t f_S(s) ds$$

con $S = \sum_{i=1}^k W_i \sim \Gamma(k,1)$ y en consecuencia $f_S(s) = \frac{1}{(k-1)!} s^{k-1} e^{-s} I_{(0,+\infty)}(s)$. Entonces,

$$P\left(\sum_{i=1}^k W_i \leq t\right) = \int_0^t \frac{1}{(k-1)!} s^{k-1} e^{-s} ds$$

Llamemos

$$A_k(t) = \int_0^t \frac{1}{(k-1)!} s^{k-1} e^{-s} ds$$

a la función de distribución acumulada de una $\Gamma(k,1)$. Integrando por partes una vez, si consideramos

$$u = \frac{s^{k-1}}{(k-1)!} \quad u' = \frac{(k-1)s^{k-2}}{(k-1)!} = \frac{s^{k-2}}{(k-2)!}$$

y

$$v' = e^{-s} \quad v = -e^{-s}$$

obtenemos

$$\begin{aligned}
 A_k(t) &= \int_0^t \frac{1}{(k-1)!} s^{k-1} e^{-s} ds = \frac{-1}{(k-1)!} s^{k-1} e^{-s} \Big|_0^t + \int_0^t \frac{1}{(k-2)!} s^{k-2} e^{-s} ds \\
 &= \frac{-1}{(k-1)!} t^{k-1} e^{-t} + A_{k-1}(t) \\
 &= \frac{-1}{(k-1)!} t^{k-1} e^{-t} + \frac{-1}{(k-2)!} t^{k-2} e^{-t} + A_{k-2}(t)
 \end{aligned}$$

Finalmente, por inducción, después de M pasos obtenemos

$$A_k(t) = \dots = -e^{-t} \sum_{i=k-M}^{k-1} \frac{t^i}{i!} + A_{k-M}(t)$$

Como

$$A_1(t) = \int_0^t e^{-s} ds = -e^{-t} + 1$$

resulta

$$A_k(t) = -e^{-t} \sum_{i=1}^{k-1} \frac{t^i}{i!} + A_1(t) = -e^{-t} \sum_{i=0}^{k-1} \frac{t^i}{i!} + 1$$

y por lo tanto

$$P(X_t \geq k) = P\left(\sum_{i=1}^k W_i \leq t\right) = A_k(t) = 1 - \sum_{i=0}^{k-1} \frac{t^i}{i!} e^{-t}$$

Si tomamos el complemento resulta

$$P(X_t < k) = P(X_t \leq k-1) = \sum_{i=0}^{k-1} \frac{t^i}{i!} e^{-t}$$

que corresponde a la función de distribución acumulada de una variable con distribución $P(t)$, tal como queríamos demostrar.

Este resultado es muy útil para generar variables aleatorias con distribución de Poisson a partir de exponenciales, a las que podemos generar fácilmente a partir de $U(0,1)$.

Supongamos que deseamos generar una variable aleatoria X con distribución $P(\lambda)$. Para ello basta utilizar la proposición anterior tomando $t = \lambda$. Podemos describir el algoritmo de la siguiente forma:

Paso 1: generamos una v.a. W_1 con distribución $E(1)$.

Paso 2: chequeamos si $W_1 \leq t$. Si ésto ocurre, continuamos con el paso siguiente. Si, en cambio, $W_1 \geq t$ terminamos y $X = 0$.

Paso 3: generamos una v.a. W_2 con distribución $E(1)$, independiente de W_1 .

Paso 4: chequeamos si $W_1 + W_2 \leq t$. Si ésto ocurre, continuamos con el paso siguiente. Si no, terminamos y $X = 1$.

Paso $2k-1$: generamos una v.a. W_k con distribución $E(1)$, independiente de W_1, W_2, \dots, W_{k-1} .

Paso $2k$: chequeamos si $W_1 + W_2 + \dots + W_k \leq t$. Si ésto ocurre seguimos, si no terminamos y $X = k$.

Etapas de una investigación

La Estadística nos permite realizar inferencias y sacar conclusiones a partir de los datos. Extrayendo la información que contienen, podremos comprender mejor las situaciones que ellos representan.

Los métodos estadísticos abarcan todas las etapas de la investigación, desde el diseño de la investigación hasta el análisis final de los datos.

Podemos distinguir tres grandes etapas:

1. **Diseño:** Planeamiento y desarrollo de las investigaciones
2. **Descripción:** Resumen y exploración de los datos
3. **Inferencia:** Predicciones y toma de decisiones sobre las características de una población en base a la información recogida en una muestra de la población.

En la etapa de **Diseño** se define cómo se desarrollará la investigación con el fin de responder las preguntas que le dieron origen. Un diseño bien realizado puede ahorrar esfuerzos en etapas posteriores y puede redundar en un análisis más sencillo. Esta etapa es crucial, pues un estudio pobremente diseñado o con datos incorrectamente recolectados o registrados puede ser incapaz de responder las preguntas que originaron el estudio.

Una vez formulado el problema, en la etapa de Diseño se definirá, entre otras cosas, la población objetivo, los tamaños de muestra, los mecanismos de selección de individuos, los criterios de inclusión y exclusión de sujetos, los métodos de asignación de tratamientos, las variables que se medirán y cómo se entrenará al equipo de trabajo para el cumplimiento del protocolo.

Los métodos de **Análisis Exploratorio** o **Estadística Descriptiva** ayudan a comprender la estructura de los datos, de manera de detectar tanto un patrón de comportamiento general como apartamientos del mismo. Una forma de realizar esto es mediante gráficos de sencilla elaboración e interpretación. Otra forma de describir los datos es resumiéndolos en uno, dos o más números que caractericen al conjunto de datos con fidelidad. Explorar los datos permitirá detectar datos erróneos o inesperados y nos ayudará a decidir qué métodos estadísticos pueden ser empleados en etapas posteriores del análisis de manera de obtener conclusiones válidas.

Finalmente, la **Inferencia Estadística** nos permite tanto hacer predicciones y estimaciones como decidir entre dos hipótesis opuestas relativas a la población de la cual provienen los datos (test de hipótesis).

La calidad de las estimaciones puede ser muy variada y está afectadas por errores. La ventaja de los métodos estadísticos es que, aplicados sobre datos obtenidos a partir de muestras aleatorias, permiten cuantificar el error que podemos cometer en una estimación o calcular la probabilidad de cometer un error al tomar una decisión en un test de hipótesis.

Para entender qué tipo de problemas consideraremos en Estadística tomemos, por ejemplo, las siguientes mediciones de la proporción de la masa de la Tierra con respecto a la Luna

Mariner II	81.3001
Mariner IV	81.3015
Mariner V	81.3006
Mariner VI	81.3011
Mariner VII	81.2997
Pioneer VI	81.3005
Pioneer VII	81.3021

En *Probabilidad* podríamos suponer que las posibles mediciones se distribuyen alrededor del verdadero valor 81.3035 siguiendo una distribución determinada y nos preguntaríamos

¿Cuál es la probabilidad de que se obtengan 7 mediciones menores que el verdadero valor?

En *Estadística*, a partir de los 7 observaciones nos preguntaríamos:

¿Son consistentes los datos con la hipótesis de que el verdadero valor es 81.3035?

¿Cuán confiable es decir que el verdadero valor está en el intervalo (81.2998, 81.3038)?

Las técnicas del análisis exploratorio nos ayudan a organizar la información que proveen los datos, de manera de detectar algún patrón de comportamiento así como también apartamientos importantes al modelo subyacente. Nos guían a la estructura subyacente en los datos de manera rápida y simple.

Estadística Descriptiva

Examinaremos los datos en forma descriptiva con el fin de:

- Organizar la información
- Sintetizar la información
- Ver sus características más relevantes
- Presentar la información

Definimos:

Población: conjunto total de los sujetos o unidades de análisis de interés en el estudio

Muestra: cualquier subconjunto de sujetos o unidades de análisis de la población en estudio.

Unidad de análisis o de observación: objeto bajo estudio. Puede ser una persona, una familia, un país, una institución o en general, cualquier objeto.

Variable: cualquier característica de la unidad de observación que interese registrar y que en el momento de ser registrada puede ser transformada en un número.

Valor de una variable, Dato, Observación o Medición: número que describe a la característica de interés en una unidad de observación particular.

Caso o Registro: conjunto de mediciones realizadas sobre una unidad de observación.

Datos cuantitativos

Esquema de Tallo y Hoja

Nos da una primera aproximación rápida a la distribución de los datos sin perder de vista las observaciones.

Ejemplo: La siguiente tabla contiene 45 observaciones correspondientes a la fuerza de compresión de cierta aleación de Aluminio-Litio.

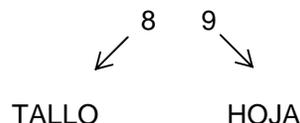
96	93	88	117	127	95	113	96
108	94	148	156	139	142	94	107
125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139
134	119	97	89	118	136	125	143
120	103	113	124	138			

- Ordenamos los datos de menor a mayor

88	89	93	94	94	95	96	96
97	103	103	104	106	107	108	111
112	112	113	113	117	117	118	119
120	120	124	125	125	125	127	127
132	134	135	136	138	139	139	142
143	148	155	155	156			

- Separamos a cada observación en dos partes: **tallo** y **hoja**
- Listamos en forma vertical y creciente los tallos y agregamos las hojas a la derecha del tallo correspondiente.

Ejemplo. Consideremos el segundo dato :



Elegimos un número de dígitos a la derecha de cada número que corresponderán a las hojas: 1 en este caso.

Separamos esos dígitos de los restantes, que constituirán los tallos. En este caso obtendremos 8 tallos, de 8 a 15.

9	3445667
10	334678
11	122337789
12	00455577
13	2456899
14	238
15	556

¿Qué podemos ver en este tipo de diagrama?

- Rango de las observaciones, valores máximo y mínimo.
- Forma de la distribución: simetría, asimetría a derecha, asimetría a izquierda y cuántos picos tiene la distribución.
- Posición del centro de la distribución y concentración de los datos.
- Desviaciones marcadas respecto al comportamiento general: outliers o valores atípicos.

Ejemplo: Los siguientes datos corresponden a tiempos de falla de cables Kevlar 49/epoxy sometidos a una presión del 90%:

TIEMPOS DE FALLA															
0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.09	0.10
0.10	0.11	0.11	0.12	0.13	0.18	0.19	0.20	0.23	0.80	0.80	0.83	0.85	0.90	0.92	0.95
0.99	1.00	1.01	1.02	1.03	1.05	1.10	1.10	1.11	1.15	1.18	1.20	1.29	1.31	1.33	1.34
1.40	1.43	1.45	1.50	1.51	1.52	1.53	1.54	1.54	1.55	1.58	1.60	1.63	1.64	1.80	1.80
1.81	2.02	2.05	2.14	2.17	2.33	3.03	3.03	3.24	4.20	4.69	7.89				

El correspondiente esquema de tallo y hoja resulta:

0	0000000000000001111111122
0	88889999
1	000001111122333444
1	55555555666888
2	00113
2	
3	002
3	
4	2
4	6
5	
5	
6	
6	
7	
7	8

En este caso cada tallo ha sido dividido en 2 líneas: en la primera se listan las hojas 0 a 4 y en la segunda las hojas 5 a 9.

Se observa asimetría a derecha y un valor alejado del resto: 7.8

El número de tallos debe ser tal que permita mostrar una imagen general de la estructura del conjunto de datos. Aunque existen algunos criterios para definir el número de tallos, la decisión depende fundamentalmente del sentido común. Demasiados detalles en general serán poco informativos, demasiado agrupamiento puede distorsionar la imagen del conjunto.

Cuando el volumen de datos es muy grande conviene usar otro tipo de gráficos que también son de fácil interpretación .

Ejemplo: Consideremos el siguiente ejemplo con datos sobre consumo diario *per cápita* de proteínas en 32 países desarrollados. Los datos se presentan ordenados de menor a mayor por simplicidad.

Consumo de proteínas per cápita en países desarrollados.

7.83	9.03	10.56
8.06	9.16	10.52
8.45	9.23	10.75
8.49	9.34	10.86
8.53	9.39	10.89
8.60	9.42	11.07
8.64	9.56	11.27
8.70	9.89	11.36
8.75	10.00	11.58
8.92	10.28	11.76
8.93	10.41	

Seleccionando como tallo la unidad obtenemos el gráfico de tallo-hojas de la izquierda de la figura. En este gráfico se acumula un número importante de hojas en cada tallo, por lo que podríamos estar perdiendo información acerca de la estructura de los datos. En el gráfico de la derecha, cada tallo ha sido dividido en dos líneas, en la primera se listan las hojas 0 a 4 y en la segunda as hojas 5 a 9.

Como puede observarse, al expandir la escala se observan más detalles y parece haber dos “grupos” de países, uno con mayor consumo per cápita de proteínas y otro con menor consumo, ya que la distribución de la variable tiene dos picos.

Variación del número de tallos. Datos de consumo de proteínas per cápita.

7		8		7		8
8		0 4 4 5 6 6 7 7 9 9		8		0 4 4
9		0 1 2 3 3 4 5 8		8		5 6 6 7 7 9 9
10		0 2 4 5 5 7 8 8		9		0 1 2 3 3 4
11		0 2 3 5 7		9		5 8
				10		0 2 4
				10		5 5 7 8 8
				11		0 2 3
				11		5 7

El problema de expandir la escala es que podrían comenzar a aparecer detalles superfluos, o simplemente atribuibles al azar.

Gráfico de tallo-hojas espalda con espalda. Comparación de grupos.

Los gráficos de tallo-hojas son útiles para comparar la distribución de una variable en dos condiciones o grupos. El gráfico se denomina tallo-hojas espalda con espalda porque ambos grupos comparten los tallos.

A continuación se muestra un gráfico de la presión arterial sistólica (PAS) a los 30 minutos de comenzada la anestesia en pacientes sometidos a dos técnicas anestésicas diferentes a las que nos referiremos como T1 y T2.

Comparación de la presión arterial sistólica en pacientes sometidos a dos técnicas anestésicas (30 minutos del inicio de la anestesia).

T1	T2
5	47
6	2
74	7 37
963	8 778999
660	9 0358
9662	10 222
821	11 37
70	12
2	13
	14
	15
4	16

El gráfico nos muestra las siguientes características de la PAS en los dos grupos de pacientes.

- La distribución de PAS tiene *forma* similar en ambos grupos: Un pico o moda y forma simétrica y aproximadamente acampanada.
- Diferencias en *posición*. Los pacientes del grupo T1 tienen niveles de PAS levemente mayores que los pacientes del grupo T2.
- Similar *dispersión*. Los valores de PAS de los pacientes de ambos grupos se encuentran en rangos aproximadamente iguales, salvo por el valor atípico (*outlier*) que se observa en el grupo T1.

Histograma

- Se divide el rango de los datos en **intervalos o clases**, que no se superpongan. Las clases deben ser **excluyentes** y **exhaustivas**.
- Se cuenta la cantidad de datos en cada intervalo o clase, es decir la **frecuencia**. También se puede usar para cada intervalo la

$$\text{frecuencia relativa} = \frac{\text{frecuencia}}{\text{cantidad total de datos}}$$

- Se grafica el histograma en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área sea proporcional a la frecuencia relativa de dicho intervalo.

Observaciones:

- No existen criterios óptimos para elegir la cantidad de intervalos. En general, entre 8 y 15 intervalos deberían ser suficientes. Utilizar muchos o muy pocos intervalos puede ser poco informativo. Se debe buscar un equilibrio entre un histograma muy irregular y uno demasiado suavizado.
- No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Ésto facilita su interpretación.
- El histograma representa la frecuencia o la frecuencia relativa a través del **área** y no a través de la altura.
- Es recomendable tomar

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

De esta manera el área es 1 y dos histogramas son fácilmente comparables independientemente de la cantidad de observaciones en las que se basa cada uno.

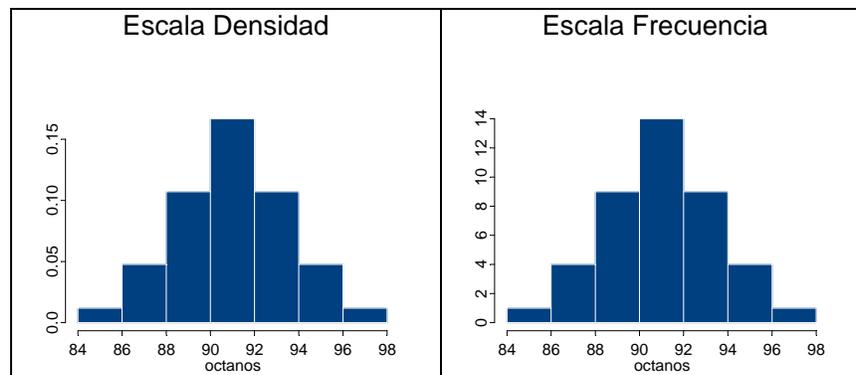
Ejemplo: Los siguientes datos corresponden a Porcentajes de Octanos en Naftas:

85.3	87.5	87.8	88.5	89.9	90.4	91.8	92.7
86.7	87.8	88.2	88.6	90.3	91.0	91.8	93.2
88.3	88.3	89.0	89.2	90.4	91.0	92.3	93.3
89.9	90.1	90.1	90.8	90.9	91.1	92.7	93.4
91.2	91.5	92.6	92.7	93.3	94.2	94.7	94.2
95.6	96.1						

Los agrupamos en 7 clases:

Clase	Frecuencia f_i	Frecuencia relativa fr_i
[84, 86]	1	0.02380952
(86, 88]	4	0.09523810
(88, 90]	9	0.21428571
(90,92]	14	0.33333333
(92,94]	9	0.21428571
(94,96]	4	0.09523810
(96,98]	1	0.02380952
Total	42	1

Histogramas para datos de OCTANOS



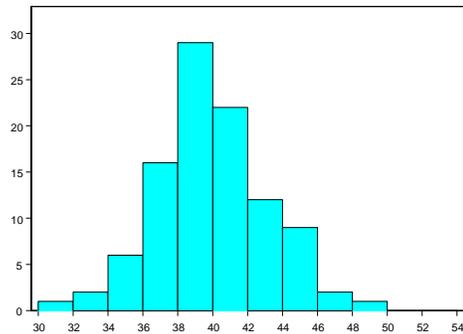
En general, si el histograma es muy irregular puede ser imposible descubrir la forma. En ese caso es conveniente tomar intervalos más anchos.

¿Qué formas puede tener un histograma?

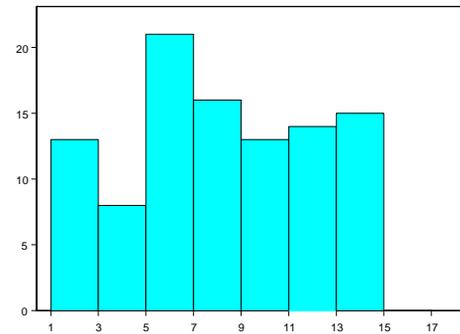
Un aspecto a tener en cuenta en la distribución de los datos es la simetría. Un conjunto de datos que no se distribuye simétricamente, se dice que es **asimétrico**. La asimetría puede verse en el esquema de Tallo y Hoja o en el Histograma y también puede apreciarse a través de la posición relativa entre media y mediana. Más adelante, en un boxplot lo veremos a través de la posición relativa entre la mediana y los cuartos.

En los siguientes gráficos mostramos algunas de las formas posibles que puede tener un histograma:

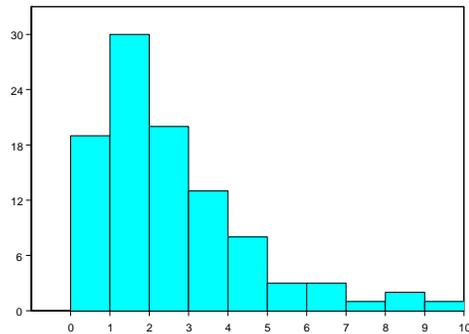
Distribución acampanada



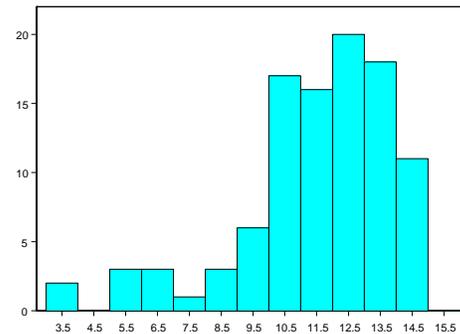
Distribución uniforme



Asimetría a derecha



Asimetría a izquierda



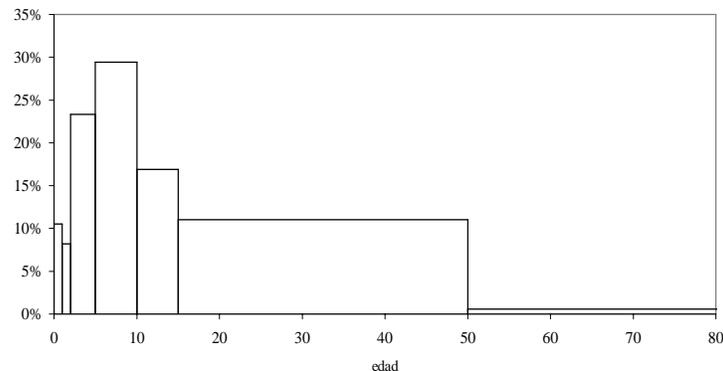
Histograma con intervalos de distinta longitud

Los datos de la siguiente tabla presentan los casos de rubéola notificados al SINAVE durante el año 2000 según grupos de edad. Notemos que los intervalos de edad tienen diferente longitud.

Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

Intervalo (años)	Frecuencia (f_i)	Frecuencia relativa (f_r)
[0, 1)	497	10.5%
[1, 2)	387	8.2%
[2, 5)	1100	23.3%
[5, 10)	1389	29.4%
[10, 15)	798	16.9%
[15, 50)	521	11.0%
≥ 50	28	0.6%
Total	4720	100.00%

Si **erróneamente** se construye un histograma considerando como altura de la barra la frecuencia relativa se obtiene la gráfica siguiente. La última categoría de edad se truncó arbitrariamente en 80 años para poder representarla.



A partir de este gráfico concluiríamos que la proporción de casos es notablemente mayor en los grupos de 2 a 5 años, de 5 a 10 años o de 10 a 15 años que en los grupos de menores de 1 año o de 1 a 2 años. Además, la proporción de casos en el grupo de 15 a 50 años impresiona como notable.

El problema es que en la imagen visual asociamos la frecuencia de casos con el área de la barra, por ello parece haber más notificaciones de gente de 15 a 50 que de cualquier otro grupo de edad.

Recordemos que la barra debe tener una altura tal que el área (base x altura) sea igual a la frecuencia (o a la frecuencia relativa). Es decir,

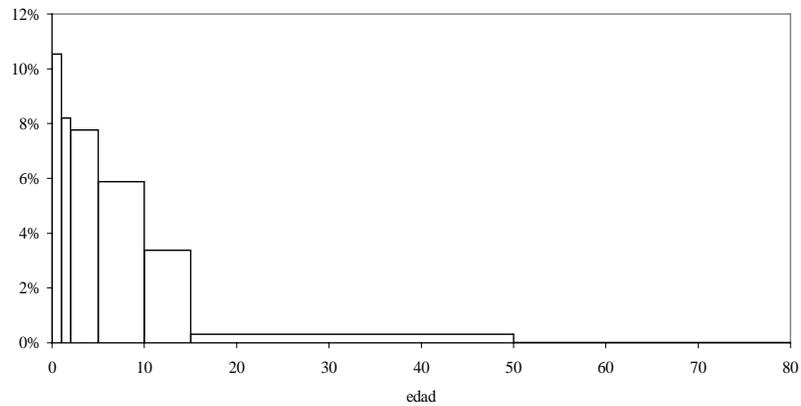
$$\text{altura de la barra} = \frac{\text{frecuencia en el intervalo}}{\text{longitud del intervalo}}.$$

De este modo el área de la barra coincide con la frecuencia en el intervalo. La altura de la barra definida de este modo se denomina *escala densidad* porque indica el número de datos por unidad de la variable. La última columna de la siguiente tabla muestra la escala densidad para los datos de rubéola y la figura siguiente presenta el histograma que se obtiene usando la escala densidad.

Escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000.
 Fuente: SINAVE.

Categoría (años)	Frecuencia (f _i)	Frecuencia relativa (f _r)	Escala densidad
[0, 1)	497	10.5%	10.53%
[1, 2)	387	8.2%	8.20%
[2, 5)	1100	23.3%	7.77%
[5, 10)	1389	29.4%	5.89%
[10, 15)	798	16.9%	3.38%
[15, 50)	521	11.0%	0.32%
≥ 50	28	0.6%	0.01%
Total	4720	100.00%	--

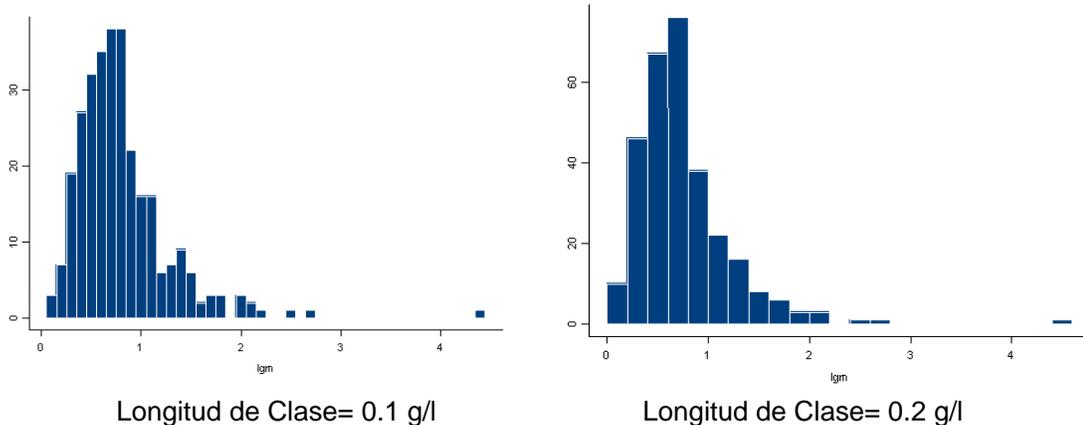
Histograma usando escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

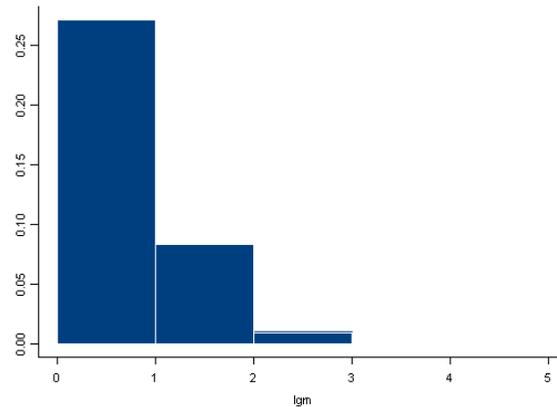


En este gráfico, el porcentaje de casos de rubéola notificados para cada grupo está representado en el *área de la barra*. El histograma muestra que una alta proporción de casos ocurre en menores de 5 años y que la proporción desciende a medida que aumenta la edad. En este gráfico estamos representando la “densidad de notificaciones” por cada año de edad.

El siguiente ejemplo nos muestra cómo varía el aspecto del histograma según la longitud de las clases.

Ejemplo: Concentración de I_{mg}





Longitud de clase=1g/l

Medidas de Resumen

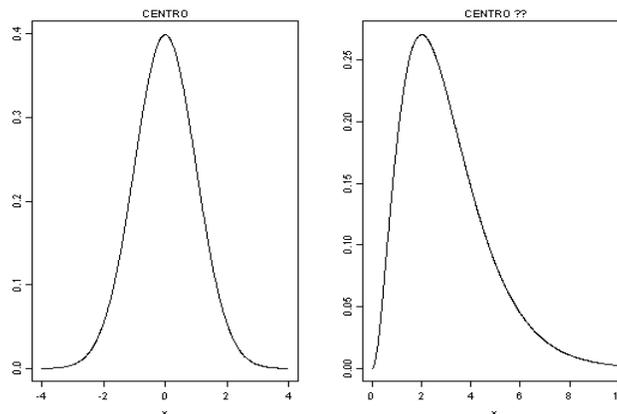
Resumiremos la información de los datos provenientes de variables numéricas mediante medidas de fácil interpretación que reflejen sus características más relevantes. La medida a elegir dependerá de cada problema.

Medidas de Posición o Centrado

Un modo de resumir un conjunto de datos numéricos es a través de un número que represente a todos, en el sentido de ser un valor *típico* para el conjunto.

La pregunta que intentamos responder es: *¿Cuál es el valor central o que mejor representa a los datos?*

Si la distribución es simétrica diferentes medidas darán resultados similares. Si es asimétrica no existe un centro evidente y diferentes criterios para resumir los datos pueden diferir considerablemente, en tanto tratan de captar diferentes aspectos de los mismos.



Supongamos que tenemos un conjunto de n datos que genéricamente representaremos por:

$$x_1, x_2, \dots, x_n$$

Promedio o Media Muestral:

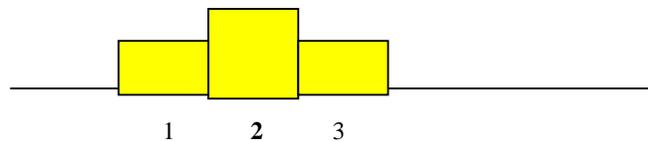
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Es el punto de equilibrio del conjunto de datos.

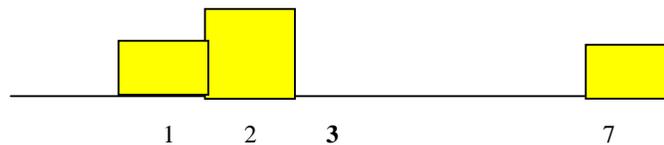
Ejemplo: Fuerza de compresión de cierta aleación de Aluminio-Litio

$$\bar{x} = \frac{\sum_{i=1}^{45} x_i}{45} = \frac{5350}{45} = 118.89$$

Ejemplo: Supongamos que las observaciones son: 1, 2, 2, 3. En este caso $\bar{x} = 2$.



Si reemplazamos el valor 3 por 7, las observaciones son: 1, 2, 2, 7 y $\bar{x} = 3$.



La media muestral es una medida muy sensible a la presencia de datos anómalos (outliers).

Mediana Muestral: Es una medida del centro de los datos en tanto divide a la muestra ordenada en dos partes de igual tamaño. Deja la mitad de los datos a cada lado.

Sean los estadísticos de orden muestrales:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Definimos como mediana

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{si } n = 2k + 1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

La mediana es resistente a la presencia de datos atípicos. También puede ser útil cuando algunos datos han sido censurados.

Ejemplos:

1) Supongamos que los datos son: 3, 5, 2, 4, 6, 8, 7, 7, 6 . Como $n = 9$, $(n+1)/2 = 5$.

Ordenamos la muestra: 2 3 4 5 6 6 7 7 8


 $\tilde{x} = 6$

2) Supongamos que los datos son: 3, 5, 2, 4, 6, 8, 7, 7. Como $n = 8$, $(n+1)/2 = 4.5$ y por lo tanto la mediana muestral es el promedio de las observaciones que ocupan las posiciones 4 y 5 en la muestra ordenada.

2 3 4 5 6 7 7 8

 $\tilde{x} = 5.5$

Ejercicios: 1) Consideremos los dos conjuntos de datos siguientes:

x's: 1,2,2,3	$\bar{x} = 2$	$\tilde{x} = 2$
y's: 1,2,2,7	$\bar{y} = 3$	$\tilde{y} = 2$

¿Qué pasa si, en el segundo caso, se registra 70 en lugar de 7?

2) Dada una muestra de salarios de cierta población, ¿sería más adecuado tomar la media o la mediana muestral para representarla?

Media α - Podada: Es un promedio calculado sobre los datos una vez que se han eliminado $\alpha \cdot 100\%$ de los datos más pequeños y $\alpha \cdot 100\%$ de los datos más grandes. Es una medida intermedia entre la media y la mediana. Formalmente podemos definirla como:

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \dots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

es decir, se obtiene promediando los datos luego de eliminar un número de observaciones en cada extremo de la muestra ordenada igual a la parte entera de $(n\alpha)$.

Otra posible manera de definirla es eliminando $(n\alpha)$ datos en cada extremo si $(n\alpha)$ es entero y, cuando no lo es, interpolando entre dos medias α -podadas, una en la cual se podan $[n\alpha]$ en cada extremo y otra en la que se podan $[n\alpha]+1$ datos en cada extremo.

Ejemplos: 1) Sea el siguiente conjunto de 10 observaciones, ya ordenadas

2 5 8 10 14 17 21 25 28 40

y calculemos la media 0.10-podada. Debemos podar 1 dato en cada extremo y calcular el promedio de los 8 datos restantes, es decir

$$\bar{x}_{0.10} = \frac{5 + 8 + 10 + 14 + 17 + 21 + 25 + 28}{8} = \frac{128}{8} = 16$$

2) Sea el siguiente conjunto de 12 observaciones, ya ordenadas

1 2 5 8 10 14 17 21 25 28 40 45

y calculemos la media 0.10-podada. Usando la definición dada inicialmente, debemos podar $[12 \cdot 0.10] = [1.2] = 1$ dato en cada extremo y calcular el promedio de los 10 datos restantes, es decir

$$\bar{x}_{0.10} = \frac{2 + 5 + 8 + 10 + 14 + 17 + 21 + 25 + 28 + 40}{10} = \frac{170}{10} = 17$$

Con la segunda definición, deberíamos calcular dos medias, una podando una observación en cada extremo de la muestra ordenada y otra podando dos observaciones en cada extremo, e interpolar linealmente entre ambas medias. Es decir, calculamos

$$\begin{aligned}\bar{x}_1 &= \frac{2 + 5 + 8 + 10 + 14 + 17 + 21 + 25 + 28 + 40}{10} = \frac{170}{10} = 17 \\ \bar{x}_2 &= \frac{5 + 8 + 10 + 14 + 17 + 21 + 25 + 28}{8} = \frac{128}{8} = 16\end{aligned}$$

y la media podada se obtiene como la ordenada correspondiente a $x = 1.2$ en la recta que pasa por $(1, 17)$ y $(2, 16)$:

$$\bar{x}_{0.10} = 16.8$$

Observemos que la media es una media α -podada con $\alpha = 0$ y la mediana una media podada con α tan próximo a 0.5 como sea posible. En ese sentido, la media podada es una medida intermedia entre la media y la mediana. Es más resistente a datos atípicos que la media.

¿Cómo elegimos α ?

Dependiendo de cuantos outliers se pretende excluir y de cuán robusta queremos que sea la medida de posición. Como dijimos, cuando seleccionamos $\alpha = 0$ tenemos la media, si elegimos el máximo valor posible para α (lo más cercano posible a 0.5) obtenemos la mediana. Cualquier poda intermedia representa un compromiso entre ambas. Una elección bastante común es $\alpha = 0.10$, que excluye un 20% de los datos.

Ejemplo: En este ejemplo calcularemos las tres medidas resumen. Los datos siguientes, ya ordenados, corresponden al número de pulsaciones por minuto en pacientes con asma durante un espasmo:

40 120 120 125 136 150 150 150 150 167

Las correspondientes medidas son:

$$\bar{x} = 130.8 \quad \tilde{x} = 143 \quad \bar{x}_{0.10} = 137.625$$

Si la distribución es simétrica la mediana y la media identifican al mismo punto. Sin embargo, si la distribución de los datos es asimétrica, esperamos que la relación entre ambas siga el siguiente patrón:

$$\text{Asimetría derecha (cola larga hacia la derecha)} \quad \Rightarrow \quad \bar{x} > \tilde{x}$$

$$\text{Asimetría izquierda (cola larga hacia la izquierda)} \quad \Rightarrow \quad \bar{x} < \tilde{x}$$

La mediana puede ser útil cuando algunos datos son censurados. En estos casos es imposible calcular la media muestral, sin embargo suele ser posible computar la mediana.

Ejemplos: a) Tiempo de supervivencia (en meses) de pacientes con cierta patología. Los datos que se indican entre paréntesis tienen censura a derecha, es decir, se sabe que el paciente sobrevivió ese tiempo, pero no se conoce el tiempo real de supervivencia.

1 5 10 12 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

Como $n = 15$ la mediana es el octavo dato, por lo tanto $\tilde{X} = 28$. Es posible calcularla aunque haya datos censurados, porque los mismos no participan en el cálculo de la mediana. Por ejemplo, aunque no conocemos exactamente el tiempo que sobrevivió el paciente cuyo dato es (45) sabemos que en esta muestra ese dato ocupará el lugar 11 o uno superior.

b) Si, en cambio, los datos son:

1 5 10 (12) 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

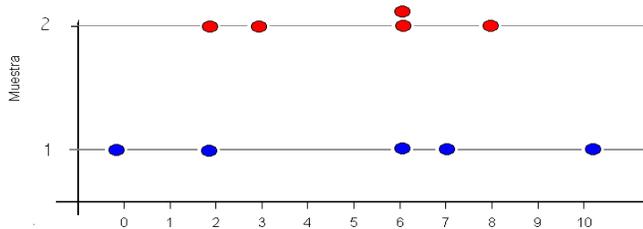
no es posible calcular la mediana debido al dato indicado como (12). Sabemos que este paciente sobrevivió por lo menos 12 meses, pero desconocemos el verdadero valor, el que puede ocupar cualquier posición entre la cuarta y la última.

Medidas de Dispersión o Variabilidad

¿Cuán dispersos están los datos? ¿Cuán cercanos son los datos al valor típico?

Grafiquemos los dos conjuntos de datos siguientes y calculemos para cada uno de ellos su media y su mediana:

x's: 0 2 6 7 10
 y's: 2 3 6 6 8



$$\bar{x} = \bar{y} = 5$$

$$\tilde{x} = \tilde{y} = 6$$

A pesar de tener igual media e igual mediana, los conjuntos de datos difieren ¿Cómo medir la diferencia observada?

Rango Muestral: Es la diferencia entre el valor más grande y el más pequeño de los datos:

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Ejemplo: en nuestros conjuntos de datos:

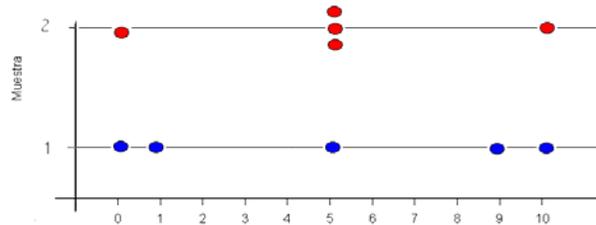
$$\text{Rango}(X) = 10 \quad \text{Rango}(Y) = 6$$

Esta medida es muy sensible a la presencia de outliers. Además no capta la dispersión interna del conjunto de datos.

Veamos otro ejemplo: Sean los siguientes conjuntos de datos

x 's: 0 1 5 9 10

y 's: 0 0 5 5 10



Si calculamos la media, la mediana y el rango muestral de ambos conjuntos, obtenemos:

$$\bar{x} = \bar{y} \quad \tilde{x} = \tilde{y} \quad \text{Rango}(x) = \text{Rango}(y).$$

Es decir, que las 3 medidas coinciden, pero la dispersión no es la misma. Propondremos otra medida de variabilidad.

Varianza Muestral: Mide la variabilidad de los datos alrededor de la media muestral.

$$\text{Varianza muestral} = S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$$\text{Desvío Estándar Muestral} = S = \sqrt{S^2}$$

Ejemplo: En los dos conjuntos de datos anteriores obtenemos:

$$S^2_x = 20.5 \quad S_x = 4.258$$

$$S^2_y = 12.5 \quad S_y = 3.536$$

- El desvío estándar tiene las mismas unidades que los datos, mientras que la varianza no.
- Al basarse en promedios, estas medidas son sensibles a la presencia de datos atípicos. Por ejemplo, si en la muestra de los y 's cambiamos el 10 por un 15 obtenemos $S^2_y = 30$ y $S_y = 5.477$, mientras que si lo cambiamos por un 20 obtenemos $S^2_y = 57.5$ y $S_y = 7.583$.

Coefficiente de Variación: Es una medida que relaciona el desvío standard con la media de una muestra.

$$CV = \frac{S}{\bar{x}}$$

Es una medida que está en desuso, ya que no tiene propiedades estadísticas muy interesantes. Sin embargo no depende de las unidades y si lo multiplicamos por 100 nos da una idea de la variabilidad relativa.

Distancia Intercuartil: Es una medida más resistente que el desvío estándar, basada en el rango de los datos centrales de la muestra.

Comenzaremos por definir los **percentiles**. El percentil $\alpha \cdot 100\%$ de la muestra ($0 < \alpha < 1$) es el valor por debajo del cual se encuentra el $\alpha \cdot 100\%$ de los datos en la muestra ordenada.

Para calcularlo:

- Ordenamos la muestra de menor a mayor
- Buscamos el dato que ocupa la posición $\alpha \cdot (n + 1)$. Si este número no es entero se interpolan los dos adyacentes.

Ejemplo: Consideremos los siguientes 19 datos ordenados:

1 1 2 2 3 4 4 5 5 6 7 7 8 8 9 9 10 10 11

↑ ↑ ↑

Percentil	Posición	Valor	
10%	$0.10(19+1) = 2$	1	
25%	$0.25(19+1) = 5$	3	Cuartil Inferior
50%	$0.50(19+1) = 10$	6	Mediana
75%	$0.75(19+1) = 15$	9	Cuartil Superior
95%	$0.95(19+1) = 19$	11	

Notemos que el percentil 50% (o segundo cuartil) coincide con la mediana. Llamaremos cuartil inferior (o primer cuartil) al percentil 25% y cuartil superior (o tercer cuartil) al percentil 75%.

Los cuartiles y la mediana dividen a la muestra ordenada en cuatro partes igualmente pobladas (aproximadamente un 25 % de los datos en cada una de ellas). Entre los cuartiles se halla aproximadamente el 50% central de los datos y el rango de éstos es:

$$d_i = \text{distancia intercuartil} = \text{cuartil superior} - \text{cuartil inferior}.$$

Observación: Si en el ejemplo cambiáramos el último dato por 110, la distancia intercuartil no cambiaría, mientras que el desvío pasaría de 3.2 a 24.13!!!!

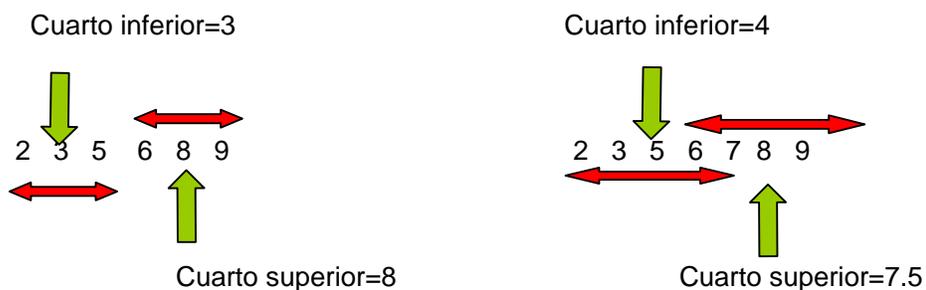
Cuartos y Distancia entre Cuartos: Medidas muy cercanas a los cuartiles inferior y superior son el cuarto inferior y el cuarto superior. Se calculan de la siguiente manera:

- Se ordena la muestra y se calcula la mediana de los datos.
- Dividimos a la muestra ordenada en dos partes: la **primera** corresponde a los datos más pequeños que la mediana y la **segunda** parte a los datos más grandes que la mediana
- Si el tamaño de la muestra es *par*, el **cuarto inferior** es la mediana de la primera mitad, mientras que el **cuarto superior** es la mediana de la segunda mitad.
- Si el tamaño de la muestra es *impar*, a la primera y a la segunda parte se las expande agregándose a cada una de ellas la mediana de todos los datos. El **cuarto inferior** es la mediana de la primera parte expandida y el **cuarto superior** es la mediana de la segunda parte expandida. Es decir, en el caso impar, la mediana interviene en el cómputo de los dos cuartos.

Definimos la distancia entre cuartos como:

$$d_c = \text{distancia entre cuartos} = \text{cuarto superior} - \text{cuarto inferior.}$$

Ejemplo: Sean las siguientes muestras ordenadas



Desvío Absoluto Mediano (Desviación absoluta respecto de la Mediana): Es una versión robusta del desvío estándar basada en la mediana. Definimos la MAD como:

$$MAD = \text{mediana}(|x_i - \tilde{x}|)$$

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana.
- Calculamos la distancia de cada dato a la mediana.
- Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
- Buscamos la mediana de las distancias sin signo.

Observación: Si deseamos comparar la distancia intercuartil y la MAD con el desvío standard es conveniente dividir las por constantes adecuadas. En ese caso se compara a S con

$$\frac{MAD}{0.675} \quad \text{ó} \quad \frac{d_i}{1.35}$$

Números de Resumen: Los 5 números de resumen de la distribución de un conjunto de datos consisten en el **mínimo**, el **cuartil inferior**, la **mediana**, el **cuartil superior** y el **máximo**.

Ejemplo: Los siguientes datos corresponden a tiempos de CPU (en segundos) de 25 trabajos enviados a un server y seleccionados al azar.

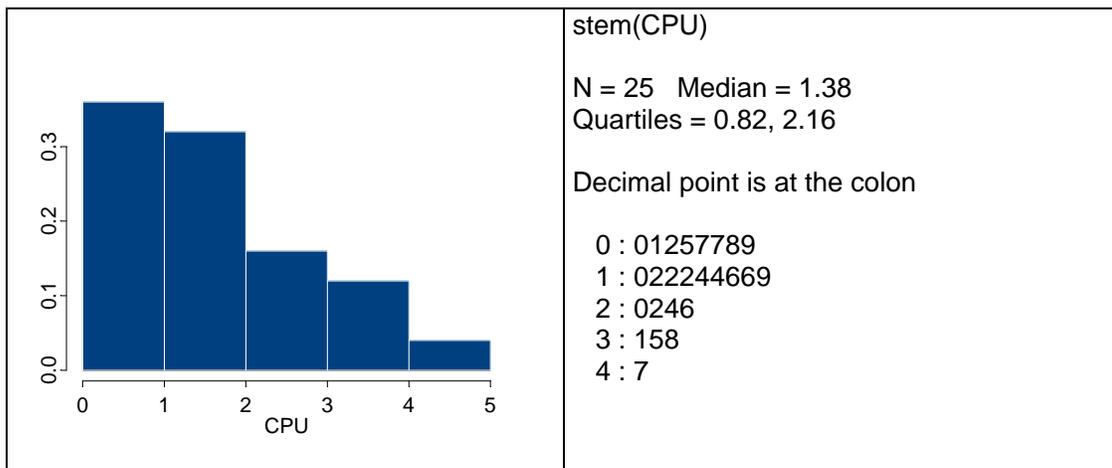
CPU				
1.17	1.23	0.15	0.19	0.92
1.61	3.76	2.41	0.82	0.75
1.16	1.94	0.71	0.47	2.59
1.38	0.96	0.02	2.16	3.07
3.53	4.75	1.59	2.01	1.40

Calculamos los 5 números resumen y la media muestral para este conjunto de datos, utilizando el software R.

```
> summary(server1)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.02 0.82    1.38  1.63   2.16  4.75
```

Realizamos un esquema de Tallo y Hoja y graficamos un histograma para este conjunto de datos:



Todas las medidas y los gráficos muestran que se trata de una distribución asimétrica con cola a derecha.

Box-Plots

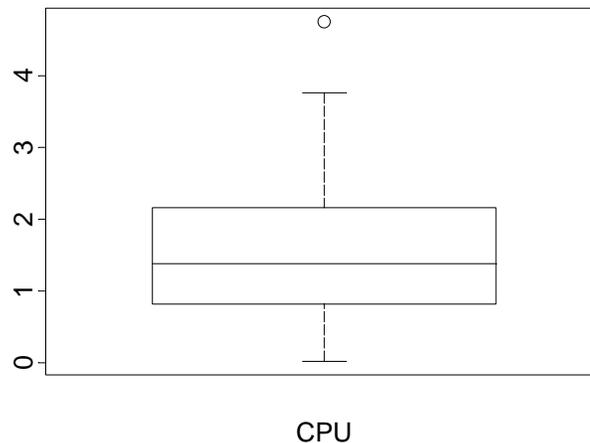
Con las medidas anteriores podemos construir un gráfico de fácil realización y lectura.

¿Cómo lo hacemos? Vamos a dar una versión, pero vale la pena advertir que hay variaciones de un programa a otro.

1. Representamos una escala vertical u horizontal
2. Dibujamos una caja cuyos extremos son los cuartiles y dentro de ella un segmento que corresponde a la mediana.
3. A partir de cada extremo dibujamos un segmento hasta el dato más alejado que está a lo sumo $1.5 d_I$ del extremo de la caja. Estos segmentos se llaman bigotes.
4. Marcamos con * a aquellos datos que están entre $1.5 d_I$ y $3 d_I$ de cada extremo y con \circ a aquellos que están a más de $3 d_I$ de cada extremo. Algunos paquetes, como el R, indican a todos los outliers de la misma forma.

Observación: Muchos paquetes estadísticos realizan el boxplot usando los cuartos y la distancia entre cuartos en lugar de la distancia intercuartil. Como estas medidas son muy próximas, en general los resultados son análogos. Lo importante es que entre los cuartos o entre los cuartiles yace aproximadamente el 50% central de los datos.

Ejemplo: El box-plot correspondiente a los tiempos de CPU es el siguiente



Es interesante observar que en el boxplot se indica a uno de los datos como outlier, mientras que en el análisis anterior esto no parecía evidente.

A partir de un box-plot podemos apreciar los siguientes aspectos de la distribución de un conjunto de datos:

- posición
- dispersión
- asimetría
- longitud de las colas
- puntos anómalos o outliers.

Los box-plots son especialmente útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.

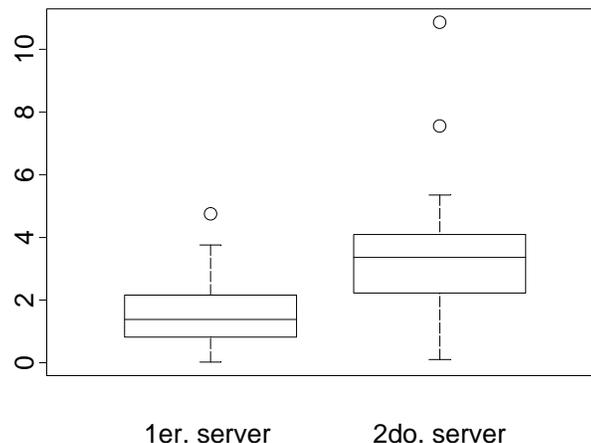
Outliers: Los métodos que hemos visto nos permiten identificar puntos atípicos que pueden aparecer en una o más variables. Su detección es importante pues pueden determinar o influenciar fuertemente los resultados de un análisis estadístico clásico, dado que muchas de las técnicas habitualmente usadas son muy sensibles a la presencia de datos atípicos.

Los outliers deben ser cuidadosamente inspeccionados. Si no hay evidencia de error y su valor es posible no deberían ser eliminados. Asimismo, la presencia de outliers puede indicar que la escala elegida no es la más adecuada.

Boxplots Paralelos

Una aplicación muy útil de los boxplots es la comparación de la distribución de dos o más conjuntos de datos graficando en una escala común los boxplots de cada una de las muestras. En este sentido los boxplots se muestran como un método muy efectivo de presentar y resumir los datos, tal como veremos en el siguiente ejemplo.

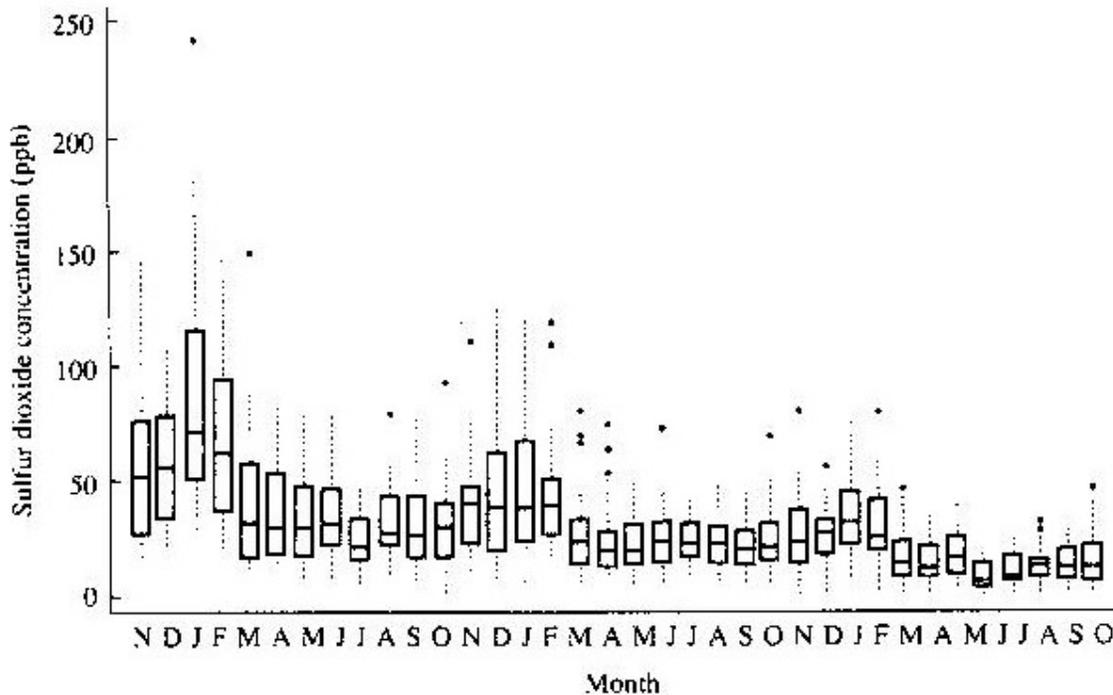
Ejemplo: Supongamos que se dispone de otros 25 datos correspondientes a tiempos de CPU enviados a otro server. Si realizamos boxplots paralelos para ambos conjuntos de datos obtenemos el siguiente gráfico. La simple comparación de los boxplots obtenidos revela que los trabajos enviados al segundo server son más largos. De hecho, el 75% de los trabajos muestreados en el segundo server tienen tiempos de CPU mayores que el cuartil superior de los trabajos muestreados en el primer server.



Ejemplo: Los siguientes boxplots corresponden a datos de concentración máxima diaria, en partes por mil millones de dióxido de azufre en Bayonne, en el estado de Nueva Jersey, desde noviembre de 1969 hasta octubre de 1972 agrupados por meses. Hay 36 grupos de datos, cada uno de tamaño aproximadamente 30.

Los boxplots muestran algunas características de estos datos en forma muy rápida.

Hay una reducción general de la concentración de dióxido de azufre a lo largo del tiempo debida a la conversión gradual en la zona al uso de combustibles con baja concentración de azufre. Esta disminución es más fuerte para los cuartiles superiores. También se muestran concentraciones más elevadas para los meses de invierno debido al uso de calderas a petróleo. Claramente se ve un efecto cíclico y amortiguado. Los boxplots muestran una distribución asimétrica a derecha, con presencia de outliers en algunos meses, y que la dispersión de la distribución es mayor cuando el nivel general de la concentración es más alto.



QQ-plot (Normal Probability Plot): El QQ-plot es un gráfico que nos sirve para evaluar la cercanía a una distribución dada, en particular a la distribución normal.

Consideremos la muestra aleatoria: X_1, X_2, \dots, X_n y los correspondientes estadísticos de orden

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Observemos que $X_{(1)} = \min(X_1, X_2, \dots, X_n)$, mientras que $X_{(n)} = \max(X_1, X_2, \dots, X_n)$.

En particular, si U_1, U_2, \dots, U_n son v.a. i.i.d tales que $U_i \sim U(0,1)$, se puede demostrar que

$$E(U_{(i)}) = \frac{i}{n+1}.$$

Por lo tanto esperamos que, si la distribución subyacente fuese Uniforme y graficásemos $U_{(1)}, \dots, U_{(n)}$ vs sus valores esperados $\frac{1}{n+1}, \dots, \frac{n}{n+1}$, el gráfico debería parecerse a una recta.

Por otro lado, sabemos que si X es una variable continua con función de distribución F estrictamente creciente, entonces

$$Y = F(X) \sim U(0,1)$$

Esto sugiere que si suponemos que $X_i \sim F$, entonces podemos graficar

$$F(X_{(i)}) \quad \text{vs} \quad \frac{i}{n+1}$$

o equivalentemente

$$X_{(i)} \quad \text{vs} \quad F^{-1}\left(\frac{i}{n+1}\right).$$

Observemos que si F es de la forma

$$F(x) = G\left(\frac{x - \mu}{\sigma}\right),$$

o sea, si depende de un parámetro de posición y uno de escala, como es el caso de la normal, podemos graficar

$$\frac{X_{(i)} - \mu}{\sigma} \quad \text{vs} \quad G^{-1}\left(\frac{i}{n+1}\right)$$

o bien

$$X_{(i)} \quad \text{vs} \quad G^{-1}\left(\frac{i}{n+1}\right)$$

Como,

$$X_{(i)} \cong \sigma \cdot G^{-1}\left(\frac{i}{n+1}\right) + \mu$$

el gráfico será aproximadamente una recta.

Notemos que si F^{-1} es la inversa de F , entonces el p -ésimo percentil de F , x_p , es tal que

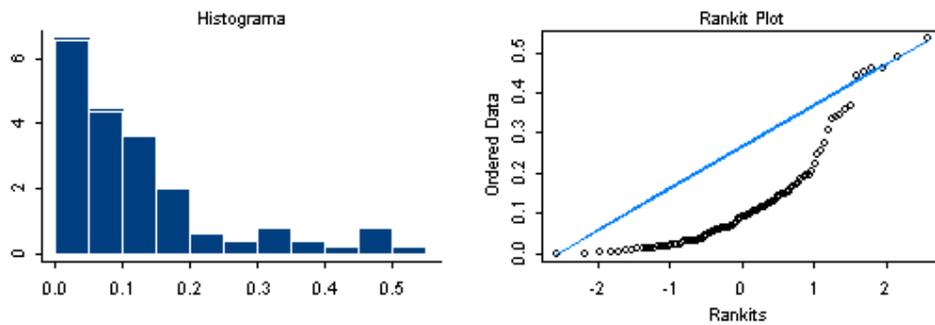
$$F(x_p) = p \Rightarrow x_p = F^{-1}(p)$$

por lo tanto, $F^{-1}\left(\frac{i}{n+1}\right)$ es el $\frac{i}{n+1}$ -percentil de F .

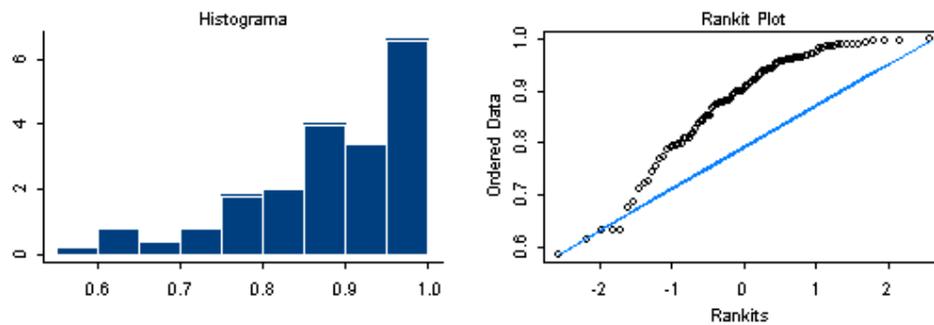
En el QQ-plot se grafican en el eje de abscisas los percentiles de la distribución teórica (en nuestro caso normal) y en el eje de ordenadas las observaciones ordenadas, que pueden ser vistas como percentiles empíricos.

En los siguientes gráficos ilustramos el uso de estas técnicas gráficas con algunos ejemplos. Cabe observar que algunos paquetes estadísticos representan a los percentiles teóricos de la distribución normal en el eje de abscisas y otros en el eje de ordenadas

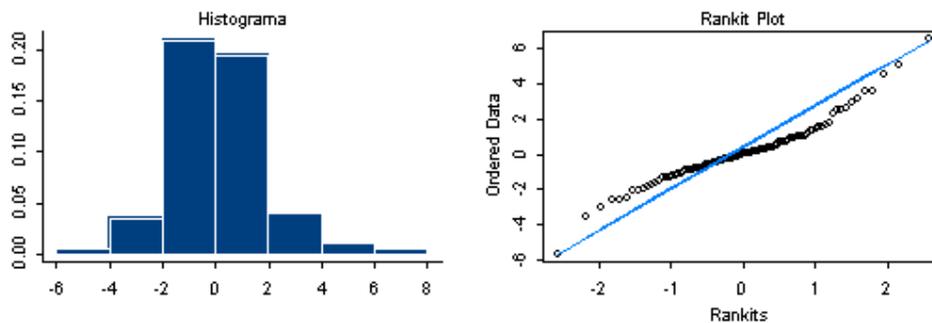
Asimétrica a Derecha



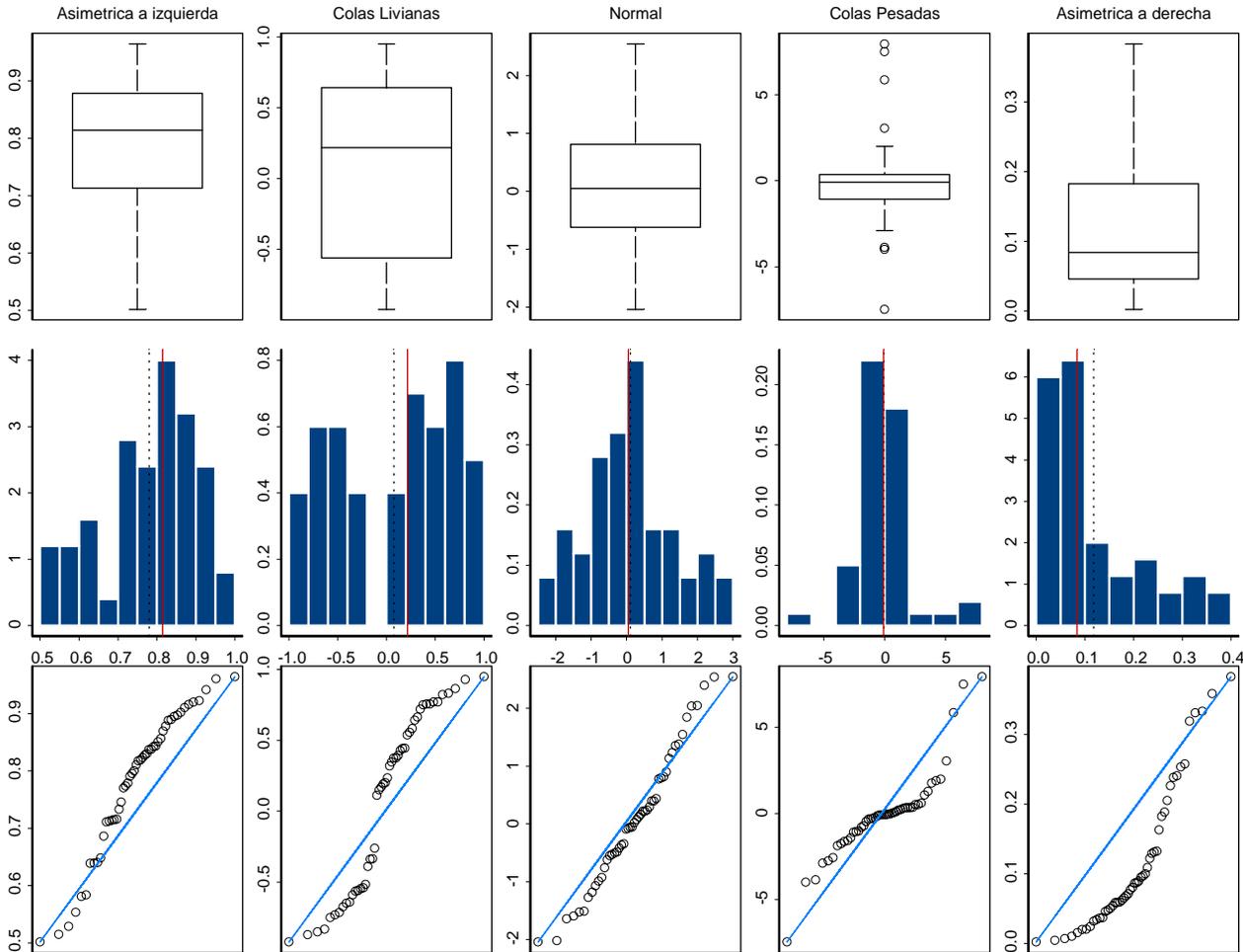
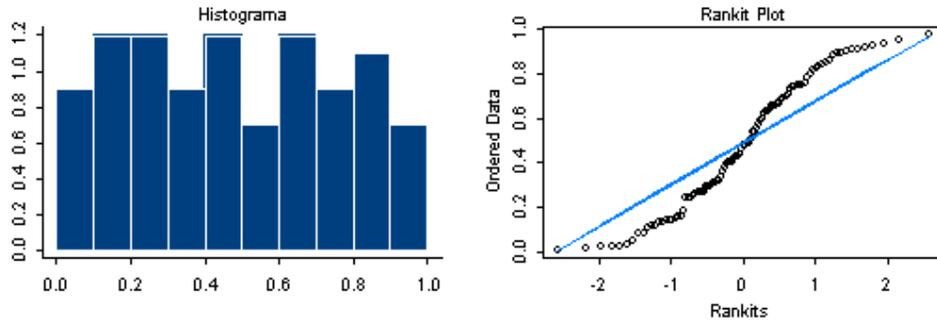
Asimétrica a Izquierda



Simétrica con Colas Pesadas



Simétrica con Colas Livianas



Inferencia estadística - Estimación puntual

La estadística provee técnicas que permiten obtener conclusiones generales a partir de un conjunto limitado – pero representativo – de datos. Cuando inferimos no tenemos garantía de que la conclusión que obtenemos sea exactamente correcta. Sin embargo, la estadística permite cuantificar el error asociado a la estimación.

La mayoría de las distribuciones de probabilidad dependen de cierto número de parámetros. Por ejemplo: $P(\lambda)$, $N(\mu, \sigma^2)$, $Bi(n, p)$, etc. Salvo que estos parámetros se conozcan, deben estimarse a partir de los datos.

El objetivo de la *estimación puntual* es usar una muestra para obtener números que, en algún sentido, sean los que mejor representan a los verdaderos valores de los parámetros de interés.

Supongamos que se selecciona una muestra de tamaño n de una población. Antes de obtener la muestra no sabemos cuál será el valor de cada observación. Así, la primera observación puede ser considerada una v.a. X_1 , la segunda una v.a. X_2 , etc. Por lo tanto, antes de obtener la muestra denotaremos X_1, X_2, \dots, X_n a las observaciones y, una vez obtenida la muestra, denotaremos x_1, x_2, \dots, x_n a los valores observados.

Del mismo modo, antes de obtener una muestra, cualquier función de ella será una v.a., por ejemplo: $\bar{X}, \tilde{X}, S^2, \max(X_1, \dots, X_n)$, etc. Una vez obtenida la muestra los valores calculados serán denotados $\bar{x}, \tilde{x}, s^2, \max(x_1, \dots, x_n)$, etc.

Definición: Un estimador puntual de un parámetro θ es un valor que puede ser considerado representativo de θ y se indicará $\hat{\theta}$. Se obtiene a partir de alguna función de la muestra.

Ejemplo: Con el fin de estudiar si un dado es o no equilibrado, se arroja el dado 100 veces en forma independiente, obteniéndose 21 ases. ¿Qué valor podría utilizarse, en base a esa información, como estimación de la probabilidad de ases? Parece razonable utilizar la frecuencia relativa de ases.

En este caso, si llamamos p a la probabilidad que queremos estimar, $\hat{p} = \frac{21}{100} = 0.21$

Métodos de estimación puntual

¿Cómo obtener estimadores para un problema dado? Estudiaremos dos métodos que proporcionan estimadores puntuales: el método de momentos y el método de máxima verosimilitud.

Método de momentos: La idea básica consiste en igualar ciertas características muestrales con las correspondientes características poblacionales. Recordemos la siguiente definición.

Definición: Sea X una v.a. con función de probabilidad puntual $p_X(x)$ en el caso discreto o función de densidad $f_X(x)$ en el caso continuo. Se denomina **momento de orden k** ($k \in \mathbb{N}$) o momento poblacional de orden k a $E(X^k)$, es decir

$$E(X^k) = \begin{cases} \sum x^k p_X(x) & \text{en el caso discreto} \\ \int_{-\infty}^{\infty} x^k f_X(x) dx & \text{en el caso continuo} \end{cases}$$

si esas esperanzas existen.

Como ya hemos visto cuando estudiamos función generadora de momentos de una variable aleatoria, los momentos están relacionados con los parámetros de la distribución asociada.

Definición: Dada una muestra aleatoria X_1, X_2, \dots, X_n , se denomina **momento muestral de orden k** a

$$\frac{\sum_{i=1}^n X_i^k}{n}$$

Definición: Sea X_1, X_2, \dots, X_n una m.a. de una distribución con función de probabilidad puntual o función de densidad que depende de m parámetros $\theta_1, \theta_2, \dots, \theta_m$. Los estimadores de momentos de $\theta_1, \theta_2, \dots, \theta_m$ son los valores $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ que se obtienen igualando m momentos poblacionales con los correspondientes momentos muestrales. En general, se obtienen resolviendo el siguiente sistema de ecuaciones

$$\frac{\sum_{i=1}^n X_i^k}{n} = E(X^k) \quad k = 1, 2, \dots, m$$

Ejemplos: 1) Sea X_1, X_2, \dots, X_n una m.a. de una distribución exponencial de parámetro λ . Como hay un solo parámetro a estimar, basta plantear una ecuación basada en el primer momento.

$$\frac{\sum_{i=1}^n X_i}{n} = E(X) \Rightarrow \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{\lambda} \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i} \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}$$

2) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $\Gamma(\alpha, \lambda)$. Como hay dos parámetros a estimar, planteamos un sistema de ecuaciones basadas en el primer y en el segundo momento.

Usando que si $X \sim \Gamma(\alpha, \lambda)$, $E(X) = \frac{\alpha}{\lambda}$ y $V(X) = \frac{\alpha}{\lambda^2}$ y la relación:
 $V(X) = E(X^2) - (E(X))^2$,

$$\begin{cases} \frac{\sum_{i=1}^n X_i}{n} = E(X) \\ \frac{\sum_{i=1}^n X_i^2}{n} = E(X^2) \end{cases} \Rightarrow \begin{cases} \frac{\sum_{i=1}^n X_i}{n} = \frac{\alpha}{\lambda} \\ \frac{\sum_{i=1}^n X_i^2}{n} = \frac{\alpha}{\lambda^2} + \left(\frac{\alpha}{\lambda}\right)^2 \end{cases}$$

Reemplazando $\frac{\alpha}{\lambda} = \bar{X}$, en la segunda ecuación, se obtiene:

$$\frac{\sum_{i=1}^n X_i^2}{n} = \frac{\bar{X}}{\lambda} + \bar{X}^2$$

y, despejando λ :

$$\frac{\bar{X}}{\lambda} = \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \Rightarrow \hat{\lambda} = \frac{\bar{X}}{\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2}$$

Finalmente, reemplazando el estimador de λ en la primera ecuación, obtenemos el estimador de α :

$$\hat{\alpha} = \frac{\bar{X}^2}{\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2}$$

3) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(0, \theta)$. Como hay un único parámetro a estimar, planteamos una ecuación basada en el primer momento.

$$\frac{\sum_{i=1}^n X_i}{n} = E(X) = \frac{\theta}{2} \quad \Rightarrow \quad \hat{\theta} = 2\bar{X}$$

4) Veamos por último un ejemplo que nos muestra que no siempre podemos utilizar los momentos en el orden natural. Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(-\theta, \theta)$. Como hay un único parámetro a estimar, parece natural plantear una ecuación basada en el primer momento. Sin embargo, si lo hacemos,

$$\frac{\sum_{i=1}^n X_i}{n} = E(X) = 0$$

Observamos que el primer momento poblacional no depende de θ y por lo tanto no podemos despejar a partir de esta ecuación el estimador del parámetro. En este caso, es necesario plantear una ecuación basada en el segundo momento:

$$\frac{\sum_{i=1}^n X_i^2}{n} = E(X^2) = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3} \quad \Rightarrow \quad \hat{\theta} = \sqrt{\frac{3 \sum_{i=1}^n X_i^2}{n}}$$

Método de máxima verosimilitud: Este método fue introducido por Fisher en la década de 1920. Se basa en la idea de hallar los valores de los parámetros que hacen que la probabilidad de obtener una muestra dada sea máxima.

Ejemplo: Se realiza una encuesta de opinión a una m.a. de 20 personas. Se les formula una única pregunta que será respondida por Si o por NO. Sean X_1, X_2, \dots, X_{20} las v.a. correspondientes a la respuesta, tales que

$$X_i = \begin{cases} 1 & \text{si la persona } i \text{ responde SI} \\ 0 & \text{si la persona } i \text{ responde NO} \end{cases}$$

para $i=1, 2, \dots, 20$ y sea $p = P(X_i = 1)$.

Observemos que las v.a. X_i son independientes y cada una de ellas tiene distribución $\text{Bi}(1,p)$. Entonces, la función de probabilidad conjunta del vector $(X_1, X_2, \dots, X_{20})$ es

$$p(x_1, x_2, \dots, x_{20}) = p^{x_1} (1-p)^{1-x_1} p^{x_2} (1-p)^{1-x_2} \dots p^{x_{20}} (1-p)^{1-x_{20}}$$

Si en la muestra obtenida se observan 7 NO's (0) y 13 SI's (1), sería

$$p(x_1, x_2, \dots, x_{20}) = p^{13} (1-p)^7$$

La pregunta es: ¿qué valor de p hace que los valores muestrales obtenidos sean los más probables?

Es decir, buscamos el valor de p que hace máxima $p(x_1, x_2, \dots, x_{20})$ o equivalentemente $\ln p(x_1, x_2, \dots, x_{20})$ ya que \ln es una función monótona creciente. Debemos maximizar la siguiente función de p

$$g(p) = \ln p(x_1, x_2, \dots, x_{20}) = 13 \ln(p) + 7 \ln(1-p)$$

Para ello, como esta función es derivable respecto de p , buscamos los posibles puntos críticos, igualando a 0 la derivada primera.

$$0 = \frac{\partial g(p)}{\partial p} = \frac{13}{p} - \frac{7}{1-p} = \frac{13(1-p) - 7p}{p(1-p)} = \frac{13 - 20p}{p(1-p)} \Leftrightarrow 13 - 20p = 0 \Leftrightarrow \hat{p} = \frac{13}{20}$$

Este valor es en efecto el que maximiza $g(p)$ pues

$$\left. \frac{\partial^2 g(p)}{\partial p^2} \right|_{p=13/20} = -\frac{13}{p^2} - \frac{7}{(1-p)^2} \Big|_{p=13/20} < 0$$

Definición: Sean X_1, X_2, \dots, X_n v.a. con función de probabilidad conjunta $p_{\vec{X}}(x_1, x_2, \dots, x_n)$ o función de densidad conjunta $f_{\vec{X}}(x_1, x_2, \dots, x_n)$ que depende de m parámetros $\theta_1, \theta_2, \dots, \theta_m$. Cuando (x_1, x_2, \dots, x_n) son los valores observados y la función

de probabilidad o de densidad conjunta se considera función de los parámetros $\theta_1, \theta_2, \dots, \theta_m$, se denomina función de verosimilitud y se denota $L(\theta_1, \theta_2, \dots, \theta_m)$.

Los estimadores de máxima verosimilitud (**EMV**) de $\theta_1, \theta_2, \dots, \theta_m$ son los valores $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ que maximizan la función de verosimilitud, o sea los valores tales que

$$L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m) \geq L(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_m) \quad \forall \tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_m$$

La forma general de los EMV se obtiene reemplazando los valores observados x_i por las v.a. X_i .

Ejemplos: 1) Sea X_1, X_2, \dots, X_n una m.a. de una distribución exponencial de parámetro λ .

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

por lo tanto, la función de verosimilitud es

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

Observemos que no incluimos los indicadores porque, dado que el rango de la v.a. no depende del parámetro a estimar, podemos suponer que todas las observaciones son no negativas.

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n x_i$$

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \quad \Leftrightarrow \quad \hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}$$

Verificar que el punto crítico obtenido es en efecto un máximo.

Observemos que en este caso el EMV coincide con el de momentos.

2) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$.

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

Por lo tanto la función de verosimilitud es

$$L(\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

y maximizarla equivale a maximizar su logaritmo

$$\ln L(\mu, \sigma) = -n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^n (x_i - \mu) = 0 \\ -n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \mu = \frac{\sum_{i=1}^n x_i}{n} \\ \sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \end{cases}$$

y, reemplazando el valor estimado de μ en la segunda ecuación, se obtienen los EMV de los parámetros

$$\hat{\mu} = \bar{X} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

3) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(0, \theta)$.

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\theta} I_{(0, \theta)}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^n I_{(0, \theta)}(x_i)$$

y la función de verosimilitud es

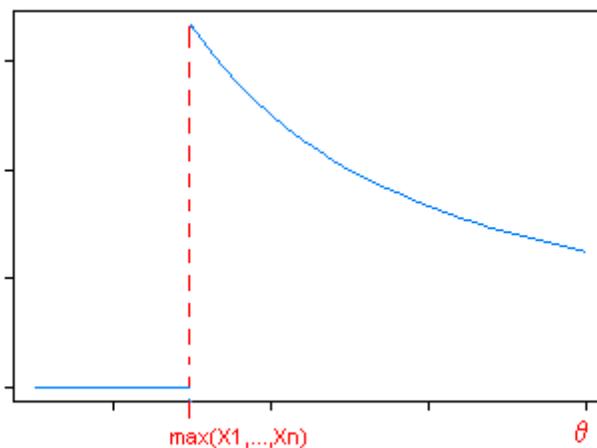
$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^n I_{(0, \theta)}(x_i)$$

Observemos que, en este caso, no es posible tomar logaritmo ni derivar porque el parámetro (argumento de la función de verosimilitud) determina el soporte de la densidad. Analicemos cómo es esta función para hallar su máximo

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } 0 < x_i < \theta \quad \forall i \\ 0 & \text{en caso contrario} \end{cases} = \begin{cases} \frac{1}{\theta^n} & \text{si } \max_{1 \leq i \leq n}(x_i) < \theta \\ 0 & \text{en caso contrario} \end{cases}$$

$$= \begin{cases} \frac{1}{\theta^n} & \text{si } \theta > \max_{1 \leq i \leq n}(x_i) \\ 0 & \text{si } \theta \leq \max_{1 \leq i \leq n}(x_i) \end{cases}$$

Grafiquemos $L(\theta)$ como función de θ .



Como se puede observar, el máximo de la función de verosimilitud se alcanza en $\theta = \max_{1 \leq i \leq n}(x_i)$ y por lo tanto el EMV del parámetro es

$$\hat{\theta} = \max_{1 \leq i \leq n}(X_i)$$

Propiedad de Invarianza de los EMV: Sea $\hat{\theta}$ el EMV de θ y sea h una función inyectiva con dominio en el rango de valores posibles de θ , entonces el EMV de $h(\theta)$ es $h(\hat{\theta})$. Por ejemplo, en el caso de una m.a. de una distribución $N(\mu, \sigma^2)$ hemos visto que el EMV de σ es

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

entonces el EMV de σ^2 es

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

pues la función $h(x)=x^2$ es inyectiva si su dominio se restringe a los reales positivos, es decir si $h : \mathfrak{R}^{\geq 0} \rightarrow \mathfrak{R}$.

En general, sean $\hat{\theta}_1, \dots, \hat{\theta}_m$ los EMV de $\theta_1, \dots, \theta_m$ y sea una función $h : \mathfrak{R}^m \rightarrow \mathfrak{R}$, ¿bajo qué condiciones el EMV de $h(\theta_1, \dots, \theta_m)$ es $h(\hat{\theta}_1, \dots, \hat{\theta}_m)$? Esta propiedad, denominada propiedad de invarianza de los EMV, se cumple si la función h puede ser completada a una función inyectiva.

Propiedades de los estimadores y criterios de selección

Observemos que, dada una muestra X_1, X_2, \dots, X_n , donde $X_i \sim F_\theta$, un estimador puntual del parámetro θ , obtenido en base a ella, es una v.a. $\hat{\theta}$. La diferencia

$$\hat{\theta} - \theta$$

es el error de estimación y una estimación será más precisa cuanto menor sea este error. Este error es también una v.a. dado que depende de la muestra obtenida. Para algunas muestras será positivo, para otras negativo. Una propiedad deseable es que la esperanza del error sea 0, es decir que “en promedio” el error obtenido al estimar a partir de diferentes muestras sea cero.

Definición: Un estimador puntual $\hat{\theta}$ del parámetro θ es **insesgado** si

$$E_\theta(\hat{\theta}) = \theta \quad \forall \theta$$

Si $\hat{\theta}$ no es insesgado, se denomina sesgo de $\hat{\theta}$ a $b(\hat{\theta}) = E_{\theta}(\hat{\theta}) - \theta$.

Por lo tanto, un estimador es insesgado si su distribución tiene como valor esperado al parámetro a estimar.

Definición: Un estimador puntual $\hat{\theta}$ del parámetro θ basado en una muestra X_1, \dots, X_n , es **asintóticamente insesgado** si

$$E_{\theta}(\hat{\theta}) \xrightarrow{n \rightarrow \infty} \theta \quad \forall \theta$$

Ejemplos: 1) Sea X : número de éxitos en n repeticiones de un experimento binomial con probabilidad de éxito igual a p . Entonces $X \sim \text{Bi}(n, p)$ y hemos visto que el EMV de p es $\hat{p} = X/n$, o sea la frecuencia relativa de éxitos. Verifiquemos que este estimador es insesgado.

$$E_P(\hat{p}) = E_P\left(\frac{X}{n}\right) = \frac{E_P(X)}{n} = \frac{np}{n} = p \quad \forall p$$

y, por lo tanto, es insesgado.

2) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$. Los EMV de μ y σ^2 son

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Como $E_{\mu, \sigma^2}(\hat{\mu}) = \mu \quad \forall \mu$, este estimador es insesgado.

Verifiquemos que el estimador de la varianza no lo es.

$$\begin{aligned} E_{\mu, \sigma^2}(\hat{\sigma}^2) &= E_{\mu, \sigma^2} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \right) = \frac{1}{n} E_{\mu, \sigma^2} \left(\sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \right) \\ &= \frac{1}{n} E_{\mu, \sigma^2} \left(\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \right) = \frac{1}{n} E_{\mu, \sigma^2} \left(\sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \right) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} E_{\mu, \sigma^2} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \frac{1}{n} E_{\mu, \sigma^2} \left(\sum_{i=1}^n X_i^2 \right) - E_{\mu, \sigma^2} (\bar{X}^2) = \frac{n}{n} E_{\mu, \sigma^2} (X_1^2) - E_{\mu, \sigma^2} (\bar{X}^2) \\
 &= \left[V_{\mu, \sigma^2} (X_1) + \left(E_{\mu, \sigma^2} (X_1) \right)^2 \right] - \left[V_{\mu, \sigma^2} (\bar{X}) + \left(E_{\mu, \sigma^2} (\bar{X}) \right)^2 \right] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n} \sigma^2
 \end{aligned}$$

Por lo tanto el EMV de la varianza no es insesgado, pero es asintóticamente insesgado ya que su esperanza tiende a σ^2 cuando el tamaño de la muestra tiende a infinito.

Ejercicio: Verificar que la varianza muestral $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ es un estimador insesgado de la varianza poblacional cualquiera sea la distribución.

3) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(0, \theta)$. El estimador de momentos de θ es $\hat{\theta} = 2\bar{X}$ y el EMV es $\hat{\theta} = \max_{1 \leq i \leq n} (X_i)$.

El estimador de momentos es insesgado. En efecto,

$$E_{\theta}(\hat{\theta}) = 2E_{\theta}(\bar{X}) = 2 \frac{\theta}{2} = \theta \quad \forall \theta$$

Verificaremos que el EMV no lo es. Para ello, necesitamos obtener la densidad de la v.a. $U = \max_{1 \leq i \leq n} (X_i)$.

Recordemos que, si X_1, X_2, \dots, X_n es una m.a. de una distribución $U(0, \theta)$, entonces

$$F_U(u) = (F_X(u))^n = \begin{cases} 0 & \text{si } u \leq 0 \\ \left(\frac{u}{\theta}\right)^n & \text{si } 0 < u < \theta \\ 1 & \text{si } u \geq \theta \end{cases}$$

entonces

$$f_U(u) = n \left(\frac{u}{\theta}\right)^{n-1} \frac{1}{\theta} I_{(0, \theta)}(u).$$

Calculemos la esperanza del EMV.

$$E_{\theta}(\max(X_i)) = E_{\theta}(U) = \int_0^{\theta} u n \left(\frac{u}{\theta}\right)^{n-1} \frac{1}{\theta} du = \frac{n}{\theta^n} \int_0^{\theta} u^n du = \frac{n}{\theta^n} \frac{u^{n+1}}{n+1} \Big|_0^{\theta} = \frac{n}{n+1} \theta$$

Entonces, el EMV no es insesgado pero es asintóticamente insesgado.

Cuando hay más de un estimador insesgado para un mismo parámetro, ¿cómo decidimos cuál conviene usar? Por ejemplo, sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$. Es inmediato verificar que los siguientes son todos estimadores insesgados de μ :

$$\begin{aligned}\hat{\mu}_1 &= \bar{X} \\ \hat{\mu}_2 &= \frac{X_1 + X_2}{2} \\ \hat{\mu}_3 &= X_1\end{aligned}$$

Las varianzas de estos estimadores son

$$\begin{aligned}V_{\mu, \sigma^2}(\hat{\mu}_1) &= \frac{\sigma^2}{n} \\ V_{\mu, \sigma^2}(\hat{\mu}_2) &= \frac{\sigma^2}{2} \\ V_{\mu, \sigma^2}(\hat{\mu}_3) &= \sigma^2\end{aligned}$$

y parece natural elegir el estimador más preciso, es decir el de menor varianza.

Principio de estimación insesgada de mínima varianza: Entre todos los estimadores insesgados de θ , elegir el de menor varianza. El estimador resultante se denomina **IMVU** (insesgado de mínima varianza uniformemente). Existe una metodología que permite hallar estimadores IMVU en muchas situaciones.

Teorema: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$. Entonces \bar{X} es estimador IMVU de μ .

A partir de este resultado deducimos que, si se tiene evidencia de que la m.a. proviene de una distribución Normal, parece conveniente usar \bar{X} como estimador de μ . Sin embargo, si los datos no son Normales este estimador podría llegar a ser una pésima elección.

Ejemplo: Sean las siguientes distribuciones simétricas alrededor del parámetro μ

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

b) Cauchy de parámetro μ : $f(x) = \frac{1}{\pi(1 + (x - \mu)^2)}$

c) $U(\mu - 1, \mu + 1)$: $f(x) = \frac{1}{2}I_{(\mu-1, \mu+1)}(x)$

La distribución de Cauchy tiene forma de campana como la distribución Normal, pero tiene colas más pesadas que ésta. La distribución Uniforme no tiene colas, por lo tanto podríamos decir que tiene colas más livianas que la Normal.

Consideremos los siguientes estimadores de μ :

$$\hat{\mu}_1 = \bar{X} \quad \hat{\mu}_2 = \tilde{X} \quad \hat{\mu}_3 = \frac{\max(X_i) + \min(X_i)}{2}$$

En el caso a), $\hat{\mu}_1$ es IMVU y por lo tanto, es la elección correcta.

En el caso b), $\hat{\mu}_1$ y $\hat{\mu}_3$ son malos porque ambos son muy sensibles a la presencia de observaciones atípicas y la distribución Cauchy produce una importante proporción de ellas. Por lo tanto la mejor elección entre estos tres estimadores sería $\hat{\mu}_2$. También podríamos utilizar una media podada.

En el caso c) el mejor estimador es $\hat{\mu}_3$ porque la distribución no tiene colas.

Error standard de un estimador: Al informar el resultado de una estimación puntual es necesario brindar información sobre la precisión de la estimación.

Definición: El error standard de un estimador $\hat{\theta}$ es su desviación standard, es decir

$$\sigma_{\hat{\theta}} = \sqrt{V_{\theta}(\hat{\theta})}$$

Si el error standard depende de parámetros desconocidos, éstos se reemplazan por un estimador y se obtiene el error standard estimado.

Ejemplo: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$. Entonces \bar{X} es el EMV de μ y su error standard es

$$\sigma_{\bar{X}} = \sqrt{V_{\mu, \sigma^2}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}}$$

Como depende del parámetro σ , podemos reemplazarlo por la varianza muestral y obtenemos el error standard estimado

$$\hat{\sigma}_{\bar{X}} = \sqrt{\frac{S_X^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n(n-1)}}$$

Definición: Sea $\hat{\theta}$ un estimador de θ , su error cuadrático medio es

$$ECM_{\theta}(\hat{\theta}) = E_{\theta} \left[(\hat{\theta} - \theta)^2 \right]$$

Si el estimador $\hat{\theta}$ es insesgado el error cuadrático medio es igual a la varianza del estimador.

Proposición: $ECM_{\theta}(\hat{\theta}) = V_{\theta}(\hat{\theta}) + [b(\hat{\theta})]^2$, siendo $b(\hat{\theta}) = E_{\theta}(\hat{\theta}) - \theta$ el sesgo del estimador.

Dem:

$$\begin{aligned} ECM_{\theta}(\hat{\theta}) &= E_{\theta} \left[(\hat{\theta} - \theta)^2 \right] = E_{\theta} \left[(\hat{\theta} - E_{\theta}(\hat{\theta}) + E_{\theta}(\hat{\theta}) - \theta)^2 \right] \\ &= E_{\theta} \left[(\hat{\theta} - E_{\theta}(\hat{\theta}))^2 + (E_{\theta}(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - E_{\theta}(\hat{\theta}))(E_{\theta}(\hat{\theta}) - \theta) \right] \\ &= E_{\theta} \left[(\hat{\theta} - E_{\theta}(\hat{\theta}))^2 \right] + E_{\theta} \left[(E_{\theta}(\hat{\theta}) - \theta)^2 \right] + 2E_{\theta} \left[(\hat{\theta} - E_{\theta}(\hat{\theta}))(E_{\theta}(\hat{\theta}) - \theta) \right] \end{aligned}$$

Usando que la esperanza de una v.a. es una constante y la esperanza de una constante es igual a ésta, se obtiene

$$ECM_{\theta}(\hat{\theta}) = \underbrace{E_{\theta} \left[(\hat{\theta} - E_{\theta}(\hat{\theta}))^2 \right]}_{V_{\theta}(\hat{\theta})} + \underbrace{(E_{\theta}(\hat{\theta}) - \theta)^2}_{(b(\hat{\theta}))^2} + \underbrace{2(E_{\theta}(\hat{\theta}) - \theta)(E_{\theta}(\hat{\theta}) - E_{\theta}(\hat{\theta}))}_0$$

y, por lo tanto, $ECM_{\theta}(\hat{\theta}) = V_{\theta}(\hat{\theta}) + [b(\hat{\theta})]^2$, como queríamos demostrar.

Principio de estimación de menor error cuadrático medio: Dados dos o más estimadores del parámetro θ , elegir el de menor ECM.

Este principio se reduce, en el caso de estimadores insesgados, al de mínima varianza entre los insesgados mencionado más arriba, ya que el error cuadrático medio se reduce a la varianza cuando un estimador es insesgado. Sin embargo, nos permite además seleccionar, por ejemplo, entre un estimador insesgado y otro que no lo es, en base a la varianza y al sesgo. Si el estimador sesgado tiene una varianza mucho menor que el insesgado, podría ser preferible su uso.

Definición: Sea X_1, X_2, \dots, X_n una m.a de una distribución que depende de un parámetro θ y sea $\hat{\theta}_n$ un estimador puntual de θ basado en esa muestra. Diremos que $\{\hat{\theta}_n\}$ es una sucesión **consistente** (o más brevemente que $\hat{\theta}_n$ es un estimador consistente de θ) si

$$\hat{\theta}_n \xrightarrow{p} \theta$$

es decir, si $\forall \varepsilon > 0, P\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \xrightarrow{n \rightarrow \infty} 0$.

Ejemplo: Sea X_1, X_2, \dots, X_n una m.a de una distribución con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 < \infty$, entonces \bar{X} es un estimador consistente de μ . En efecto, aplicando la desigualdad de Chebyshev,

$$P\left(|\bar{X} - \mu| > \varepsilon\right) \leq \frac{V(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0$$

Ejercicio: Verificar que, en este ejemplo, $\hat{\mu} = \frac{X_1 + X_2}{2}$ no es consistente de μ .

Proposición: Sea X_1, X_2, \dots, X_n una m.a de una distribución que depende de un parámetro θ y sea $\hat{\theta}_n$ un estimador de θ basado en la muestra de tamaño n . Si

a) $E_\theta(\hat{\theta}_n) \xrightarrow{n \rightarrow \infty} \theta$ (o sea, si el estimador es asintóticamente insesgado)

b) $V_\theta(\hat{\theta}_n) \xrightarrow{n \rightarrow \infty} 0$

entonces, $\hat{\theta}_n$ es consistente de θ .

Dem: Si el estimador es insesgado, la demostración es inmediata, a partir de la desigualdad de Chebyshev,. No daremos la demostración en el caso general.

Ejemplos: 1) Sea X_1, X_2, \dots, X_n una m.a de una distribución con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 < \infty$, entonces \bar{X} es un estimador consistente de μ . En efecto, $E(\bar{X}) = \mu$ y $V(\bar{X}) = \frac{\sigma^2}{n}$. Por lo tanto, se satisfacen las dos condiciones de la Proposición.

2) Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(0, \theta)$. Hemos demostrado antes que el EMV de θ , $\hat{\theta} = \max_{1 \leq i \leq n} (X_i)$ es asintóticamente insesgado pues $E_\theta(\hat{\theta}) = \frac{n}{n+1}\theta$. Para probar que es consistente, verificaremos que su varianza tiende a cero cuando el tamaño de la muestra tiende a infinito. Pero

$$V_\theta(\hat{\theta}) = E_\theta(\hat{\theta}^2) - [E_\theta(\hat{\theta})]^2 = E_\theta(\hat{\theta}^2) - \left[\frac{n}{n+1}\theta \right]^2$$

entonces, debemos calcular la esperanza del cuadrado de la v.a. $U = \max_{1 \leq i \leq n} (X_i)$. Recordando que su densidad está dada por

$$f_U(u) = n \left(\frac{u}{\theta} \right)^{n-1} \frac{1}{\theta} I_{(0, \theta)}(u)$$

$$E_\theta(U^2) = \int_0^\theta u^2 n \left(\frac{u}{\theta} \right)^{n-1} \frac{1}{\theta} du = \frac{n}{\theta^n} \int_0^\theta u^{n+1} du = \frac{n}{\theta^n} \frac{u^{n+2}}{n+2} \Big|_0^\theta = \frac{n}{n+2} \theta^2.$$

Entonces,

$$V_\theta(\hat{\theta}) = \frac{n}{n+2} \theta^2 - \left(\frac{n}{n+1} \right)^2 \theta^2 = \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) \theta^2 = \frac{n}{(n+2)(n+1)^2} \theta^2 \xrightarrow{n \rightarrow \infty} 0$$

Por lo tanto, el EMV es consistente.

3) El último ejemplo que veremos ilustra como demostrar la consistencia de un estimador a partir de la Ley de los Grandes Números y de las propiedades de la convergencia en probabilidad.

En primer lugar recordemos que si $X_1, X_2, \dots, X_n, \dots$ e $Y_1, Y_2, \dots, Y_n, \dots$ son sucesiones de v.a. tales que $X_n \xrightarrow{p} a$ e $Y_n \xrightarrow{p} b$, entonces:

a) $X_n \pm Y_n \xrightarrow{p} a \pm b$

b) $X_n Y_n \xrightarrow{p} ab$

- c) $\frac{X_n}{Y_n} \xrightarrow{p} \frac{a}{b}$ si $b \neq 0$
- d) $g(X_n) \xrightarrow{p} g(a)$ si g es una función continua en a .
- e) si c_n es una sucesión numérica tal que $c_n \rightarrow c$, entonces $c_n X_n \xrightarrow{p} ca$

Sea X_1, X_2, \dots, X_n una m.a de una distribución con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 < \infty$, demostraremos que la varianza muestral S_X^2 es un estimador consistente de la varianza poblacional.

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

Por la Ley de los Grandes Números $\bar{X} \xrightarrow{p} \mu$, entonces por la propiedad d) $\bar{X}^2 \xrightarrow{p} \mu^2$.

Por otra parte, aplicando nuevamente la Ley de los Grandes Números

$$\frac{\sum_{i=1}^n X_i^2}{n} \xrightarrow{p} E_{\mu, \sigma^2}(X^2) = V_{\mu, \sigma^2}(X) + [E_{\mu, \sigma^2}(X)]^2 = \sigma^2 + \mu^2$$

Como además $\frac{n}{n-1} \rightarrow 1$, se obtiene

$$S_X^2 = \frac{n}{n-1} \left(\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \right) \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

y por lo tanto la varianza muestral es un estimador consistente de σ^2 .

Inferencia estadística – Intervalos de confianza

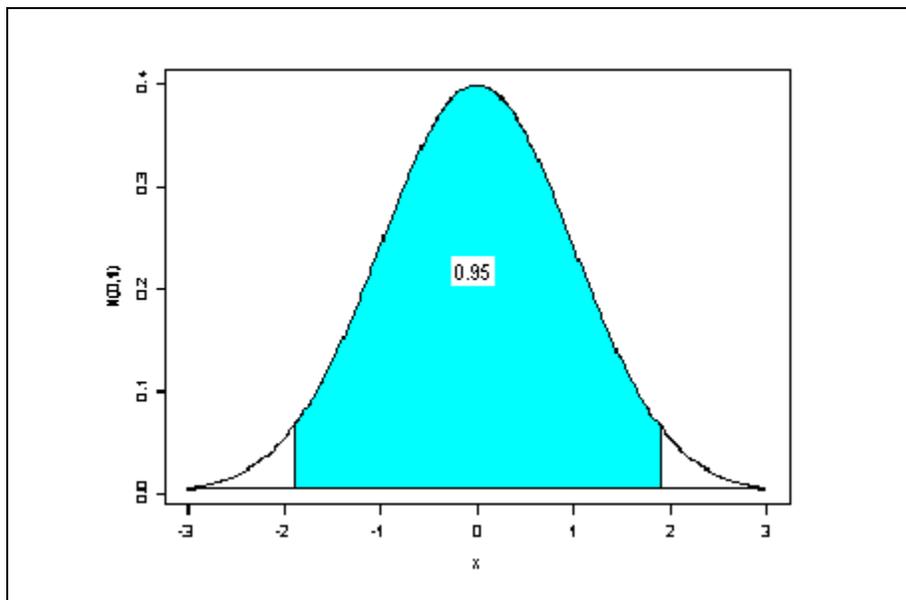
Cuando se obtiene una estimación puntual de un parámetro, es conveniente acompañar dicha estimación por una “medida” de la precisión de la estimación. Un modo de hacerlo es informar el estimador y su error standard. Otro modo es reemplazar la estimación puntual por un intervalo de valores posibles para el parámetro.

Ejemplo: Supongamos que tenemos una m.a. X_1, X_2, \dots, X_n de una distribución $N(\mu, \sigma_o^2)$ con varianza σ_o^2 conocida. Por ser los datos normales, sabemos que

$$\bar{X} \sim N\left(\mu, \frac{\sigma_o^2}{n}\right) \Leftrightarrow \frac{\bar{X} - \mu}{\frac{\sigma_o}{\sqrt{n}}} \sim N(0,1)$$

y, por lo tanto, sabemos que la probabilidad de que $\sqrt{n} \frac{\bar{X} - \mu}{\sigma_o}$ se encuentre entre -1.96 y 1.96 es 0.95 , es decir

$$P\left(-1.96 \leq \sqrt{n} \frac{\bar{X} - \mu}{\sigma_o} \leq 1.96\right) = 0.95$$



A partir de esta expresión obtenemos

$$P\left(-1.96 \frac{\sigma_o}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma_o}{\sqrt{n}}\right) = 0.95 \quad \Leftrightarrow \quad P\left(\bar{X} - 1.96 \frac{\sigma_o}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma_o}{\sqrt{n}}\right) = 0.95$$

Es decir, que la probabilidad de que el intervalo

$$\left[\bar{X} - 1.96 \frac{\sigma_o}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma_o}{\sqrt{n}} \right]$$

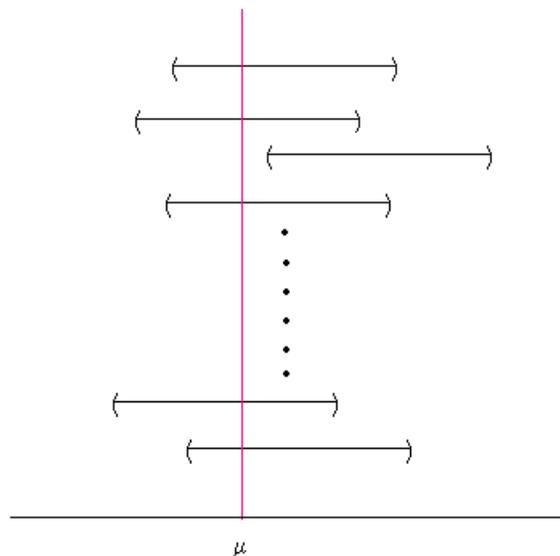
contenga al verdadero valor del parámetro μ es 0.95. Este intervalo se denomina intervalo de confianza para μ de nivel de confianza 0.95.

Definición: Sea X_1, X_2, \dots, X_n una m.a. de una distribución que depende de un parámetro θ . Dadas dos funciones de la muestra $a(X_1, X_2, \dots, X_n)$ y $b(X_1, X_2, \dots, X_n)$ tales que

$$P(a(X_1, X_2, \dots, X_n) \leq \theta \leq b(X_1, X_2, \dots, X_n)) = 1 - \alpha$$

con α pequeño (por ejemplo, 0.10, 0.05, 0.01), el intervalo $[a(X_1, X_2, \dots, X_n), b(X_1, X_2, \dots, X_n)]$ se denomina **intervalo de confianza de nivel $1 - \alpha$** para el parámetro θ .

Interpretación: Supongamos que, en base a diferentes muestras calculamos los correspondientes intervalos de confianza para θ . Entonces el $(1 - \alpha)$ 100% de ellos contendrán al verdadero valor θ .



Observaciones: 1) No es correcto decir “la probabilidad de que θ pertenezca al intervalo (a,b) es $1 - \alpha$ ” porque θ no es una variable aleatoria. El intervalo es aleatorio ya que sus extremos son funciones de la muestra y por lo tanto, debemos decir “la probabilidad de que el intervalo (a,b) contenga al parámetro θ es $1 - \alpha$ ”

2) Una vez construido el intervalo a partir de una muestra dada, ya no tiene sentido hablar de probabilidad. En todo caso, tenemos “confianza” de que el intervalo contenga a θ . La confianza está puesta en el método de construcción de los intervalos, que nos asegura que $(1 - \alpha)$ 100% de las muestras producirán intervalos que contienen a θ .

Intervalos de confianza para los parámetros de una distribución normal

Distribución t: Sean dos v.a. $Z \sim N(0,1)$ y $U \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ independientes, entonces

$$T = \frac{Z}{\sqrt{U/n}} \sim t_n$$

Se dice que T tiene distribución t de Student con n grados de libertad. Esta distribución está tabulada para diferentes valores de n . Su densidad es simétrica respecto al 0 y tiene forma de campana, pero tiene colas más pesadas que la distribución normal standard. Cuando n tiende a infinito, la distribución de Student tiende a la distribución normal standard.

Proposición: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$, entonces

$$\text{a) } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Leftrightarrow \quad \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

$$\text{b) } \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \text{con } S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

c) \bar{X} y S^2 son independientes

$$\text{d) } \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

Dem: a) Ya hemos visto que cualquier combinación de v.a. normales independientes es normal y el promedio es una combinación lineal particular.

b) y c) Están fuera del alcance de este curso.

d) Resulta de a) b) y c) pues

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0,1) \quad \text{y} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

son v.a. independientes. Entonces, por definición de la distribución t de Student,

$$\frac{\sqrt{n} \frac{\bar{X} - \mu}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

Intervalo de confianza para la media de la distribución normal con varianza conocida: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma_o^2)$, con varianza σ_o^2 conocida, entonces

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma_o} \sim N(0,1)$$

$$P\left(-z_{\alpha/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{\sigma_o} \leq z_{\alpha/2}\right) = 1 - \alpha$$

de donde se deduce el siguiente intervalo de confianza de nivel $1 - \alpha$ para μ ,

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}} \right] \quad (1)$$

Intervalo de confianza para la media de la distribución normal con varianza desconocida: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$, entonces

$$\sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

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

de donde se deduce el siguiente intervalo de confianza de nivel $1 - \alpha$ para μ ,

$$\left[\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right]$$

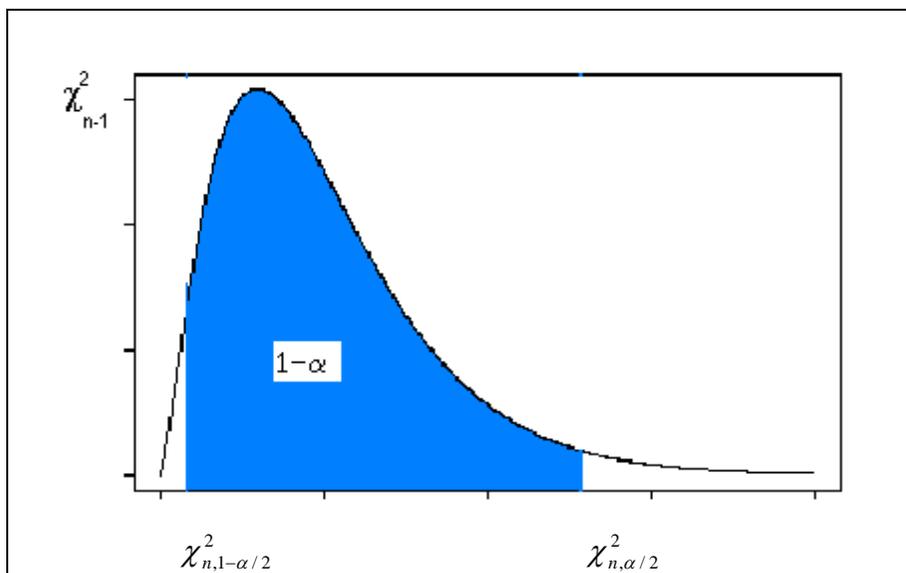
Intervalo de confianza para la varianza de la distribución normal con media conocida: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu_0, \sigma^2)$, con media μ_0 conocida, entonces

$$\frac{X_i - \mu_0}{\sigma} \sim N(0,1) \quad \forall 1 \leq i \leq n \quad \Rightarrow \quad \left(\frac{X_i - \mu_0}{\sigma} \right)^2 \sim \chi_1^2 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) \quad \forall 1 \leq i \leq n$$

Como además las v.a. son independientes

$$\sum_{i=1}^n \left(\frac{X_i - \mu_0}{\sigma} \right)^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

¿Cómo elegimos los percentiles de la distribución χ^2 que encierran un área igual a $1 - \alpha$?



Los elegimos de manera tal que quede un área igual a $\alpha/2$ en cada extremo. Entonces,

$$P\left(\chi_{n,1-\alpha/2}^2 \leq \frac{\sum_{i=1}^n (X_i - \mu_o)^2}{\sigma^2} \leq \chi_{n,\alpha/2}^2\right) = 1 - \alpha$$

Se obtiene el siguiente intervalo

$$\left[\frac{\sum_{i=1}^n (X_i - \mu_o)^2}{\chi_{n,\alpha/2}^2}, \frac{\sum_{i=1}^n (X_i - \mu_o)^2}{\chi_{n,1-\alpha/2}^2} \right]$$

Intervalo de confianza para la varianza de la distribución normal con media desconocida: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$, entonces

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Por lo tanto,

$$P\left(\chi_{n-1,1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}^2\right) = 1 - \alpha$$

Se obtiene el siguiente intervalo

$$\left[\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} \right]$$

Ejemplos: Sea X_1, X_2, \dots, X_{49} una m.a., $X_i \sim N(\mu, \sigma^2)$.

- a) Supongamos que el verdadero valor del desvío standard es $\sigma_o = 35$ y que se observa $\bar{x} = 160$. Construyamos un intervalo de confianza para la media de nivel 0.95.

Como las v.a. son normales y la varianza es conocida, el intervalo para μ será de la forma

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}} \right)$$

con $z_{\alpha/2} = z_{0,025} = 1.96$, $\sigma_o = 35$, $n = 49$ y valor observado de \bar{X} igual a 160. Obtenemos

$$\left(160 - 1.96 \frac{35}{\sqrt{49}}, 160 + 1.96 \frac{35}{\sqrt{49}} \right) = (160 - 9.8, 160 + 9.8) = (150.2, 169.8)$$

b) Supongamos ahora que la varianza es desconocida pero que el valor observado de S es $s=35$. El correspondiente intervalo de confianza para μ será de la forma

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

con $t_{n-1,\alpha/2} = t_{48,0,025} = 2.01$. Obtenemos

$$\left(160 - 2.01 \frac{35}{\sqrt{49}}, 160 + 2.01 \frac{35}{\sqrt{49}} \right) = (160 - 10.05, 160 + 10.05) = (149.95, 170.05)$$

Notemos que es más ancho que el anterior

c) Suponiendo como antes que observamos $\bar{x} = 160$ y $s = 35$, hallemos un intervalo de confianza para σ^2 de nivel 0.95.

Por tratarse de una muestra normal con media desconocida, el intervalo para σ^2 será de la forma

$$\left(\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} \right)$$

con $\chi_{n-1,\alpha/2}^2 = \chi_{48,0,025}^2 = 69.02$ y $\chi_{n-1,1-\alpha/2}^2 = \chi_{48,0,975}^2 = 30.75$. Obtenemos

$$\left(\frac{48 \cdot 35^2}{69.02}, \frac{48 \cdot 35^2}{30.75} \right) = (851.93, 1912.20)$$

y un intervalo de confianza para σ de nivel 0.95 será

$$\left(\sqrt{\frac{48 \cdot 35^2}{69.02}}, \sqrt{\frac{48 \cdot 35^2}{30.75}} \right) = (\sqrt{851.93}, \sqrt{1912.20}) = (29.19, 43.73)$$

Esto último resulta de aplicar una función monótona creciente a cada extremo del intervalo para σ^2

Determinación del tamaño de muestra: Consideremos el intervalo de confianza para μ con varianza conocida en el caso de una m.a. normal. La longitud del intervalo obtenido (1) es

$$L = 2z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}}$$

y depende de

- nivel de confianza (α)
- varianza o desvío standard de las observaciones (σ_o)
- tamaño de la muestra (n)

Un modo de obtener mayor precisión, es decir un intervalo más angosto, es aumentando el tamaño de la muestra. Si se desea una longitud menor o igual que L_o , entonces

$$L = 2z_{\alpha/2} \frac{\sigma_o}{\sqrt{n}} \leq L_o \Leftrightarrow \sqrt{n} \geq \frac{2z_{\alpha/2}\sigma_o}{L_o} \Leftrightarrow n \geq \left(\frac{2z_{\alpha/2}\sigma_o}{L_o} \right)^2$$

Ejemplo: Supongamos que $\sigma_o = 35$, ¿qué tamaño de muestra se requiere como mínimo para obtener un intervalo de nivel 0.95 de longitud menor o igual que 10?.

En este caso, $L_o = 10$, $\sigma_o = 35$ y $z_{0.025} = 1.96$, entonces

$$n \geq \left(\frac{2 \cdot 1.96 \cdot 35}{10} \right)^2 = 188.23 \Rightarrow n \geq 189$$

En el caso de varianza desconocida el problema es más complejo porque el percentil t también depende del tamaño de muestra.

Método general para obtener intervalos de confianza:

Sea X_1, X_2, \dots, X_n una m.a. de una distribución que depende de un parámetro θ . Supongamos que existe una función $T(X_1, X_2, \dots, X_n, \theta)$ (es decir, una función de la muestra y del parámetro) cuya distribución no depende de θ ni de ningún otro parámetro desconocido. Entonces, existen dos valores a y b tales que

$$P(a \leq T(X_1, X_2, \dots, X_n, \theta) \leq b) = 1 - \alpha$$

y, a partir de esta expresión, es posible obtener un intervalo de confianza para θ .

La función $T(X_1, X_2, \dots, X_n, \theta)$ se denomina pivote.

Ejemplo: Sea X_1, X_2, \dots, X_n una m.a. de una distribución exponencial de parámetro λ . Hemos demostrado que

$$\sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$$

Usando este resultado y que, si $V \sim \Gamma(\alpha, \lambda)$ y $a > 0$ entonces $aV \sim \Gamma\left(\alpha, \frac{\lambda}{a}\right)$, se puede demostrar que

$$2\lambda \sum_{i=1}^n X_i \sim \chi_{2n}^2 = \Gamma\left(\frac{2n}{2}, \frac{1}{2}\right)$$

Usando como pivote la función $T(X_1, X_2, \dots, X_n, \lambda) = 2\lambda \sum_{i=1}^n X_i$, podemos obtener un intervalo de confianza de nivel $1 - \alpha$ para el parámetro λ .

$$P\left(\chi_{2n, 1-\alpha/2}^2 \leq 2\lambda \sum_{i=1}^n X_i \leq \chi_{2n, \alpha/2}^2\right) = 1 - \alpha$$

$$P\left(\frac{\chi_{2n, 1-\alpha/2}^2}{2 \sum_{i=1}^n X_i} \leq \lambda \leq \frac{\chi_{2n, \alpha/2}^2}{2 \sum_{i=1}^n X_i}\right) = 1 - \alpha$$

y el intervalo requerido es

$$\left[\frac{\chi_{2n, 1-\alpha/2}^2}{2 \sum_{i=1}^n X_i}, \frac{\chi_{2n, \alpha/2}^2}{2 \sum_{i=1}^n X_i} \right]$$

Ejemplo: Sea X_1, X_2, \dots, X_n una m.a. de una distribución $U(0, \theta)$. Para obtener un intervalo de confianza para θ , recordemos que el EMV de θ es $\hat{\theta} = \max(X_1, \dots, X_n)$ y probemos que la distribución de $\hat{\theta}/\theta$ no depende de θ .

Llamemos V a la v.a. $\max(X_1, \dots, X_n)$. Recordemos que, si X_1, X_2, \dots, X_n es una m.a. de una distribución F_X , entonces la función de distribución de V está dada por

$$F_V(v) = (F_{X_i}(v))^n$$

Queremos demostrar que la distribución de V/θ . no depende de θ .

$$F_{V/\theta}(w) = P\left(\frac{V}{\theta} \leq w\right) = P(V \leq \theta w) = F_V(\theta w) = (F_{X_i}(\theta w))^n$$

Como, en nuestro caso, $X_i \sim U(0, \theta)$,

$$F_{V/\theta}(w) = (F_{X_i}(\theta w))^n = \begin{cases} 0 & \text{si } \theta w \leq 0 \\ \left(\frac{\theta w}{\theta}\right)^n & \text{si } 0 < \theta w < \theta \\ 1 & \text{si } \theta w \geq \theta \end{cases} = \begin{cases} 0 & \text{si } w \leq 0 \\ w^n & \text{si } 0 < w < 1 \\ 1 & \text{si } w \geq 1 \end{cases}$$

Por lo tanto, la distribución de V/θ no depende de θ . Derivando, se obtiene la densidad de V/θ

$$f_{V/\theta}(w) = n w^{n-1} I_{(0,1)}(w)$$

Utilizando $T(X_1, X_2, \dots, X_n, \theta) = \frac{\max(X_1, \dots, X_n)}{\theta}$ como pivote, obtendremos un intervalo de confianza para θ de nivel $1 - \alpha$. Buscamos a y b tales que

$$P\left(a \leq \frac{\max(X_1, \dots, X_n)}{\theta} \leq b\right) = 1 - \alpha \quad (2)$$

y, obtenemos el siguiente intervalo

$$\left[\frac{\max(X_1, \dots, X_n)}{b}, \frac{\max(X_1, \dots, X_n)}{a} \right]$$

¿Cómo elegimos a y b ? Observando (2), debemos hallar a y b , $0 < a < b < 1$, tales que

$$\int_a^b n w^{n-1} dw = w^n \Big|_a^b = b^n - a^n = 1 - \alpha \quad (3)$$

Obviamente hay infinitas soluciones de esta ecuación, pero podríamos elegir la solución que produce el intervalo de menor longitud esperada, es decir, buscar a y b que minimicen $E(L)$ sujeto a la condición (3), siendo

$$L = \max(X_1, \dots, X_n) \left(\frac{1}{a} - \frac{1}{b} \right)$$

Como ya hemos demostrado que $E(\max(X_1, \dots, X_n)) = \frac{n}{n+1} \theta$, debemos minimizar

$$\frac{n}{n+1} \theta \left(\frac{1}{a} - \frac{1}{b} \right) \quad (4)$$

sujeto a la condición $b^n - a^n = 1 - \alpha$.

Esto puede hacerse utilizando multiplicadores de Lagrange o bien, despejando de esta última expresión a en función de b , reemplazándola en (4) y minimizando la expresión resultante respecto de a .

El intervalo de mínima longitud esperada es

$$\left(\frac{\max(X_1, \dots, X_n)}{1}, \frac{\max(X_1, \dots, X_n)}{\sqrt[n]{\alpha}} \right)$$

Intervalos de confianza de nivel asintótico $1 - \alpha$:

En muchos problemas no es posible encontrar intervalos de confianza de nivel exacto $1 - \alpha$, o bien son de muy difícil construcción. En otros casos disponemos de muy poca información sobre la distribución de las variables aleatorias. En estos dos tipos de situaciones es posible obtener intervalos de confianza de nivel aproximado cuando el tamaño de la muestra es grande.

Definición: Sea X_1, X_2, \dots, X_n una m.a. de una distribución que depende de un parámetro θ . Dadas dos sucesiones $\{a_n(X_1, X_2, \dots, X_n)\}$ y $\{b_n(X_1, X_2, \dots, X_n)\}$ tales que

$$\lim_{n \rightarrow \infty} P(a_n(X_1, X_2, \dots, X_n) \leq \theta \leq b_n(X_1, X_2, \dots, X_n)) = 1 - \alpha$$

la sucesión de intervalos $[a_n(X_1, X_2, \dots, X_n), b_n(X_1, X_2, \dots, X_n)]$ es una sucesión de **intervalos de confianza de nivel asintótico $1 - \alpha$** para el parámetro θ . También se dice que, si n es suficientemente grande, el intervalo $[a_n(X_1, X_2, \dots, X_n), b_n(X_1, X_2, \dots, X_n)]$ tiene nivel aproximado $1 - \alpha$.

¿Porqué calcular intervalos de nivel asintótico?

- Porque no es posible encontrar una función pivote que no dependa del parámetro
- Porque no se conoce la distribución exacta de la función pivote
- Porque en general es más fácil encontrar la distribución asintótica que la exacta de la función pivote

Ejemplos: 1) Sea X_1, X_2, \dots, X_n una m.a. de una distribución F con $E(X_i) = \mu$ y $V(X_i) = \sigma^2 < \infty$. Buscamos un intervalo de confianza para μ .

Sabemos que \bar{X} es un estimador insesgado y consistente de μ . No conocemos su distribución exacta porque no conocemos la de X_i , pero sabemos que

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

Si σ^2 es conocido, esta función podría servir de pivote para el intervalo de nivel aproximado, pero ¿qué usamos si σ^2 es desconocido?.

Propiedad:

$$\left. \begin{array}{l} Y_n \xrightarrow{d} Y \\ U_n \xrightarrow{p} a \end{array} \right\} \Rightarrow U_n Y_n \xrightarrow{d} aY$$

Como $s \xrightarrow{p} \sigma$ por ser un estimador consistente, entonces $\frac{s}{\sigma} \xrightarrow{p} 1$ y $\frac{\sigma}{s} \xrightarrow{p} 1$.

Luego,

$$\left. \begin{array}{l} \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1) \\ \frac{\sigma}{s} \xrightarrow{p} 1 \end{array} \right\} \Rightarrow \sqrt{n} \frac{\bar{X} - \mu}{s} \xrightarrow{d} N(0,1)$$

A partir de este resultado,

$$P\left(-z_{\alpha/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{s} \leq z_{\alpha/2}\right) \rightarrow 1 - \alpha$$

y se obtiene el siguiente intervalo de nivel aproximado $1 - \alpha$

$$\left[\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Intervalo de confianza de nivel asintótico $1 - \alpha$ para el parámetro p de la distribución Binomial:

Sea X_1, X_2, \dots, X_n una m.a. de una distribución $Bi(1, p)$. Entonces

$X = \sum_{i=1}^n X_i \sim Bi(n, p)$. Queremos construir un intervalo de nivel asintótico $1 - \alpha$ para p .

Recordemos que, por el TCL,

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^n X_i}{n} \stackrel{(a)}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

y, por lo tanto

$$P\left(-z_{\alpha/2} \leq \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) \cong 1 - \alpha \quad (5)$$

Hay dos formas de obtener un intervalo para p a partir de esta última expresión.

- a) Como, por la Ley de los Grandes Números, $\frac{X}{n} = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{p} p$ podemos aplicar la Propiedad enunciada antes y reemplazar en el denominador el pivote p por su estimador. Entonces

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

$$P\left(\frac{X}{n} - z_{\alpha/2} \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}} \leq p \leq \frac{X}{n} + z_{\alpha/2} \sqrt{\frac{\frac{X}{n} \left(1 - \frac{X}{n}\right)}{n}}\right) \cong 1 - \alpha$$

obteniendo un intervalo para p de nivel aproximado $1 - \alpha$.

b) Reescribimos la expresión (5) en la forma

$$P\left(\left|\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq z_{\alpha/2}\right) \cong 1 - \alpha \quad \Leftrightarrow \quad P\left(\frac{\left(\frac{X}{n} - p\right)^2}{\frac{p(1-p)}{n}} \leq z_{\alpha/2}^2\right) \cong 1 - \alpha$$

Observemos que

$$\begin{aligned} \frac{\left(\frac{X}{n} - p\right)^2}{\frac{p(1-p)}{n}} \leq z_{\alpha/2}^2 &\Leftrightarrow \left(\frac{X}{n} - p\right)^2 \leq z_{\alpha/2}^2 \frac{p(1-p)}{n} \\ &\Leftrightarrow \left(\frac{X}{n}\right)^2 - 2p\frac{X}{n} + p^2 - z_{\alpha/2}^2 \frac{p(1-p)}{n} \leq 0 \\ &\Leftrightarrow p^2\left(1 + \frac{z_{\alpha/2}^2}{n}\right) - p\left(\frac{2X}{n} + \frac{z_{\alpha/2}^2}{n}\right) + \left(\frac{X}{n}\right)^2 \leq 0 \end{aligned}$$

Buscamos las raíces de esta ecuación de segundo grado, que llamaremos \hat{p}_1 y \hat{p}_2 y el intervalo de nivel aproximado $1 - \alpha$ para p será

$$[\hat{p}_1, \hat{p}_2].$$

Inferencia estadística – Tests de hipótesis

Hasta ahora hemos visto como obtener, a partir de una muestra, un estimador puntual o un intervalo de confianza para un parámetro θ . Frecuentemente el objetivo del estudio es decidir, en base a la información que provee la muestra, entre dos hipótesis relativas a un parámetro.

Ejemplo: Supongamos que el consumo promedio de nafta de los motores utilizados por una empresa automotriz en uno de sus modelos es de 10 litros cada 100 km. Se presenta un proyecto de mejora del motor que produciría una disminución en el consumo pero, por razones de costo, se considera viable el proyecto si la reducción lleva el consumo a un valor menor de 9 litros cada 100 km.

Para estudiar la conveniencia o no de aplicar la mejora a los motores, se aplica esta mejora a una muestra de 25 motores, los cuáles se ponen a funcionar en igualdad de condiciones durante un periodo fijo. El consumo promedio observado es de 8.9 litros cada 100 km. ¿Proveen estos datos evidencia de que vale la pena incorporar la mejora al motor o se deben simplemente al azar?

Supongamos que el consumo de nafta de los motores es una v.a. con distribución normal con varianza igual a 1 y que la muestra es aleatoria, es decir que los 25 consumos son independientes. Es decir, supongamos que X_1, \dots, X_{25} es una m.a., $X_i \sim N(\mu, 1)$. Entonces

$$\bar{X} \sim N\left(\mu, \frac{1}{25}\right) \Leftrightarrow \frac{X - \mu}{\sqrt{1/25}} \sim N(0, 1)$$

Si la media verdadera del consumo en el motor mejorado fuese de 9 litros cada 100 km., ¿cuál es la probabilidad de que una v.a. normal con media 9 y varianza 1/25 tome un valor igual o menor que el observado, 8.9?

$$P(\bar{X} \leq 8.9) = P\left(\frac{\bar{X} - 9}{1/5} \leq \frac{8.9 - 9}{1/5}\right) = \Phi(-0.5) = 0.309 \cong 0.31$$

Esta probabilidad se denomina **p-valor**.

Si el consumo promedio observado hubiese sido $\bar{X} = 8.6$ litros cada 100 km, entonces

$$P(\bar{X} \leq 8.6) = P\left(\frac{\bar{X} - 9}{1/5} \leq \frac{8.6 - 9}{1/5}\right) = \Phi(-2) = 0.023,$$

es decir que, en este último caso, hubiese sido muy poco probable que se observase un valor promedio de 8.6 si la media verdadera es 9.

¿Qué es lo que estamos tratando de decidir? Nuestras hipótesis se refieren a μ , y se

podrían enunciar así:

- i) $\mu = 9$ litros cada 100 km. En este caso no se implementa la mejora a los motores
- ii) $\mu < 9$ litros cada 100 km. En este caso conviene implementar la mejora a los motores

A la primera hipótesis se la denomina **hipótesis nula** y se designa H_0 . Esta hipótesis implica que no hay efecto, es la hipótesis del status quo, o sea del no cambio respecto a la situación inicial. La segunda hipótesis se denomina **hipótesis alternativa** y se designa H_1 . Se la suele llamar la hipótesis del investigador.

Expresadas en términos del parámetro de interés las hipótesis del ejemplo serán

$$H_0: \mu = 9 \quad \text{vs} \quad H_1: \mu < 9$$

Un test es una regla de decisión basada en un **estadístico** o función de la muestra, en este caso \bar{X} , y en una **zona de rechazo**, es decir un conjunto de valores para los cuáles se rechaza la hipótesis nula H_0 .

¿Cómo se elige la zona de rechazo? Observemos que al tomar una decisión en base a una muestra, podemos cometer dos tipos de error.

	No se rechaza H_0	Se rechaza H_0
H_0 es cierta	OK	Error tipo I
H_0 no es cierta	Error tipo II	OK

Debido a la variabilidad muestral, es imposible construir tests en los cuáles estemos absolutamente seguros de tomar la decisión correcta,. Lo que podemos hacer es tratar de mantener bajas las probabilidades de error.

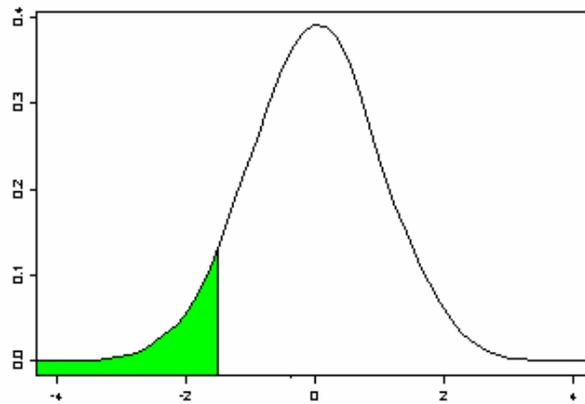
Llamaremos **nivel de significación del test**, y lo designaremos α , a la probabilidad de error tipo I (en realidad a la máxima probabilidad de error tipo I) y designaremos β a la probabilidad de error tipo II.

Como el estadístico se construye bajo la condición de que H_0 es verdadera, lo que podemos controlar es la probabilidad de error tipo I. Elegiremos la zona de rechazo del test de manera que la probabilidad de error tipo I sea un valor α predeterminado.

Volviendo al ejemplo, sabemos que, si H_0 es cierta,

$$\frac{\bar{X} - 9}{1/5} \sim N(0, 1)$$

Si queremos que el test tenga nivel de significación $\alpha = 0.05$, rechazaríamos H_0 si

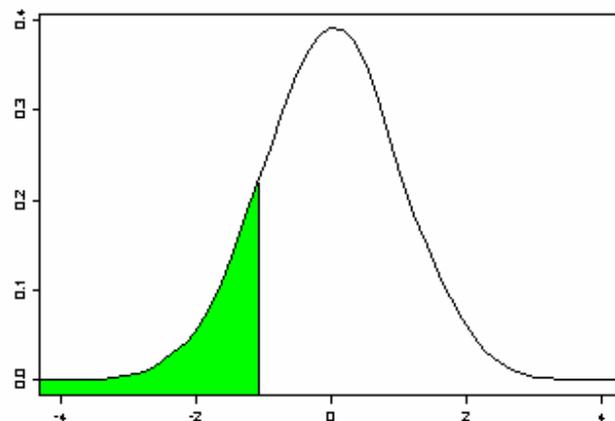


$$\frac{\bar{X} - 9}{1/5} \leq -1.64.$$

Esta es la zona de rechazo del test de nivel 0.05. Si observamos un promedio igual a 8.9, el valor del estadístico es -0.5 y por lo tanto no se rechaza H_0 , mientras que si observamos un promedio igual a 8.6, el valor del estadístico es -2 y se rechaza H_0 .

Si queremos que el test tenga nivel de significación $\alpha = 0.10$, rechazaríamos H_0 si

$$\frac{\bar{X} - 9}{1/5} \leq -1.28$$



Esta es la zona de rechazo del test de nivel 0.10.

Como hemos visto, al seleccionar la región de rechazo controlamos la probabilidad de error tipo I, pero ¿qué ocurre con el error tipo II?

Supongamos que en nuestro ejemplo, observamos un consumo promedio en la muestra de tamaño 25 igual a 8.9 litros cada 100 km y trabajamos con el test de nivel 0.05. En este caso, no rechazamos H_0 (tampoco lo haríamos con el test de nivel 0.10) y por lo tanto, si la mejora en el motor fuese real, podríamos estar cometiendo un error de tipo II.

Por ejemplo, si la modificación en el motor reduce el consumo a 8.5 litros cada 100 km, ¿cuál es la probabilidad de cometer un error tipo II?

$$P_{\mu=8.5} \left(\frac{\bar{X} - 9}{1/5} > -1.64 \right) = P_{\mu=8.5} \left(\bar{X} > -1.64 \cdot \frac{1}{5} + 9 \right) = P_{\mu=8.5} \left(\frac{\bar{X} - 8.5}{1/5} > \frac{-1.64 \cdot \frac{1}{5} + 9 - 8.5}{1/5} \right)$$

$$P_{\mu=8.5} \left(\frac{\bar{X} - 8.5}{1/5} > 0.86 \right) = 1 - \Phi(0.86) = 1 - 0.805 = 0.195$$

Es decir, que la probabilidad de error tipo II para el valor de $\mu = 8.5$ es aproximadamente 0.20.

Definición: La **función de potencia** de un test, $\pi(\mu)$, es la probabilidad de rechazar la hipótesis nula cuando el valor verdadero del parámetro es μ .

Utilizando la función de potencia es posible obtener una expresión general para los dos tipos de errores, pues

$$\pi(\mu) = \begin{cases} \alpha(\mu) & \text{si } \mu \in H_0 \\ 1 - \beta(\mu) & \text{si } \mu \in H_1 \end{cases}$$

donde $\alpha(\mu)$ y $\beta(\mu)$ denota las probabilidades de error tipo I y tipo II respectivamente cuando el verdadero valor del parámetro es μ .

Tipos de hipótesis a testear:

Hipótesis unilaterales:

$$H_0: \theta = \theta_0 \text{ (ó } \theta \leq \theta_0) \quad \text{vs} \quad H_1: \theta > \theta_0$$

$$H_0: \theta = \theta_0 \text{ (ó } \theta \geq \theta_0) \quad \text{vs} \quad H_1: \theta < \theta_0$$

Hipótesis bilaterales:

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta \neq \theta_0$$

La forma de la región de rechazo dependerá de la hipótesis alternativa a testear. Así, en el ejemplo presentado anteriormente, la zona de rechazo consiste en un intervalo de valores en la cola izquierda de la distribución porque la hipótesis alternativa es de la forma $\mu < \mu_0$.

Tests de hipótesis de nivel α para los parámetros de la distribución normal:

Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$.

Tests para la media cuando la varianza es conocida: Supongamos que $\sigma^2 = \sigma_0^2$ es conocida y consideremos las siguientes hipótesis

- a) $H_0: \mu = \mu_0$ (ó $\mu \leq \mu_0$) vs $H_1: \mu > \mu_0$
- b) $H_0: \mu = \mu_0$ (ó $\mu \geq \mu_0$) vs $H_1: \mu < \mu_0$
- c) $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

Estadístico del test: $T = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma_0}$. Bajo $H_0: \mu = \mu_0$, $T \sim N(0,1)$.

Región de rechazo: Como dijimos, la zona de rechazo depende de la hipótesis alternativa. Estará dada, en cada caso, por

- a) $T \geq z_\alpha$
- b) $T \leq -z_\alpha$
- c) $|T| \geq z_{\alpha/2}$

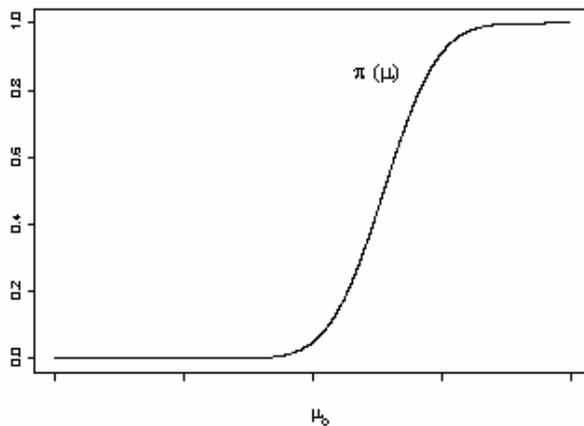
Observemos que, así como la forma de la región de rechazo depende de la alternativa, su tamaño depende del nivel. Por ejemplo, consideremos el caso c). Como la alternativa es $\mu \neq \mu_0$, la forma de la región es $|T| \geq K$, pero como la probabilidad de rechazar H_0 siendo cierta, o sea la probabilidad de Error tipo I, debe ser α ,

$$P_{\mu_0} \left(\left| \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma_0} \right| \geq K \right) = \alpha \Leftrightarrow 1 - P_{\mu_0} \left(\left| \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma_0} \right| < K \right) = 1 - P_{\mu_0} \left(-K < \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma_0} < K \right) = \alpha$$

$$\Leftrightarrow 1 - \Phi(K) + \Phi(-K) = \alpha \Leftrightarrow 2(1 - \Phi(K)) = \alpha \Leftrightarrow \Phi(K) = 1 - \frac{\alpha}{2} \Leftrightarrow K = z_{\alpha/2}$$

Función de potencia: La notación P_μ , como ya hemos visto, indicará la probabilidad cuando el valor verdadero del parámetro es μ . Hallaremos la función de potencia para cada uno de los tests planteados.

$$\begin{aligned} \text{a) } \pi(\mu) &= P_\mu \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \right) = P_\mu \left(\frac{\bar{X} - \mu + \mu - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \right) \\ &= P_\mu \left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \geq z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right) = 1 - \Phi \left(z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right) \end{aligned}$$

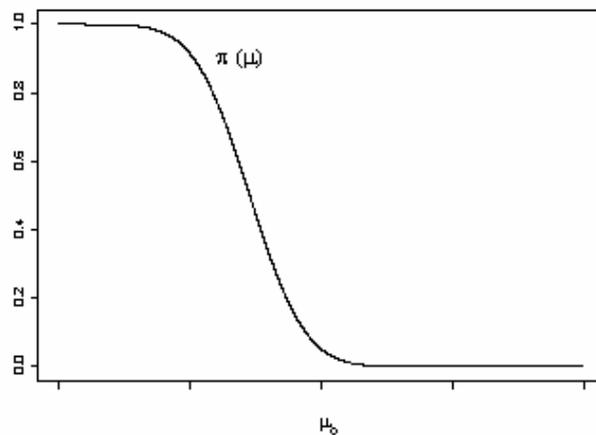


Observemos que esta función es creciente y $\pi(\mu_0) = \alpha$, entonces, si $\mu < \mu_0$, $\pi(\mu) < \alpha$. Por esta razón el test también es de nivel α para las hipótesis

$$H_0: \mu \leq \mu_0 \quad \text{vs} \quad H_1: \mu > \mu_0$$

en el sentido de que la probabilidad de error tipo I es a lo sumo α .

$$\begin{aligned} \text{b) } \pi(\mu) &= P_{\mu} \left(\frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \leq -z_{\alpha} \right) = P_{\mu} \left(\frac{\bar{X} - \mu + \mu - \mu_0}{\sigma_0 / \sqrt{n}} \leq -z_{\alpha} \right) \\ &= P_{\mu} \left(\frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} \leq -z_{\alpha} + \frac{\mu_0 - \mu}{\sigma_0 / \sqrt{n}} \right) = \Phi \left(-z_{\alpha} + \frac{\mu_0 - \mu}{\sigma_0 / \sqrt{n}} \right) \end{aligned}$$



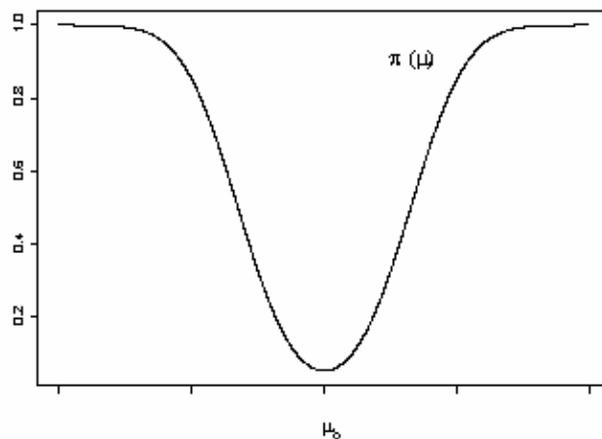
Observemos que esta función es decreciente y $\pi(\mu_0) = \alpha$, entonces, si $\mu > \mu_0$, $\pi(\mu) < \alpha$. Por esta razón el test también es de nivel α para las hipótesis

$$H_0: \mu \geq \mu_0 \quad \text{vs} \quad H_1: \mu > \mu_0$$

en el sentido de que la probabilidad de error tipo I es a lo sumo α .

$$\begin{aligned} \text{c) } \pi(\mu) &= P_{\mu} \left(\left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| \geq z_{\alpha/2} \right) = 1 - P_{\mu} \left(\left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| < z_{\alpha/2} \right) \\ &= 1 - P_{\mu} \left(-z_{\alpha/2} < \frac{\bar{X} - \mu + \mu - \mu_0}{\sigma_0 / \sqrt{n}} < z_{\alpha/2} \right) \end{aligned}$$

$$\begin{aligned}
 &= 1 - P_{\mu} \left(-z_{\alpha/2} + \frac{\mu_o - \mu}{\sigma_o/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma_o/\sqrt{n}} < z_{\alpha/2} + \frac{\mu_o - \mu}{\sigma_o/\sqrt{n}} \right) \\
 &= 1 - \Phi \left(z_{\alpha/2} + \frac{\mu_o - \mu}{\sigma_o/\sqrt{n}} \right) + \Phi \left(-z_{\alpha/2} + \frac{\mu_o - \mu}{\sigma_o/\sqrt{n}} \right)
 \end{aligned}$$



Observemos que esta función decrece hasta μ_o donde $\pi(\mu_o) = \alpha$ y crece a partir de allí.

Tamaño de muestra requerido para obtener una probabilidad de error tipo II dada para un valor $\mu = \mu_1$ (fijo) en la alternativa: Recordemos que el error de tipo II se define como “aceptar la hipótesis nula H_o cuando es falsa”. Buscamos el valor de n para que la probabilidad de error tipo II sea menor que β cuando $\mu = \mu_1$ es un valor fijo en H_1 .

$$\begin{aligned}
 \text{a) } P_{\mu_1} \left(\frac{\bar{X} - \mu_o}{\sigma_o/\sqrt{n}} < z_{\alpha} \right) &\leq \beta \Leftrightarrow 1 - \pi(\mu_1) \leq \beta \Leftrightarrow \pi(\mu_1) \geq 1 - \beta \\
 \Leftrightarrow 1 - \Phi \left(z_{\alpha} + \frac{\mu_o - \mu_1}{\sigma_o/\sqrt{n}} \right) &\geq 1 - \beta \Leftrightarrow \Phi \left(z_{\alpha} + \frac{\mu_o - \mu_1}{\sigma_o/\sqrt{n}} \right) \leq \beta \Leftrightarrow z_{\alpha} + \frac{\mu_o - \mu_1}{\sigma_o/\sqrt{n}} \leq z_{1-\beta}
 \end{aligned}$$

Observemos que en este caso la alternativa es $H_1: \mu > \mu_0$, por lo tanto, $\mu_0 - \mu_1 < 0$ y se obtiene

$$n \geq \left(\frac{(z_\alpha - z_{1-\beta})\sigma_o}{\mu_1 - \mu_0} \right)^2 = \left(\frac{(z_\alpha + z_\beta)\sigma_o}{\mu_1 - \mu_0} \right)^2$$

$$b) P_{\mu_1} \left(\frac{\bar{X} - \mu_0}{\sigma_o / \sqrt{n}} > -z_\alpha \right) \leq \beta \Leftrightarrow 1 - \pi(\mu_1) \leq \beta \Leftrightarrow \pi(\mu_1) \geq 1 - \beta$$

$$\Leftrightarrow \Phi \left(-z_\alpha + \frac{\mu_0 - \mu_1}{\sigma_o / \sqrt{n}} \right) \geq 1 - \beta \Leftrightarrow -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma_o / \sqrt{n}} \geq z_\beta$$

Observemos que en este caso la alternativa es $H_1: \mu < \mu_0$, por lo tanto, $\mu_0 - \mu_1 > 0$ y se obtiene

$$n \geq \left(\frac{(z_\alpha + z_\beta)\sigma_o}{\mu_0 - \mu_1} \right)^2$$

c) Para el caso bilateral, el cálculo del tamaño de muestra se hace en forma aproximada, despreciando la más pequeña de las dos probabilidades.

Tests para la media cuando la varianza es desconocida: Supongamos ahora que la varianza es desconocida y consideremos las mismas hipótesis sobre μ .

- | | | |
|---|----|-----------------------|
| a) $H_0: \mu = \mu_0$ (ó $\mu \leq \mu_0$) | vs | $H_1: \mu > \mu_0$ |
| b) $H_0: \mu = \mu_0$ (ó $\mu \geq \mu_0$) | vs | $H_1: \mu < \mu_0$ |
| c) $H_0: \mu = \mu_0$ | vs | $H_1: \mu \neq \mu_0$ |

Estadístico del test: $T = \sqrt{n} \frac{\bar{X} - \mu_0}{S}$. Bajo $H_0: \mu = \mu_0$, $T \sim t_{n-1}$

Región de rechazo: Como siempre la forma de la zona de rechazo depende de la hipótesis alternativa. Estará dada, en cada caso, por

- $T \geq t_{n-1, \alpha}$
- $T \leq -t_{n-1, \alpha}$

c) $|T| \geq t_{n-1, \alpha/2}$

El tamaño de la zona de rechazo depende del nivel. Por ejemplo, consideremos el caso a). Como la alternativa es $\mu > \mu_0$, la forma de la región es $T \geq K$, pero como la probabilidad de rechazar H_0 siendo cierta, o sea la probabilidad de Error tipo I, debe ser α ,

$$P_{\mu_0} \left(\sqrt{n} \frac{\bar{X} - \mu_0}{S} \geq K \right) = \alpha \Leftrightarrow 1 - P_{\mu_0} \left(\sqrt{n} \frac{\bar{X} - \mu_0}{S} \leq K \right) = \alpha$$

$$\Leftrightarrow 1 - F_T(K) = \alpha \Leftrightarrow F_T(K) = 1 - \alpha \Leftrightarrow K = t_{n-1, \alpha}$$

donde F_T designa la función de distribución de una v.a. t con $n-1$ grados de libertad.

Función de potencia y cálculo del tamaño de muestra para obtener una probabilidad de error tipo II dada: La función de potencia de este test es complicada porque la distribución del estadístico cuando $\mu \neq \mu_0$ es una distribución t no central. Aunque hay tablas y gráficos que permiten obtener probabilidades para una distribución de este tipo, no los estudiaremos en este curso. Por la misma razón, no calcularemos tamaño de muestra para obtener una probabilidad de error tipo II dada para una alternativa fija.

Respecto al p-valor, cuando se utilizan tablas sólo es posible obtener una cota, ya que las tablas proveen solamente algunos valores críticos de la distribución t .

Tests para la varianza cuando la media es desconocida: Las hipótesis a testear son

- a) $H_0: \sigma^2 = \sigma_0^2$ (ó $\sigma^2 \leq \sigma_0^2$) vs $H_1: \sigma^2 > \sigma_0^2$
 b) $H_0: \sigma^2 = \sigma_0^2$ (ó $\sigma^2 \geq \sigma_0^2$) vs $H_1: \sigma^2 < \sigma_0^2$
 c) $H_0: \sigma^2 = \sigma_0^2$ vs $H_1: \sigma^2 \neq \sigma_0^2$

Estadístico del test: $U = \frac{(n-1)S^2}{\sigma_0^2}$. Bajo $H_0: \sigma^2 = \sigma_0^2$, $U \sim \chi_{n-1}^2$.

Región de rechazo: Como siempre la forma de la zona de rechazo depende de la hipótesis alternativa. En este caso, estará dada por

- a) $U \geq \chi_{n-1, \alpha}^2$
 b) $U \leq \chi_{n-1, 1-\alpha}^2$
 c) $U \geq \chi_{n-1, \alpha/2}^2$ ó $U \leq \chi_{n-1, 1-\alpha/2}^2$

El tamaño de la zona de rechazo depende del nivel. Por ejemplo, consideremos el caso b). Como la alternativa es $\sigma^2 < \sigma_0^2$, la forma de la región es $U \leq K$, pero como la

probabilidad de rechazar H_0 siendo cierta ($P(\text{Error tipo I})$) debe ser α ,

$$P_{\sigma_0^2} \left(\frac{(n-1)S^2}{\sigma_0^2} \leq K \right) = \alpha \Leftrightarrow K = \chi_{n-1, \alpha}^2$$

Función de potencia: La obtendremos sólo para el caso b). Los otros dos casos se obtienen en forma similar.

$$\pi(\sigma_1^2) = P_{\sigma_1^2} \left(\frac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1, 1-\alpha}^2 \right) = P_{\sigma_1^2} \left(\frac{(n-1)S^2}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi_{n-1, 1-\alpha}^2 \right) = F_{\chi_{n-1}^2} \left(\frac{\sigma_0^2}{\sigma_1^2} \chi_{n-1, 1-\alpha}^2 \right)$$

donde $F_{\chi_{n-1}^2}$ indica la función de distribución chi-cuadrado con $n-1$ grados de libertad.

Utilizando tablas sólo es posible obtener una cota para la potencia ya que las tablas proveen solamente algunos valores críticos de la distribución χ^2 .

Por la misma razón, al calcular el p-valor utilizando tablas, sólo es posible obtener una cota.

Ejercicio: ¿Qué estadístico utilizaría en el caso en que la media μ fuese conocida?. ¿Cuál es la distribución de dicho estadístico? ¿Cómo se modifican las regiones de rechazo y la función de potencia de los tests?

Ejemplo: Se toman 25 determinaciones de la temperatura en cierto sector de un reactor, obteniéndose

$$\bar{x} = 243^\circ C \quad \text{y} \quad s = 2.8^\circ C$$

Interesa saber, a nivel 0.05

- si existe evidencia para decidir que la temperatura media en ese sector del reactor es menor que $250^\circ C$.
- si existe evidencia para decidir que la varianza de la temperatura en ese sector del reactor es mayor que $(2^\circ C)^2$.

a) Las hipótesis a testear son

$$H_0: \mu = 250 \text{ (ó } \mu \geq 250) \quad \text{vs} \quad H_1: \mu < 250$$

El estadístico del test será $T = \sqrt{n} \frac{\bar{X} - 250}{S}$,

y la región de rechazo estará dada por los valores de T tales que

$$T = \sqrt{n} \frac{\bar{X} - 250}{S} \leq -t_{n-1, 0.05}$$

En nuestro caso, $n = 25$ y por lo tanto $-t_{24, 0.05} = -1.71$. Como el valor observado de T es -12.5 , se rechaza H_0 , es decir hay evidencia de que la temperatura media del reactor es menor que $250^\circ C$.

b) Las hipótesis a testear son

$$H_0: \sigma^2 = 4 \text{ (ó } \sigma^2 \leq 4) \quad \text{vs} \quad H_1: \sigma^2 > 4$$

El estadístico del test será $U = \frac{(n-1)S^2}{\sigma^2}$,

y la región de rechazo estará dada por los valores de U tales que

$$U = \frac{(n-1)S^2}{4} \geq \chi_{n-1, 0.05}^2$$

En nuestro caso, $n = 25$ y por lo tanto $\chi_{24, 0.05}^2 = 36.42$. Como el valor observado de U es 47.04 , se rechaza H_0 . Es decir, hay evidencia de que la varianza de la temperatura del reactor es mayor que $(2^\circ C)^2$.

Tests de hipótesis de nivel aproximado (o asintótico) α para la media de una distribución cualquiera:

Sea X_1, X_2, \dots, X_n una m.a. de una distribución con media μ y varianza $\sigma^2 < \infty$. Aplicando el Teorema Central del Límite, sabemos que

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

Además, utilizando la propiedad enunciada al construir intervalos de confianza de nivel asintótico $(1 - \alpha)$ para la media de una distribución cualquiera,

$$\left. \begin{array}{l} \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1) \\ \frac{\sigma}{S} \xrightarrow{p} 1 \end{array} \right\} \Rightarrow \sqrt{n} \frac{\bar{X} - \mu}{S} \xrightarrow{d} N(0,1)$$

Por lo tanto, si n es suficientemente grande,

$$\sqrt{n} \frac{\bar{X} - \mu}{S} \stackrel{(a)}{\sim} N(0,1)$$

Supongamos que se desea testear a nivel aproximado α alguna de las hipótesis siguientes:

- | | | |
|---|----|-----------------------|
| a) $H_0: \mu = \mu_0$ (ó $\mu \leq \mu_0$) | vs | $H_1: \mu > \mu_0$ |
| b) $H_0: \mu = \mu_0$ (ó $\mu \geq \mu_0$) | vs | $H_1: \mu < \mu_0$ |
| c) $H_0: \mu = \mu_0$ | vs | $H_1: \mu \neq \mu_0$ |

y que n es suficientemente grande. Utilizando como estadístico $T = \sqrt{n} \frac{\bar{X} - \mu_0}{s}$, las siguientes regiones de rechazo proveen tests de nivel aproximado α para cada una de las hipótesis:

- a) $T \geq z_{\alpha}$
- b) $T \leq -z_{\alpha}$
- c) $|T| \geq z_{\alpha/2}$

Función de potencia aproximada: Un estimador de la función de potencia puede obtenerse reemplazando el estadístico S por su valor observado s , o sea:

$$\begin{aligned} \pi(\mu) &= P_{\mu} \left(\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| \geq z_{\alpha/2} \right) = 1 - P_{\mu} \left(\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| \leq z_{\alpha/2} \right) \\ &= 1 - P_{\mu} \left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu + \mu - \mu_0}{s/\sqrt{n}} \leq z_{\alpha/2} \right) \\ &= 1 - P_{\mu} \left(-z_{\alpha/2} + \frac{\mu_0 - \mu}{s/\sqrt{n}} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq z_{\alpha/2} + \frac{\mu_0 - \mu}{s/\sqrt{n}} \right) \end{aligned}$$

$$\cong 1 - \Phi\left(z_{\alpha/2} + \frac{\mu_o - \mu}{s/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} + \frac{\mu_o - \mu}{s/\sqrt{n}}\right)$$

En forma similar, se obtiene la función de potencia aproximada en los otros dos casos.

Ejemplo: En algunos casos, la varianza y la media dependen del mismo parámetro y no es necesario reemplazar σ por un estimador. Por ejemplo sea X_1, X_2, \dots, X_n una m.a. de una distribución de Poisson de parámetro λ . Entonces, si n es suficientemente grande,

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \stackrel{(a)}{\sim} N(0,1)$$

Supongamos que se desea testear a nivel aproximado α

$$H_0: \lambda = \lambda_o \quad \text{vs} \quad H_1: \lambda > \lambda_o$$

Entonces, bajo H_0 ,

$$\frac{\bar{X} - \lambda_o}{\sqrt{\lambda_o/n}} \stackrel{(a)}{\sim} N(0,1)$$

y, el test con región de rechazo

$$\frac{\bar{X} - \lambda_o}{\sqrt{\lambda_o/n}} \geq z_\alpha$$

tiene nivel aproximado α .

Su función de potencia aproximada se obtiene en la forma siguiente:

$$\begin{aligned} \pi(\lambda_1) &= P_{\lambda_1} \left(\frac{\bar{X} - \lambda_o}{\sqrt{\lambda_o/n}} \geq z_\alpha \right) = P_{\lambda_1} \left(\bar{X} \geq z_\alpha \sqrt{\lambda_o/n} + \lambda_o \right) = P_{\lambda_1} \left(\frac{\bar{X} - \lambda_1}{\sqrt{\lambda_1/n}} \geq \frac{z_\alpha \sqrt{\lambda_o/n} + \lambda_o - \lambda_1}{\sqrt{\lambda_1/n}} \right) \\ &\cong 1 - \Phi \left(z_\alpha \sqrt{\frac{\lambda_o}{\lambda_1}} + \frac{\lambda_o - \lambda_1}{\sqrt{\lambda_1/n}} \right) \end{aligned}$$

Test de hipótesis de nivel aproximado (o asintótico) α para una proporción (parámetro p de la distribución binomial): Sea X_1, X_2, \dots, X_n una m.a. de una

distribución $\text{Bi}(1, p)$. Entonces, $X = \sum_{i=1}^n X_i \sim \text{Bi}(n, p)$. Aplicando el Teorema Central del Límite, si n es suficientemente grande,

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} Z \sim N(0,1)$$

siendo \bar{X} la proporción muestral o frecuencia relativa de éxitos.

Un test de nivel aproximado α para las hipótesis:

- a) $H_0: p = p_0$ vs $H_1: p > p_0$
- b) $H_0: p = p_0$ vs $H_1: p < p_0$
- c) $H_0: p = p_0$ vs $H_1: p \neq p_0$

se basa en el estadístico $\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, el cual, si H_0 es cierta, tiene distribución

aproximada $N(0,1)$. Las regiones de rechazo estarán dadas por

a) $\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_\alpha$

b) $\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_\alpha$

c) $\left| \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| \geq z_{\alpha/2}$

Ejercicio: Deducir las funciones de potencia aproximadas en los 3 casos.

Relación entre tests de hipótesis bilaterales e intervalos de confianza:

Introduciremos esta idea a través de un ejemplo. Sea X_1, X_2, \dots, X_n una m.a. de una distribución $N(\mu, \sigma^2)$. Sabemos que el intervalo de confianza para μ de nivel $1 - \alpha$ está dado por

$$\left[\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right].$$

Supongamos ahora que deseamos testear a nivel α las siguientes hipótesis:

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu \neq \mu_0$$

Dado que el intervalo construido contiene con alta probabilidad al valor verdadero de μ , si μ_0 no pertenece al intervalo, ésto nos llevaría a sospechar que la hipótesis nula es falsa.

Es decir, podríamos construir un test de nivel α rechazando H_0 si μ_0 no pertenece al intervalo de confianza, dado que

$$\begin{aligned} P(EI) &= P_{\mu_0} \left(\mu_0 \notin \left[\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right] \right) \\ &= 1 - P_{\mu_0} \left(\mu_0 \in \left[\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right] \right) = 1 - (1 - \alpha) = \alpha. \end{aligned}$$

Proposición: Sea $IC(X_1, X_2, \dots, X_n)$ un intervalo de confianza de nivel $1 - \alpha$ para un parámetro θ , obtenido a partir de una m.a. X_1, X_2, \dots, X_n . Consideremos el problema de testear las hipótesis

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta \neq \theta_0$$

El test que rechaza H_0 cuando $\theta_0 \notin IC(X_1, X_2, \dots, X_n)$, tiene nivel α .

Ejemplo: Sea X_1, X_2, \dots, X_n una m.a. de una distribución exponencial de parámetro λ .

Recordemos que, usando que $2\lambda \sum_{i=1}^n X_i \sim \chi_{2n}^2$, hemos obtenido el siguiente intervalo de confianza de nivel exacto $1 - \alpha$ para λ

$$IC_{\lambda} = \left[\frac{\chi_{2n,1-\alpha/2}^2}{2 \sum_{i=1}^n X_i}, \frac{\chi_{2n,\alpha/2}^2}{2 \sum_{i=1}^n X_i} \right]$$

Si deseamos testear las hipótesis

$$H_0: \lambda = \lambda_0 \quad \text{vs} \quad H_1: \lambda \neq \lambda_0$$

El test que rechaza H_0 si $\lambda_0 \notin IC_{\lambda}$ tiene nivel α .