Análisis descriptivo. Nociones de probabilidad e inferencia.

24 de abril de 2022

1. Tipos de variables.

Según su nivel de medición, las variables se distinguen como:

1. Categóricas o cualitativas: Los niveles indican categorías y no están ordenados. Dichas categorías se identifican con «etiquetas» que son cadenas de caracteres («A», «Plural», «3»), los números que identifican categorías no poseen ni valor ni relación. Ejemplos: (i) Género, con niveles: «singular», «plural»; (ii) Animicidad, con niveles: «1» = animado, «0» = no animado (variable categórica binaria).

2. Ordinales o cuasi-cualitativas: Los niveles indican categorías que siguen un orden. No hay distancias entre las categorías (o sea si el orden es A > B > C; B no es el doble de A y C el triple). Ejemplo: Frecuencia Léxica, con niveles: «alto», «medio», «bajo»; nivel de competencia lingüística en L2, con niveles: «A1», «A2», «B1», «B2», «C1», «C2».

3. Cuantitativas discretas: Sus valores pertenecen a los números naturales ($\mathbb{N}$ incluido el cero) . Entre valores consecutivos no existen valores intermedios. Toman un conjunto finito (Ej.: $\{0, 1, 2, 3, 4\}$) o infinito ($\{0, 1, 2, 3, \ldots\}$) numerable de valores. Surgen de un *conteo*. Ejemplo: Frecuencia Léxica absoluta, largo de una palabra (en letras).

4. Cuantitativas continuas: Entre dos determinados valores hay infinitos valores intermedios. Los valores pertenecen a los números reales ($\mathbb{R}$). Toman infinitos valores dentro de un intervalo de los números reales. Surgen de una *medición*. Ejemplo: Tiempo de reacción, Promedio de longitud de pausas.

A las variables cuantitativas se le pueden aplicar operaciones (suma, resta, etc.) y transformaciones de cambio de escala (por ejemplo, pasar a escala logarítmica). Las escalas subjetivas de progresión en un continuo a menudo se consideran cuantitativas; por ejemplo registrar la respuesta de un individuo a la pregunta: «¿Cuán familiar le resulta la palabra X a usted en una escala continua de 1 a 7?». Otros ejemplos de variables lingüísticas son: Voz, Persona, Lenguaje Nativo, Concretud (escala subjetiva en continuo de 1 a 7), Cantidad de vecinos ortográficos (Ej.: palo, malo, ralo), roles temáticos, jerarquía de relativización[1], cantidad de concordancias correctas en una entrevista oral de español L2, Promedio del número de palabras por unidad de habla, Promedio de palabras por cláusula, Proporción de cláusulas sin errores, Tiempo de habla total, Distancia entre «controlador» y «objetivo» en la concordancia (registrado como cantidad de palabras o de nodos en la jerarquía sintáctica), Tipo de error de concordancia, Tamaño de la familia morfológica de una palabra (el número de palabras complejas en las que aparece como constituyente), etc.

Veamos ahora cómo organizar los datos. En el caso de las variables cualitativas se cuentan la cantidad de casos observados para cada categoría en la muestra. Se construye así una distribución de frecuencias, donde a cada clase o categoría se le asigna su frecuencia absoluta. Por ejemplo, el siguiente cuadro ilustra la cantidad de sustantivos controladores en la concordancia según su tipo de animicidad (en un corpus de español L2). Observamos que *524* controladores nominales de la muestra son animados.

Variable	Nivel	Frecuencia
Animicidad	0 = No animado	1333
	1 = Animado	524

Cuadro 1: Ejemplo de distribución de frecuencias.

Pasemos a los datos cuantitativos. Si son discretos, los niveles se definen según los valores que puede tomar la variable. Si son continuos, es preciso construir los niveles como intervalos de clase. Luego se asignan las frecuencias absolutas a cada nivel. Por ejemplo, el cuadro *2* muestra la distribución de frecuencias para la cantidad de errores de concordancia en

[1] Se trata de una escala implicativa del tipo «las lenguas que relativizan el objeto directo también relativizan el sujeto». Ej.: Sujeto > Objeto Directo > Objeto Indirecto > Poseedor.

una tarea escrita para estudiantes de nivel intermedio de español L2. Observamos que en *10* escritos de la muestra no se cometió ningún error, en *5* escritos se cometieron *20* errores, etc.

# errores	Frecuencia
0	10
5	20
8	7
10	5
14	3
...	...

Cuadro 2: Cantidad de errores de concordancia en alumnos de nivel intermedio de español L2.

Por otro lado, el cuadro *3* indica el promedio de subordinadas en una tarea de narración oral para alumnos de nivel superior de español L2. Se observa que *15* narraciones de la muestra evidencian entre *0* y *5* subordinadas en promedio; y solamente *2* narraciones evidencian entre *15* y *20* subordinadas en promedio.

Intervalo	Frecuencia
[0, 5)	15
[5, 10),	9
[10, 15)	6
[15, 20)	2
...	...

Cuadro 3: Cantidad de errores de concordancia en alumnos de nivel intermedio de español L2.

Ahora bien, muchas veces conviene expresar los datos en términos de proporciones (frecuencias porcentuales). La suma de las frecuencias absolutas de cada nivel de una variable es igual al número total de casos. O sea que para m niveles de la variable, con sus frecuencias f_i $(i = 1, \ldots, m)$ tenemos que: $f_1 + f_2 + \ldots + f_m = n$. Para sacar las frecuencias relativas de cada nivel hacemos: $f_{r_i} = \frac{f_i}{n}$ y luego la multiplicamos por cien para sacar la frecuencia porcentual: $f_{p_i} = f_{r_i} \times 100$. Así, por ejemplo, para el cuadro *2*, la frecuencia relativa del nivel cero (errores) es: $f_{r_1} = \frac{10}{10+20+7+5+3} \simeq 0{,}22$ y su frecuencia porcentual es $f_{p_1} = 0{,}22 \times 100 = 22\,\%$. Por supuesto, la suma de las frecuencias relativas de todos los niveles es 1 y la de las frecuencias porcentuales es 100.

Por otro lado, las frecuencias absolutas acumuladas para cada nivel surgen de la suma de la frecuencia absoluta del nivel corriente con las frecuencias absolutas de los niveles anteriores de la variable: $F_k = \sum_{i=1}^{k} f_i$. Por ejemplo, para el cuadro *2* tenemos:

# errores	f_i	F_i
0	10	10
5	20	30
8	7	37
10	5	42
14	3	45
...	...	...

Cuadro 4: Cantidad de errores de concordancia en alumnos de nivel intermedio de español L2.

2. Medidas de resumen muestrales.

Son medidas (estadísticos) que resumen en un solo valor características del conjunto de valores que componen la muestra.

2.1. Medidas de tendencia central.

Indican el centro de la distribución de un conjunto de datos: $\{x_1, x_2, \ldots, x_n\}$.

Media muestral. Es el «centro de gravedad» o punto de equilibrio de los datos: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$. La suma de las distancias de los datos a la media da cero: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Preserva la dependencia lineal, o sea: $y = ax+b \Rightarrow \bar{Y} = a\bar{x}+b$. Es sensible a valores extremos o *outliers* (valores muy alejados del conjunto de los datos), por ello no es una medida *robusta* (¡Basta solamente un valor extremo para que se modifique la media muestral!). No necesariamente es un dato presente en la muestra. Por ejemplo, si tenemos la muestra: $x = \{11, 11, 15, 18, 25\}$, entonces $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(11 + 11 + 15 + 18 + 25) = \frac{80}{5} = 16$.

Mediana muestral. Medida del centro de los datos que divide a la muestra ordenada en dos partes de igual tamaño. Es un estadístico de orden porque para calcularla se requiere ordenar los datos de menor a mayor: $x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}$, donde el (1) indica la posición del dato de menor valor en la muestra ordenada (que no necesariamente es el primer dato de la muestra) y (n) es el dato de mayor valor (que no necesariamente es el último dato de la muestra). La mediana, denotada como $\tilde{X}$, es el dato que ocupa la posición $x^{\left(\frac{n+1}{2}\right)}$ si el número total de datos es *impar*. O sea que en este caso es el dato de la posición central. Por otra parte, si la muestra es *par*, la mediana es el *promedio* de los dos valores centrales de la muestra ordenada: $\frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n}{2}+1\right)}}{2}$. La definición formal de la mediana es:

$$\tilde{X} = \begin{cases} x^{\left(\frac{n+1}{2}\right)} & \textit{si } n \textit{ es impar} \\ \frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n}{2}+1\right)}}{2} & \textit{si } n \textit{ es par} \end{cases}$$

Por ejemplo, para $x = \{11, 11, 15, 18, 25\}$ donde $n = 5$ es impar y los datos ya están ordenados de menor a mayor, el valor del centro es $\tilde{X} = x^{(3)} = 15$. Por otro lado, si $x = \{11, 11, 15, 18, 25, 10\}$, con $n = 6$ par, los datos ordenados son $\{10, 11, 11, 15, 18, 25\}$, con lo cual el promedio de los valores centrales (que dejan igual tamaño a derecha e izquierda de éstos) da $\tilde{X} = \frac{11+15}{2} = \frac{26}{2} = 13$.

Si una distribución es simétrica entonces la media es igual a la mediana, $\bar{X} = \tilde{X}$. Si la distribución es asimétrica a derecha, entonces $\bar{X} > \tilde{X}$; y si es asimétrica a izquierda, entonces: $\bar{X} < \tilde{X}$. Por ende, la media se corre según la dirección de la asimetría. La mediana es *robusta* a la presencia de valores extremos.

Moda. Es el dato de la muestra que tiene mayor *frecuencia*. Por ejemplo, para el caso anterior $x = \{11, 11, 15, 18, 25, 10\}$, la moda es $Mo = 11$ porque $f_1 = 2$ y los demás valores tienen $f_{i>1} = 1$. En cambio, si tuviéramos: $x = \{11, 11, 15, 18, 25, 25\}$ habría dos modas: $Mo = 11$ y $Mo = 25$ porque para ambos $f_1 = f_4 = 2$ (recordar que el subscrito identifica el nivel de la variable). Las distribuciones pueden ser unimodales, bimodales o multimodales.

Media α- podada. Es el promedio de los valores de la muestra una vez removidos el $\alpha\,\%$ de los valores más grandes y más pequeños. Por ejemplo, si tenemos la muestra $x = \{\mathbf{1}, \mathbf{3}, 5, 5, 5, 6, 7, 7, 9, 10, 13, 14, 15, 15, 15, 15, 16, 16, \mathbf{25}, \mathbf{30}\}$ y recortamos el $\alpha = 10\,\%$ de la muestra de $n = 20$, es decir que sacamos $0{,}1 \times 20 = 2$ datos de cada lado. Entonces «podamos» los dos valores más grandes y más pequeños de la muestra ordenada, aquí: $1, 3, 25, 30$; y calculamos la media sobre el resto de los valores: $\bar{x}_{0,10} = \frac{1}{16}\sum_{i=3}^{18} x_i = \frac{(5\times3)+6+(7\times2)+9+10+13+14+(15\times4)+(16\times2)}{16} = 10{,}8125$. Se trata de una medida robusta a *outliers*.

Medidas de posición (estadísticos de orden). Hemos visto que la mediana muestral es un estadístico de orden porque se calcula como función de los datos ordenados. Recordemos que los valores muestrales ordenados de forma ascendente son $X = x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}$. Entonces el dato en primera posición $x^{(1)} = \min_{1\leq i\leq n} X$ es el mínimo y el de la última posición es el máximo: $x^{(n)} = \max_{1\leq i\leq n} X$. La diferencia entre el máximo y el mínimo es el rango muestral: $Rg\,(X) = x^{(n)} - x^{(1)}$.

Cuantil. Los cuantiles son valores (*observados o no*) del recorrido de una variable (los valores que puede tomar una variable) que dividen el conjunto de datos en partes iguales, es decir, con la misma cantidad de observaciones en cada parte. Los *cuartiles* Q dividen al conjunto de datos en cuatro partes iguales (corresponden a los cuantiles 0.25; 0.50 y 0.75); los *deciles* D lo dividen en diez partes iguales y los *percentiles* P lo hacen en cien partes iguales.

Percentil. El percentil es una medida de posición que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo. Por ejemplo, el percentil 20º es el valor bajo el cual se encuentran el 20 por ciento de las observaciones, y el 80 % restante son mayores. Formalmente, sea $0 \leq p \leq 1$, entonces el $(100\,p) - ésimo$ percentil muestral es la observación tal que np de las observaciones son menores

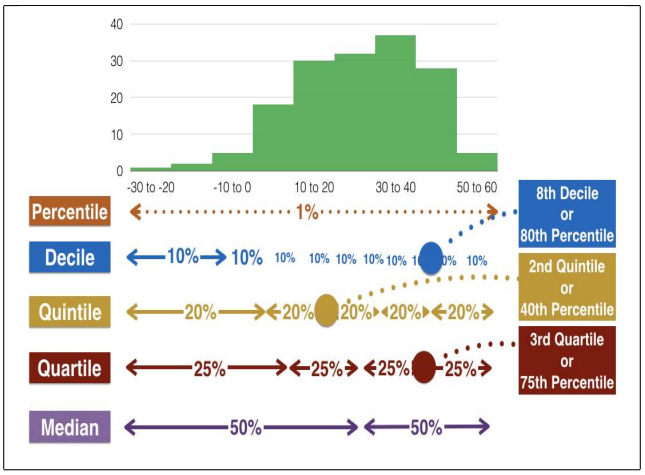
y $n\,(1-p)$ son mayores (n es el tamaño muestral). La siguiente es la definición formal, representando $\{b\}$ un número redondeado.

$$(100\,p) - \acute{e}simo\ percentil\ muestral = \begin{cases} x^{(\{np\})} & \frac{1}{2n} < p < 0{,}5 \\ x^{(n+1-\{n(1-p)\})} & 0{,}5 < p < 1 - \frac{1}{2n} \end{cases}$$

Por ejemplo, si $n = 12$, el percentil del 65 % es la novena observación de la muestra ordenada:

$$x^{(n+1-\{n(1-p)\})} = x^{(12+1-\{12(1-0{,}65)\})} = x^{(13-\{12(0{,}35)\})} = x^{(13-\{4{,}2\})} Si = x^{(13-4)} = x^{(9)}$$

Cuartil. El percentil del 25 % corresponde al primer cuartil de los datos (o sea que el 25 % de los datos ordenados se encuentra debajo del cuantil correspondiente); el del 50 % corresponde a la mediana (o sea que el 50 % de los datos ordenados se encuentra debajo del cuantil correspondiente); el del 75 % al tercer cuartil (o sea que el 75 % de los datos ordenados se encuentra debajo del cuantil correspondiente). Los cuartiles son valores (observados o no) de la variable que dividen las observaciones en cuatro partes iguales. Se denotan como Q_1, Q_2, Q_3. La diferencia entre el tercer y el primer cuartil es el *rango intercuartil*: $R.I. = Q_3 - Q_1$.



2.2. Medidas de dispersión.

Miden variabilidad o distancia al centro de los datos.

Varianza muestral, desvío estándar muestral y coeficiente de variación. Es el promedio de los cuadrados de las distancias entre las observaciones y la media muestral: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2$.

Se usa el cuadrado de los desvíos del centro de los datos para eliminar los signos (ya que, como vimos, la suma de los desvíos es cero). Se divide por $n-1$ para que sea un estadístico *insesgado* de la varianza poblacional. Si $y = ax + b \Rightarrow S_y^2 = a^2 S_x^2$. La varianza también es un promedio, por tanto, al igual que la media muestral no resulta ser un estadístico robusto a la presencia de *outliers*.

El *desvío estándar muestral* es la raíz cuadrada de la varianza muestral: $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x - \bar{X}\right)^2}$. Un desvío bajo significa que los datos muestrales tienden a estar agrupados cerca de su media muestral.

El coeficiente de variación es una medida de dispersión relativa. Mide la proporción que representa el desvío estándar de la media aritmética. Es: $CV = \frac{S}{\bar{X}}$. El CV se usa para comparar la dispersión de dos variables con medias o unidades de medición diferentes. Toma valores entre cero y uno y se lo expresa usualmente como proporción. Si está cerca de cero significa que es una muestra «compacta», con poca variabilidad de los datos y lo contrario sucede si se encuentra cerca de uno.

Vale decir que, cuando la media muestral no es adecuada como resumen de la tendencia central de una muestra tampoco la varianza, desvío o CV resultarán adecuados para medir su dispersión.

Rango muestral y rango intercuartil. Ya hemos visto que el rango es la diferencia entre el valor más grande y más pequeño de los datos. También mencionamos que el rango intercuartil es la diferencia entre el tercer y el primer cuartil. El rango muestral está influenciado por *outliers*, mientras que el intercuartil no lo está. Además dos variables podrían tener diferente dispersión alrededor de su media pero igual rango muestral, por lo cual no resulta una medida muy efectiva.

MAD [Median Absolute Deviation]. Es la mediana de los desvíos absolutos respecto de la mediana: $MAD = mediana\left(\mid x_i - \tilde{X} \mid\right)$. Es la versión robusta de la varianza muestral. Para calcularlo hay que: (i) ordenar los datos de menor a mayor; (ii) calcular la mediana; (iii) calcular la distancia de cada dato a la mediana y luego ordenar dichas distancias de menor a mayor; (iv) buscar la mediana de las distancias sin signo (en valor absoluto). Por ejemplo, en la siguiente muestra (ya ordenada) $x = \{2, 3, 5, 8, 15, 50\}$ el último dato es un *outlier*. La mediana es $\tilde{X} = \frac{5+8}{2} = \frac{13}{2} = 6{,}5$. Los desvíos respecto de la mediana son: $(-4{,}5, -3{,}5, -1{,}5, 1{,}5, 8{,}5, 43{,}5)$. Los valores absolutos de estos desvíos en orden creciente son: $(1{,}5, 1{,}5, 3{,}5, 4{,}5, 8{,}5, 43{,}5)$. Entonces la mediana de los valores absolutos de los desvíos será: $MAD = \frac{3{,}5+4{,}5}{2} = \frac{8}{2} = 4$. Para estandarizar la medida se hace: $MADN = \frac{MAD}{0{,}6745}$.

2.3. Otras medidas para caracterizar distribuciones.

Coeficiente de asimetría muestral de Fisher. Mide la asimetría de la distribución de los datos respecto de la media muestral: $sk_F(x) = \frac{\sqrt{n}\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^3}{\left[\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2\right]^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$ (μ_3 es el tercer momento central).

Para una distribución simétrica (por ejemplo, la normal): $sk_F(x) \approx 0$ y $\bar{X} = \tilde{X} = Mo$. Si $sk_F(x) > 0$ es asimétrica a derecha ($\bar{X} > \tilde{X}$) y si $sk_F(x) < 0$, asimétrica a izquierda ($\bar{X} < \tilde{X}$). Como mencionamos antes, la media muestral sigue a la dirección de la asimetría.

Coeficiente de curtosis muestral. Una curtosis grande implica una mayor concentración de valores de la variable tanto muy cerca de la moda (media) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios. Esto explica una forma de la distribución de frecuencias/probabilidad con colas más gruesas, con un centro más apuntado y una menor proporción de valores intermedios entre el pico y colas. Se define como: $k(x) = \frac{n\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^4}{\left[\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2\right]^2} = \frac{\mu_4}{\sigma^4}$ (μ_4 es el cuarto momento central). Para la normal $k(x) \approx 3$. O sea que cuando el coeficiente es superior a *tres*, la distribución será más «puntiaguda» y con colas más gruesas que la normal; y menor a *tres*, menos «puntiaguda» y con colas menos gruesas que la normal. Sin embargo una mayor curtosis no implica mayor varianza ni viceversa.

3. Gráficos.

Las variables cualitativas o cuasi-cualitativas pueden representarse mediante gráficos circulares (tipo «torta») y de barras. Bresnan & col. (2007) estudiaron la alternancia de dativo en con datos extraídos de dos corpus. En inglés el «recipiente» puede realizarse como NP (*Mary gave John the book*) o PP (*Mary gave the book to John*). La cuestión radicaba en predecir el recipiente a partir de una serie de variables: verbos involucrados, clase semántica de dichos verbos, animicidad, accesibilidad discursiva, si hay o no un pronombre, largo de la palabra, si es o no definido (los tres últimos tanto para el objeto que se mueve o transfiere como para el recipiente). Ahora bien, los verbos fueron clasificados según los siguientes niveles teniendo en cuenta sus usos en la construcción de dativo: (a) «Abstracto»: *give it some thought*; (t) «transferencia de posesión»: *give an armband, send*; (f) «tranferencia futura de posesión»: *owe, promise*; (p) «prevención de posesión»: *cost, deny*; (c) «comunicación»: *tell, give me your name*. Los gráficos circulares siguientes muestran la distribución de dichas categorías semánticas.

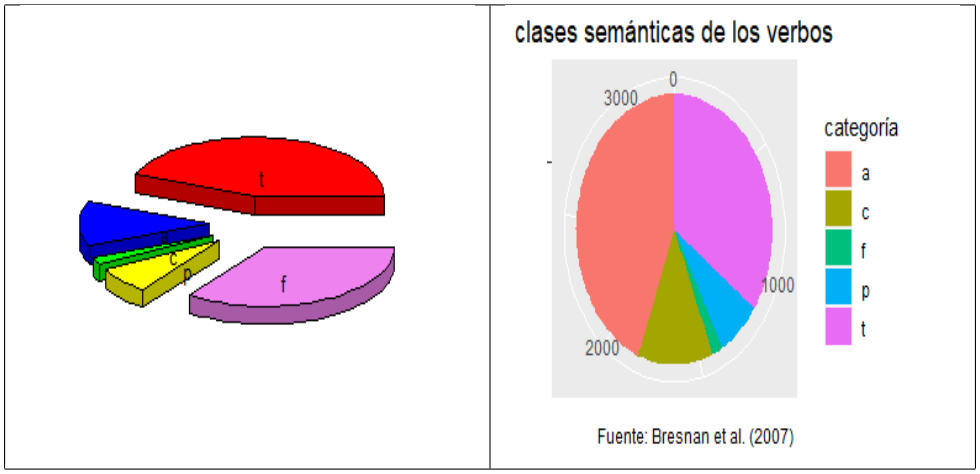


Figura 1: Gráficos circulares.

Por otro lado, el gráfico de barras siguiente ilustra las frecuencias absolutas del verbo «pay» en sus sentidos «abstracto» [a] («pay attention») y de «transferencia de posesión» [t] («pay money»). El asunto es comparar la animicidad y la accesibilidad discursiva del «recipiente» en la construcción de dativo, para ambos sentidos del verbo. Según el gráfico, el sentido de «transferencia» es más probable que ocurra con recipientes animados y que ya se conocen. Por otra parte, el sentido abstracto es más compatible con recipientes no animados. Por otra parte, la diferencia entre dado y no dado en este último caso no parece ser tan grande. Los datos están en el cuadro *5*. Por otro lado, los gráficos de barra pueden mostrar dichos datos por separado de modo adyacente o de manera superpuesta, como se ve en las figuras. Son útiles para representar la distribución de dos sub-conjuntos de individuos.

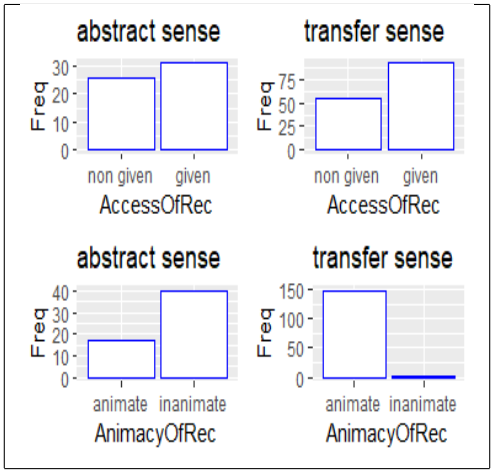


Figura 2: Gráficos de barras por separado.

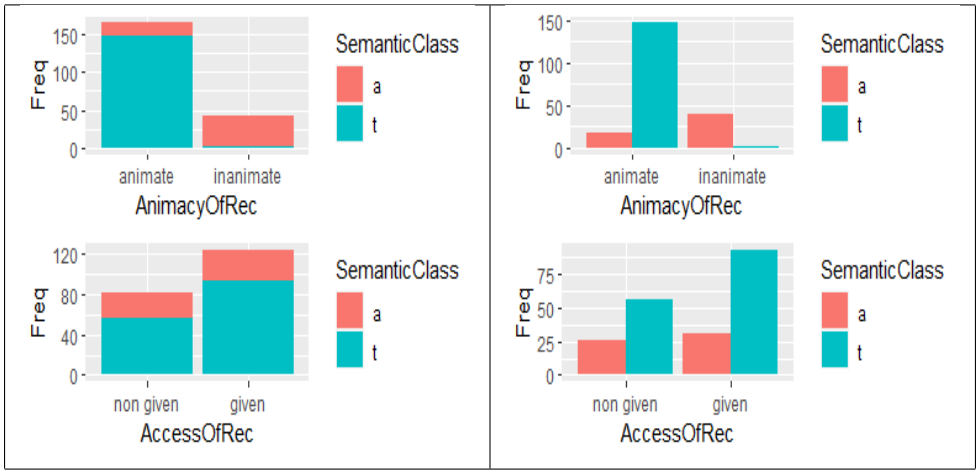


Figura 3: Gráficos de barras superpuestas y adyacentes.

Sentidos	Animate	Inanimate	Given	Non given
«pay» (transfer)	147	2	93	56
«pay» (abstract)	17	40	31	26

Cuadro 5: Categorías cruzadas entre: (i) animicidad y sentidos de «pay»; (ii) accesibilidad y sentidos de «pay».

Miremos ahora el siguiente cuadro de categorías cruzadas entre animicidad (animado / no animado) y realización del recipiente (NP o PP):

	NP	PP
Animado	517	300
No animado	33	47

Cuadro 6: Categorías cruzadas entre animicidad y realización del recipiente.

Dicha información puede representarse mediante un gráfico de mosaico del siguiente modo. El tamaño de los bloques se corresponde con la frecuencia. Por ende, se ve que los recipientes animados tienden a realizarse como NP y los inanimados como PP.

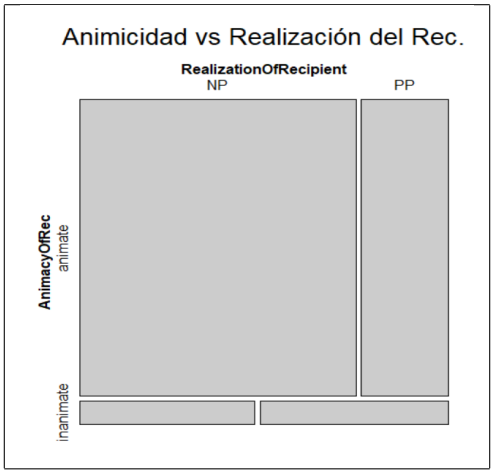


Figura 4: Gráfico Mosaico.

El gráfico de bastones se usa para representar la distribución de frecuencias de variables cuantitativas *discretas*. El siguiente gráfico muestra la distribución de frecuencias absolutas del largo (en cantidad de letras) de *81* palabras que tienen que ver con animales y plantas. Se observa que la moda está en *5* letras.

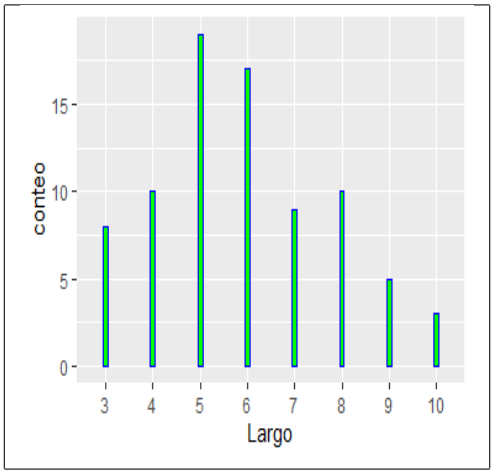
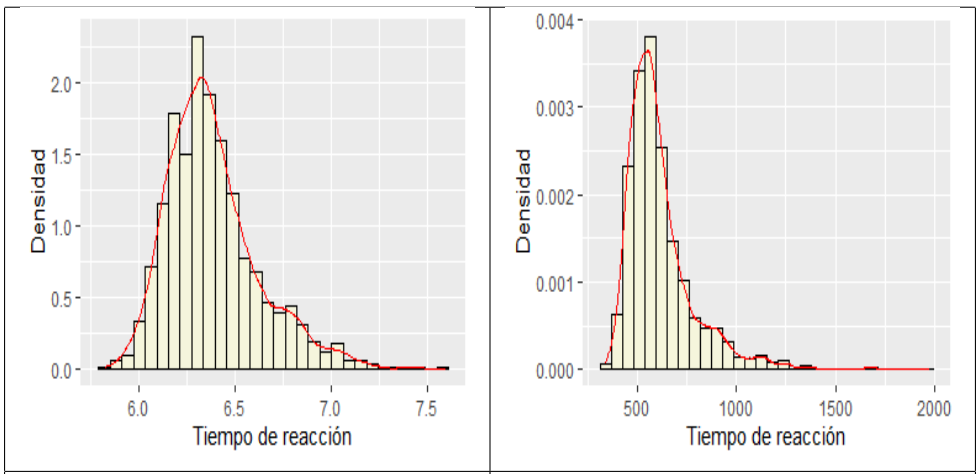


Figura 5: Gráfico de bastones: largo de 81 palabras.

Un histograma se construye dividiendo en intervalos el rango de valores de los datos y asignando a cada intervalo la cantidad o proporción de observación que caen dentro éste. Es adecuado para graficar la distribución de variables *cuantitativas continuas*. El histograma es altamente dependiente de la cantidad de intervalos o bien del largo de dichos intervalos. Algunas fórmulas para calcularlos son las siguientes. Notar que siempre dependen de la cantidad n de datos. Para la cantidad de intervalos: (i) $k = \lfloor 10log\,(n) \rfloor$; (ii) $k = \lfloor 2\sqrt{n} \rfloor$; (iii) $k = \lfloor 1 + log_2\,(n) \rfloor$; donde $\lfloor \ldots \rfloor$indica la parte entera. Para el largo de los intervalos: (i) $h_n = 3{,}49sn^{-1/3}$; (ii) $h_n = 2Rn^{-1/3}$; donde s indica el desvío estándar y R es el rango intercuartil. El siguiente histograma muestra los tiempos de reacción para *79* de las *81* palabras mencionadas antes. Está hecho con *30* intervalos. A la izquierda se muestran los tiempos de reacción en escala logarítmica y a la derecha sin dicha escala. En el gráfico también se muestra una estimación no paramétrica de la densidad. Notar que la escala logarítmica reduce la asimetría a derecha propia de los datos.



con escala logarítmica	sin escala logarítmica

Figura 6: Histograma: tiempo de reacción para 79 palabras.

Otro gráfico de suma utilidad es el *box plot* o gráfico de cajas. Se construye así: (i) se dibuja un rectángulo o «caja» cuyos extremos inferior y superior se corresponden con el primer y tercer cuartil respectivamente; (ii) dentro del rectángulo se dibuja una línea que corresponde al segundo cuartil o mediana; (iii) a partir de cada extremo se dibuja un segmento («bigote») hasta el dato más alejado que se encuentra, a lo sumo, a $1{,}5 \times RI$ del extremo de la caja. Se clasifican como *outliers moderados* a los datos más allá de $1{,}5 \times RI$ y menores a $3 \times RI$ del extremo de la caja. Por otra parte, se clasifican como *outliers extremos* aquellos datos más allá de $3 \times RI$ del extremo de la caja. A partir de un *box plot* se puede recabar información de la distribución sobre: posición (mediana), dispersión (R. I.), asimetría, *outliers*. Los *outliers* pueden deberse a errores (valores imposibles para una variable) o bien a valores de individuos pertenecientes a una población diferente a la que se desea estudiar. Si se dan estos casos es lícito eliminar dichos datos extremos. Pero si no, *es un error eliminar datos*. Dichos *outliers* pueden indicar un conjunto especial de datos con comportamiento diferente del resto que valdría la pena estudiar aparte o bien que hubiera un error con la escala de medición elegida. Para ejemplificar el *box plot* tomemos la siguiente muestra, que corresponde a la Figura 7.

$$x = \{13, 17, 23, 25, 34, 38, 42, 44, 55, 61, 67, 91, 197\}$$

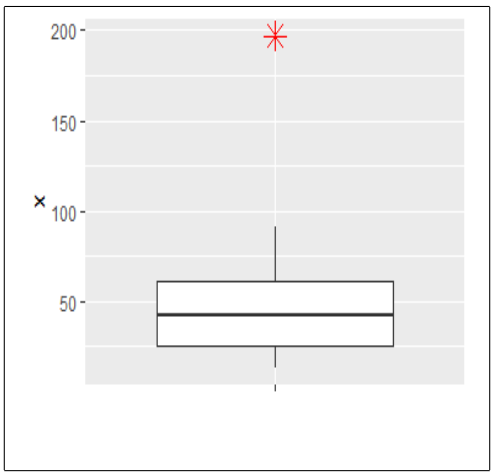


Figura 7: Boxplot de la muestra.

Analicemos con más detalle: $\bar{X} = 54{,}38$ (no está representada en el boxplot); $\tilde{X} = 42$; $Q_1 = 25$; $Q_3 = 61$; $RI = 61 - 25 = 36$; $Q_3 + 1{,}5 \times RI = 61 + 54 = 115$;$Q_1 - 1{,}5 \times RI = 25 - 54 = -29$; $Q_3 + 3 \times RI = 61 + 108 = 169$; $Q_1 - 3 \times RI = 25 - 108 = -83$. El valor adyacente superior (VAS) que corresponde al extremo del segmento superior es $VAS = 91$ porque es el valor observado inmediatamente inferior a $Q_3 + 1{,}5 \times RI = 115$. En cambio, el valor adyacente inferior (VAI) que x al extremo del segmento inferior es $VAI = 13$ porque es el valor observado inmediatamente inferior a $Q_1 - 1{,}5 \times RI = -29$. ¿Es 197 un *outlier* severo? Sí, porque $197 > Q_3 + 3 \times RI = 169$. Vemos además que la media muestral está corrida hacia el extremo superior de la caja (no representada en el gráfico); por ende, hay una asimetría a derecha. Los siguientes son tres boxplots correspondientes a tres situaciones diferentes de asimetría. Hay $n = 100$ datos en cada grupo G, con $G1 : X \sim N\left(\mu = 0, \sigma^2 = 1\right)\ [-\infty < x < \infty]$, $G2 : X \sim Exp\,(\lambda = 1)[x > 0]$, $G3 : X \sim Beta\,(\alpha = 5, \beta = 1)\ [0 < x < 1]$. En $G1$ la distribución es simétrica: la mediana está hacia el centro de la caja y los «bigotes» son aproximadamente iguales. En $G2$ la distribución es asimétrica a derecha: el «bigote» superior es más largo que el inferior y la mediana está corrida hacia el extremo inferior de la caja (además se observan *outliers*). En $G3$ la distribución es asimétrica a izquierda: el «bigote» inferior es más largo que el superior y la mediana está corrida hacia el extremo superior.

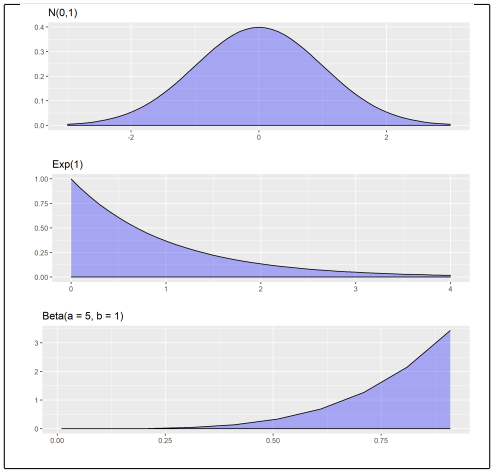


Figura 8: distribuciones que generan los datos de cada grupo: G1: N(0,1), G2: Exp(1), G3: Beta(5,1).

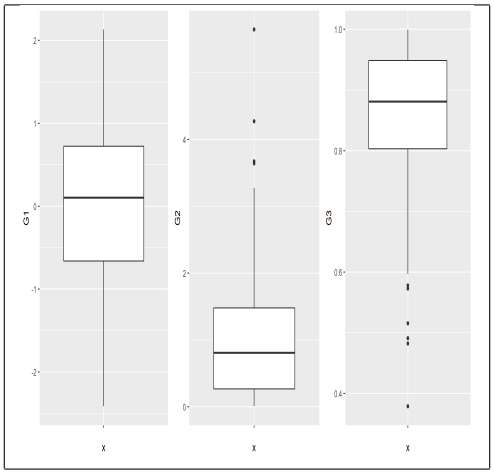


Figura 9: Boxplots simétrico (G1), asimétrico a derecha (G2), asimétrico a izquierda (G3).

Ahora veamos un ejemplo real. La siguiente figura muestra los tiempos de reacción para una tarea de decisión léxica de *79* palabras que tenían que ver con animales o plantas (cada sujeto daba latencias para las mismas *79* palabras). Se ilustra la distribución de los tiempos de reacción para las respuestas correctas e incorrectas, discriminados por lenguaje nativo (inglés) o no nativo (otro). Se observa que los *R.I.* (dispersión) en el caso de los hablantes nativos son similares, siendo la mediana de las respuestas incorrectas menor a la de las respuestas correctas (pero dicha diferencia parece poca, ¿será significativa?). Además las medianas y los *R.I.* de las respuestas son menores que en el caso de los hablantes no nativos, confirmando el hecho de que los hablantes nativos tardan menos en responder a la tarea que los no nativos. En el caso de los hablantes no nativos se ve que la dispersión (*R.I.*) y la mediana es mayor para las respuestas incorrectas que para las correctas. Por último se detectan muchos *outliers* solamente en el caso de las respuestas correctas en ambos grupos. Al tratarse de tiempos de reacción, las distribuciones son todas asimétricas a derecha.

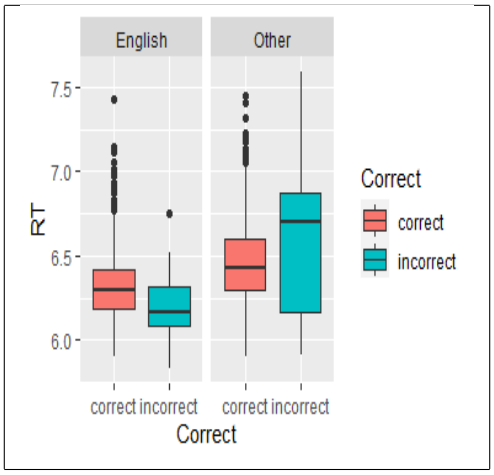
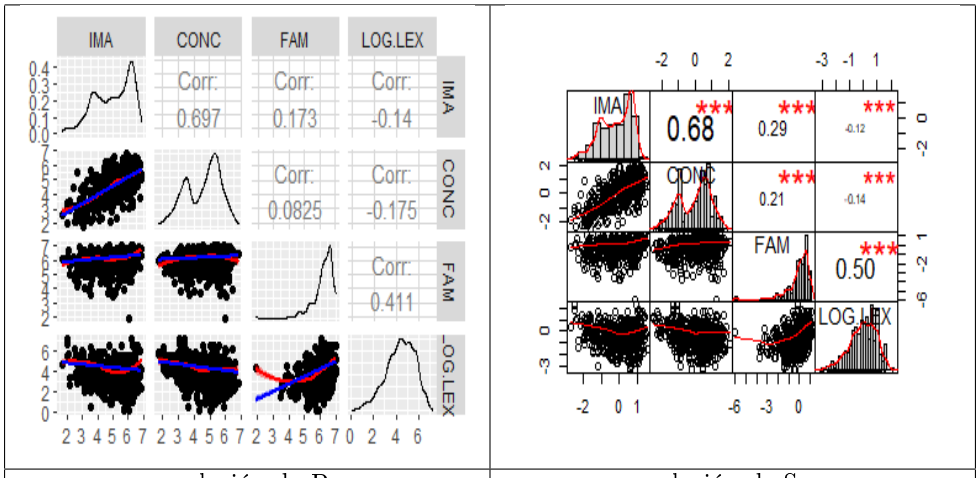


Figura 10: Boxplots de tiempos de reacción según tipo de respuesta y de hablante.

Ahora supongamos que se miden varias variables cuantitativas sobre un conjunto de individuos. Un gráfico para ver posibles asociaciones entre distintos pares de variables es el dispersograma. Sobre *1857* instancias de concordancias (de dos términos) se midieron las siguientes características del sustantivo controlador que se extrajeron de la base de datos «BuscaPalabras» (Davis & Perea, 2005): (1) Concretud (CONC): índice subjetivo en escala de 1 a 7 que indica cuán concreta es una palabra de menos (+ abstracta) a más (+ concreta); (2) Familiaridad (FAM): índice subjetivo en escala de 1 a 7, que indica cuán frecuentemente una palabra es oída, leída o producida diariamente; (3) Imaginabilidad (IMA): índice subjetivo en escala de 1 a 7 que indica la intensidad con la que una palabra evoca imágenes. Frecuencia (LEXESP): frecuencia de la palabra en el corpus «BuscaPalabras», en escala por mil. Esta última fue transformada como: $LOG.LEX = log\,(LEXESP + 1)$. Se estandarizaron todas las variables. Se muestra dispersograma en la parte inferior y las correlaciones de *Spearman* (derecha) y *Pearson* (izquierda), en la parte superior. Se observa que hay más asociación entre IMA y CONC y entre FAM y LOG.LEX. Se nota que cuanto más concreta es una palabra más imaginable resulta; y que cuanto más frecuente, más familiar. Para los dispersogramas de la izquierda se ajustó una regresión *lineal* (línea azul) y una regresión *loess* (línea roja).



con correlación de Pearson | con correlación de Spearman

Figura 11: Dispersogramas con correlograma.

4. Preprocesamiento.

Discretización. Muchas veces se hace necesario discretizar en categorías una variable continua. Mencionaremos tres métodos *no supervisados* (porque no tienen en cuenta clases para *guiar* dichos métodos). El primero es el de «intervalos iguales»: dividir el rango del atributo continuo en un número especificado de intervalos, todos del mismo tamaño. Este método puede estar afectado por *outliers* o formar intervalos con diferente frecuencia (o sea con más ítems que otros). El segundo método es el de «igual frecuencia»: poner una cantidad similar de ítems en cada intervalo. La tercera posibilidad es hacer un *clustering* y usar tantos intervalos como *clusters* encontrados. El siguiente cuadro muestra la discretización de las variables que tienen que ver con el controlador, mencionadas antes, utilizando un *clustering* por mezcla de gausianas.

Atributo	Descripción	Discretización	Casos	Ejemplos del Corpus
FAM.f	Familiaridad	$0 = [1{,}88; 5{,}74)$	348	«los sultanes» (3.46); «los archivos» (5.30)
		$1 = [5, 74; 7]$	1509	«otras deudas» (5.81); «muchas aventuras»(6)
CONC.f	Concretud	$0 = [1{,}80; 4{,}22)$	714	«estos casos» (1.88); «los lenguajes» (3.93)
		$1 = [4, 22; 6{,}82]$	1143	«las costumbres» (5.07); «los árboles» (5.87)
IMA.f	Imaginabilidad	$0 = [1{,}64; 5{,}75)$	1170	«los conocimientos» (1.64); «unas horas» (4.68)
		$1 = [5, 75; 6{,}90]$	687	«las iglesias» (5.82); «estas personas» (6.22)
LEX.f	Frecuencia: *LOG.LEX*	$0 = [1{,}16; 4{,}39)$	931	«muchas librerías» (1.93); «los cantos» (3.10)
		$1 = [4, 39; 7{,}22]$	926	«sus juegos» (5.08); «mujeres famosas» (6.20)

Cuadro 7: Discretización de atributos utilizando *clustering* por mixtura de gausianas (*mclust*).

También es preciso controlar la cantidad de ítems por categorías en una variable cualitativa, ya que si una categoría cuenta muy pocos ítems luego produce estimaciones imprecisas. Para solucionar esto es posible «fusionar» categorías para aumentar el número de casos.

Transformaciones. Para hacer comparables a las variables, por ejemplo, porque están medidas en diferentes unidades, se puede estandarizar las variables usando: $z_i = \frac{x_i - \bar{X}}{\sqrt{s^2}} = \frac{x_i - \bar{X}}{s}$. Esto funciona siempre y cuando la media y el desvío sean adecuadas como medidas de centralidad y dispersión. Si no es así se puede usar la mediana (en lugar de la media) y el rango intercuartil o el MAD (en lugar del desvío). Esta transformación cambia de escala la variable en términos de desvíos estándar a partir de la media, y hace que la variable pase a tener media *cero* y desvío *uno*. Estandarizar permite ver cuán lejos se está del centro de la distribución. Pero, ¡Cuidado! Estandarizar no significa «normalizar», es decir hacer que una variable siga una distribución normal estándar. Esto solamente sucede si la variable ya tenía distribución normal. Otra forma de detectar valores atípicos consiste en estandarizar la variable $z_i = \frac{x_i - \bar{X}}{s}$ y declarar un *outlier* si su valor absoluto es mayor a desvíos estándar: $\mid z_i \mid > 3$.

A veces se utiliza transformar los conteos como $log\,(x_i + 1)$. Se le suma una unidad porque el logaritmo de una frecuencia cero no está definido. Cambiar a la escala del logaritmo convierte una distribución en más simétrica pero solamente se puede usar para ésto si la variable posee una distribución con *asimetría a derecha*. Hacer que una variable tenga una distribución más simétrica tampoco significa «normalizar»; la distribución normal es simétrica pero muchas otras también lo son.

Otras veces se aplican transformaciones por individuos con el objeto de hacer comparables los valores de diferentes individuos. Se usa, por ejemplo, en el caso de varios «jueces» que dan una puntuación o evaluación subjetiva. Puede pasar que algunos «jueces» tiendan a dar puntuaciones muy altas o muy bajas, sesgando los datos. Se podría usar la siguiente, en la cual las puntuaciones superiores a la media resultan positivas y las inferiores a la media, negativas. Luego se divide las puntuaciones superiores por la distancia entre la media y el máximo; y a las puntuaciones inferiores, por la distancia entre la media y el mínimo.

$$T(x) = \begin{cases} \frac{x-\bar{X}}{x_{max}-\bar{X}} & si \quad x > \bar{X} \\ \frac{x-\bar{X}}{\bar{X}-x_{min}} & si \quad x < \bar{X} \end{cases}$$

Datos faltantes. Son datos «NA» («not available») que faltan, en la celda no hay ningún valor (el cero sí es un valor posible). Se han teorizado tres mecanismos que explican los datos faltantes (Enders, 2010):

1. **MAR** («Missing At Random»): Cuando la probabilidad de que ocurra el dato faltante en la variable X depende de los datos observados de las otra(s) variable(s) pero no de los datos de la variable X misma.

2. **MCAR** («Missing Completely At Random»): Cuando la probabilidad de que ocurra el dato faltante en la variable X no depende ni de los datos las otras variables ni de los datos de X. Los datos faltantes son puramente azarosos.

3. **MNAR** («Missing Not At Random»): Cuando la probabilidad de que ocurra el dato faltante en la variable X depende de los datos observados de las otra(s) variable(s) y además de los datos de la variable X misma.

Para «adivinar» el posible valor faltante se usan métodos de imputación. Generalmente, los métodos de imputación más recientes asumen que el mecanismo generador es MAR; que es menos restringido que MCAR, el cual resulta bastante implausible en la práctica. No existen tests estadísticos para distinguir entre MAR y MNAR, por lo que es necesario entender los propios datos. El paquete *mice* [*Multivariate Imputation by Chained Equations*] de *R* (Van Buuren & Groothuis-Oudshoorn, 2011) realiza imputación múltiple.

5. Nociones de probabilidades.

Un *experimento aleatorio* es aquel de cual conocemos sus posibles resultados pero no cual resultado en particular va a ocurrir. Vamos a llamar *espacio muestral* (denotado por S) al conjunto de todos los resultados posibles de un experimento aleatorio. Por ejemplo: (a) tiro un dado: $S = \{1, 2, 3, 4, 5, 6\}$; (b) mido el tiempo hasta que un niño produce la primera ocurrencia de un sustantivo: $S = \mathbb{R}_{>0}$ (números reales positivos); (c) clasifico el género de los sustantivos en español: $S = \{M, F, N\}$; (d) Si suponemos un binomio del tipo *día y noche*; *agua y aceite*, *sal y pimienta*, donde el primer elemento es p y el segundo es q, y el experimento es encontrar los posibles ordenes de ambos términos: $S = \{pp, qq, pq, qp\}$; (e) clasificamos los ordenes posibles entre sujeto, objeto y verbo en la oración simple para las lenguas del mundo: $S = \{SVO, SOV, VSO, VOS, OVS, OSV\}$. Por otro lado, un *evento* es una *colección* de posibles resultados de un experimento, incluido S mismo. Si A es un evento, entonces se puede dar que: $A \subset S$ (incluido en), $A = S$; $A = \phi$ (vacío). Por ejemplo: (a) el evento de los números naturales (incluido el cero) menores a *5*: $A = \{x : x < 5\} = \{0, 1, 2, 3, 4\}$; (b) el evento de tiradas de un dado con número par: $A = \{x : x \in par; x \in [1, 6]\} = \{2, 4, 6\}$; (c) el evento de los binomios donde p ocurre primero: $A = \{pp, pq\}$; (d) el evento donde el sujeto precede al objeto en la oración simple de las lenguas del mundo: $S = \{SVO, SOV, VSO\}$.

El *complemento de un evento* A, denotado por $\bar{A}$ o A^c es el conjunto de los elementos de S que no forman parte del evento: $\bar{\mathcal{A}} = A^C = \{x : x \notin A\}$. Por ejemplo; en (b) es $\bar{A} = \{1, 3, 5\}$; en (d) es $\bar{A} = \{VOS, OVS, OSV\}$.

Definamos la *unión de eventos* ($\cup$ = "*o*") como aquel evento que contiene a todos los elementos de ambos eventos: $A \cup B = \{x : x \in A \; o \; x \in B\}$. Por ejemplo en el ejemplo del dado, si el evento A es que sale un número par y el B es que sale un número mayor que *tres*: $A = \{2, 4, 6\}, B = \{4, 5, 6\}$, entonces $A \cup B = \{2, 4, 5, 6\}$. En el ejemplo de los órdenes de sujeto, verbo y objeto; si A es que el sujeto precede al objeto y B es que el verbo precede al objeto: $A = \{SVO, SOV, VSO\}$, $B = \{SVO, VSO, VOS\}$, entonces $A \cup B = \{SVO, VSO, SOV, VOS\}$.

Definamos la *intersección de eventos* ($\cap$ = "*y*") como aquel evento que contiene a todos los elementos comunes a ambos eventos: $A \cup B = \{x : x \in A \; y \; x \in B\}$. En el primer ejemplo del párrafo anterior $A \cap B = \{4, 6\}$ y en el segundo ejemplo, $A \cap B = \{SVO, VSO\}$. Además dos eventos son disjuntos (o excluyentes) si la intersección entre ellos está vacía: $A \cap B = \phi$. Por ejemplo, un evento y su complemento son excluyentes $A \cap \bar{A} = \phi$. Por otra parte, la unión de un evento y su complemento es igual al espacio muestral, son pues complementarios: $A \cup \bar{A} = S$.

Definamos ahora la *diferencia entre eventos* como todo lo que está en el evento A y no está en el evento B: $A - B = A \cap \bar{B}$ (análogamente: $B - A = B \cap \bar{A}$). Por otro lado, la *diferencia simétrica* entre eventos es todo lo que está en la unión de A y B pero no está en la intersección: $A \triangle B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B)$.

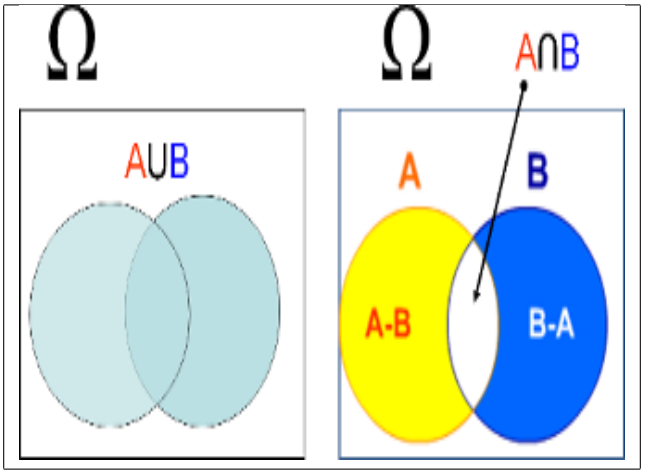


Figura 12: Diagramas de Venn

Por último veamos que para realizar una *partición* del espacio muestral S deben darse las condiciones: (a) los eventos $A_1, A_2, \ldots, A_n$ son mutuamente excluyentes; es decir que son disjuntos de a pares tal que $A_i \cap A_j = \phi$ (la intersección entre todos los pares está vacía); (b) la unión de todos los eventos es igual al espacio muestral: $\bigcup_{i=1}^{\infty} A_i = S$.

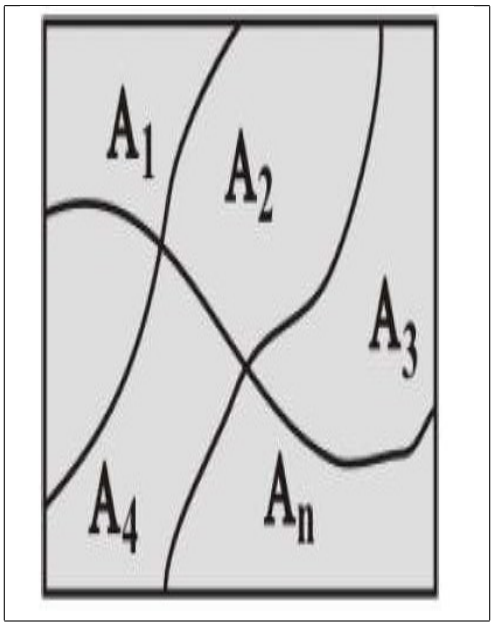


Figura 13: Partición del espacio muestral.

Sean A, B, C elementos de S. El cuadro que sigue muestra propiedades útiles:

Conmutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Asociativa	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
Distributiva	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
Leyes de De Morgan	$(A \cup B)^c = A^c \cap B^c$	$(A \cap B)^c = A^c \cup B^c$

Cuadro 8: Propiedades.

.Ahora bien, lo que queremos es calcular la probabilidad de que al realizar un experimento aleatorio ocurra el evento A, es decir: $P(A)$. Una probabilidad asocia a los eventos del espacio muestral (sub-conjuntos de S) una probabilidad entre *cero* y *uno*. Se definen los siguientes axiomas de probabilidad.

1. La probabilidad del evento A debe estar entre *cero* y *uno*: $0 \leq P(A) \leq 1$.
2. La probabilidad del espacio muestral es *uno*: $P(S) = 1$.

3. Si $A_1, A_2, \ldots, A_n$ son eventos mutuamente excluyentes ($A_i \cap A_j = \phi$) entonces la probabilidad de la unión entre los eventos es igual a la suma de las probabilidades de cada uno de los eventos: $P(A_1 \cup A_2, \ldots, \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$. Por ejemplo supongamos que la probabilidad del evento A es $P(A) = \frac{2}{20}$ y la del evento B es $P(B) = \frac{7}{20}$, entonces $P(A \cup B) = \frac{2}{20} + \frac{7}{20} = \frac{9}{20} = 0{,}45$. Pero ¡cuidado!: $P(A \cap B) = P(\phi) = 0$ porque son eventos disjuntos.

Por otra parte, la probabilidad del complemento de A es $P\left(\bar{A}\right) = 1 - P(A)$[2]. Por ejemplo: $P(A) = \frac{2}{20} = \frac{1}{10} = 0{,}1$, entonces: $P\left(\bar{A}\right) = 1 - P(A) = 1 - 0{,}1 = 0{,}9$. Otros resultados son los siguientes:

- $P\left(B \cap \bar{A}\right) = P(B) - P(A \cap B)$
- Si los eventos son *no* excluyentes: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ [ley de la suma]. Como los eventos no son excluyentes $P(A \cap B) \neq 0$, por eso hay que restarla.
- si $A \subset B \Rightarrow P(A) \leq P(B)$

La *probabilidad condicional* reduce el espacio muestral al evento condicionante. Se define como: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, notar que como $P(B)$ divide, debe darse que $P(B) \neq 0$. análogamente: $P(A) \neq 0 \Rightarrow P(B \mid A) = \frac{P(A \cap B)}{P(A)}$. Por otro lado: $P(B \mid B) = 1$ y $P(A^c \mid B) = 1 - P(A \mid B)$. Observar que al tener probabilidades estimadas como $\frac{n_i}{N}$, los «N» en en numerador y el denominador se cancelan: $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{n_1/N}{n_2/N} = \frac{n_1}{n_2}$. Se representa la probabilidad condicional en el diagrama de Venn que sigue. $P(A \mid B)$ reduce el espacio muestral a la parte pintada de verde y amarillo. Se cuentan los casos en los que se cumple B y de estos se toma la fracción en los que también se cumple A.

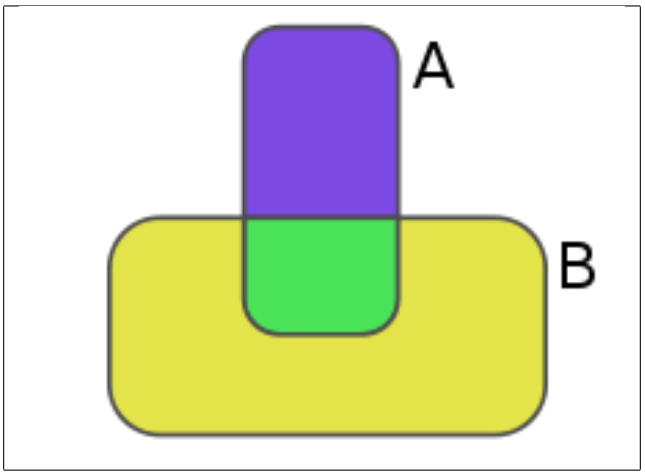


Figura 14: Probabilidad condicional.

Diacrónicamente, el objeto de una oración transitiva podía aparecer en inglés en posición post-verbal o pre-verbal. Sin embargo esto está afectado también por el «peso» del objeto, medido, por ejemplo, en cantidad de sílabas (con lo cual un pronombre sería de peso «ligero»). La siguiente tabla de contingencia muestra las probabilidades conjuntas de los objetos pre y post verbales con el hecho de que estén realizados o no como pronombres para un supuesto corpus diacrónico del inglés.

	Pronombre	No pronombre	
Objeto Preverval	0.224	0.655	0.879
Objeto Postverbal	0.014	0.107	0.121
	0.238	0.762	1

Cuadro 9: Realización del objeto y su posición en la oración

Nos preguntamos cuál es la probabilidad de que el objeto esté realizado como pronombre dado que tiene posición postverbal. Se tiene: $A = Pronombre$, $B = Postverbal$ y $P(A \cap B) = 0{,}014$, $P(B) = P(Postverbal \cap Pronombre) + P(Postverbal \cap No\ pronombre) = 0{,}014 + 0{,}107 = 0{,}121$. Entonces: $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(Pronombre \mid Postverbal) = \frac{0{,}014}{0{,}121} \approx 0{,}11$. En cambio, en posición preverbal sería: $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(Pronombre \mid Preverbal) = \frac{0{,}224}{0{,}224 + 0{,}655} = \frac{0{,}224}{0{,}879} \approx 0{,}25$.

[2] En efecto: $P(S) = P\left(A \cup \bar{A}\right) = P(A) + P\left(\bar{A}\right) = 1 \Rightarrow P\left(\bar{A}\right) = 1 - P(A)$.

Definamos ahora la regla del producto, la cual nos sirve para calcular probabilidades conjuntas, como sigue.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A \mid B)\, P(B)$$

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(B \mid A)\, P(A)$$

Pero por la propiedad conmutativa: $P(A \cap B) = P(B \cap A)$, por ende: $P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$.
Veamos ahora la regla de la cadena. Sean A, B, C eventos del espacio muestral S; $P(A) > 0$ y $P(A \cap B) > 0$; entonces:

$$P(A \cap B \cap C) = P(A)\, P(B \mid A)\, P(C \mid A \cap B)$$

Observemos ahora lo siguiente:

$$P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

$$\Rightarrow P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$\Rightarrow P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}$$

Esta es la regla de Bayes, su generalización es útil para calcular probabilidades condicionales en una partición del espacio muestral, como muestra la siguiente Figura (caso $P(A \mid B)$).

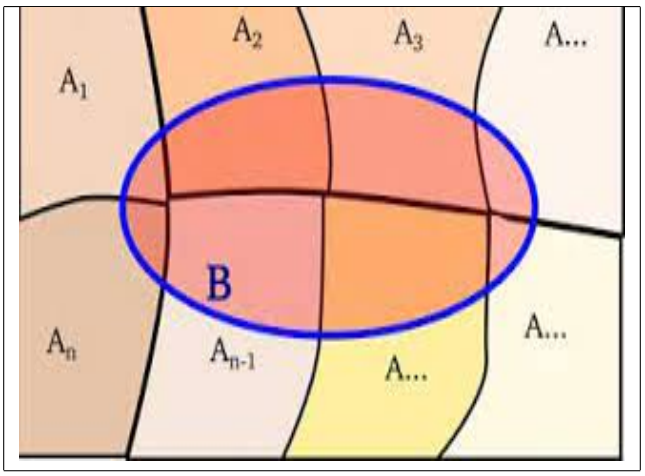


Figura 15: Regla de Bayes.

Sean $A_1, A_2, A_3, \ldots$ una partición de S para $i = 1, 2, 3, \ldots$ y $P(B) > 0$, entonces:

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{P(B)} = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{i=1}^{\infty} P(B \mid A_i)\, P(A_i)}$$

Para ver la igualdad del denominador aplicamos la regla de la suma para eventos excluyentes y luego la del producto:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + ... = P(B \mid A_1)\, P(A_1) + P(B \mid A_2)\, P(A_2) + ... = \sum_{i=1}^{\infty} P(B \mid A_i)\, P(A_i)$$

Notar que ahora no es necesario conocer $P(B)$.
Veamos un ejemplo. Consideremos el orden del objeto directo realizado como un SN animado o inanimado. Supongamos las siguientes probabilidades:

(i) probabilidad de objeto SN animado $P(Anim) = 0{,}4$;

(ii) probabilidad de objeto SN pos-verbal dado que el objeto es animado: $P(PostV \mid Anim) = 0{,}7$;

(iii) probabilidad de objeto SN post-verbal dado que el objeto es inanimado: $P(PostV \mid Inanim) = 0{,}8$.

Notar que $P(Inanim) = P(Anim^c) = 1 - P(Anim) = 1 - 0{,}4 = 0{,}6$.

Se desea calcular la probabilidad de que un objeto sea animado dado que se encuentra después del verbo, o sea: $A_1 = Anim$; $A_2 = Inanim$; $B = PostV$. Aplicando la regla de Bayes:

$$P(A_1 \mid B) = \frac{P(B \mid A_1)\,P(A_1)}{P(B)} = \frac{P(B \mid A_1)\,P(A_1)}{\sum_{i=1}^{2} P(B \mid A_i)\,P(A_i)} = \frac{P(B \mid A_1)\,P(A_1)}{P(B \mid A_1)\,P(A_1) + P(B \mid A_2)\,P(A_2)}$$

$$\Leftrightarrow P(Anim \mid PostV) = \frac{P(PostV \mid Anim)\,P(Anim)}{P(PostV)} = \frac{P(PostV \mid Anim)\,P(Anim)}{P(PostV \mid Anim)\,P(Anim) + P(PostV \mid Inanim)\,P(Inanim)}$$

Las probabilidades del numerador ya las tenemos. Respecto del denominador:

$$\begin{aligned} P(PostV) = &\quad P(PostV \mid Anim)\,P(Anim) + P(PostV \mid Inanim)\,P(Inanim) \\ P(PostV) = &\quad (0{,}7 \times 0{,}4) + (0{,}8 \times 0{,}6) \end{aligned}$$

Por ende:

$$P(Anim \mid PostV) = \frac{P(PostV \mid Anim)\,P(Anim)}{P(PostV)} = \frac{0{,}7 \times 0{,}4}{(0{,}7 \times 0{,}4) + (0{,}8 \times 0{,}6)} = \frac{0{,}28}{0{,}76} \approx 0{,}3684$$

Por último, introduzcamos la noción de independencia. Dos eventos A y B son independientes si habiendo sucedido B, no cambia la probabilidad de A. Con lo cual: $P(A \mid B) = P(A)$ y $P(B \mid A) = P(B)$. Además, bajo independencia, la probabilidad conjunta es igual al producto de las probabilidades de dichos eventos: $P(A \cap B) = P(A)\,P(B)$. Por ejemplo si $A = correr$ y $B = loco$ (Ej.: «correr como un loco»), con: $P(correr) = 10^{-3}$; $P(loco) = 10^{-6}$ y $P(correr, loco) = 10^{-7}$. Se ve que las palabras *correr y loco* no son independientes porque $10^{-3} \times 10^{-6} = 10^{-9}$ que es diferente de la probabilidad conjunta 10^{-7}.

Por otra parte, si dos eventos A y B son independientes también lo son: (i) A y B^c; (ii) A^cy B; (iii) A^cy B^c. Asimismo, si se tuvieran tres eventos o más, para que haya independencia es necesario demostrar que hay independencia entre todos los pares posibles de eventos, todas las ternas, etc. Por ejemplo los eventos A, B, C son mutuamente independientes si y solo si lo son: A y B; A y C; B y C; A, B, C. Además dos eventos A y B son independientes condicionados a otra información C si se da que : $P(A \cap B \mid C) = P(A \mid C)\,P(B \mid C)$.

6. Algunas aplicaciones (optativo).

Modelos de lenguaje. Las probabilidades condicionales se usan en los modelos de «n-gramas». Dicho modelo se define como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(w_n \mid w_1, w_2, \ldots, w_{n-1}) = \frac{P(w_1, w_2, \ldots, w_{n-1}, w_n)}{P(w_1, w_2, \ldots, w_{n-1})}$$

Sea $C(w_1, \ldots, w_n)$ la frecuencia de un «n-grama». Estimamos los modelos como: $P_{MLE}(w_1, \ldots, w_n) = \frac{C(w_1,\ldots,w_n)}{N}$ y $P_{MLE}(w_n \mid w_1, \ldots, w_{n-1}) = \frac{C(w_1,\ldots,w_n)}{C(w_1,\ldots,w_{n-1})}$. El siguiente cuadro muestra las frecuencias absolutas de mono-grama (una palabra) y bi-gramas (dos palabras) para la oración: *(person) she was inferior to both sisters* en un corpus.

w_1	$C(w_1)$	w_1w_2	$C(w_1w_2)$
person	223	person she	2
she	6917	she was	843
was	9409	was inferior	0
inferior	33	inferior to	7
to	20042	to both	9
both	317	both sisters	2

Cuadro 10: Frecuencias de w_1y $w_1w_2$para *She was inferior to both sisters*

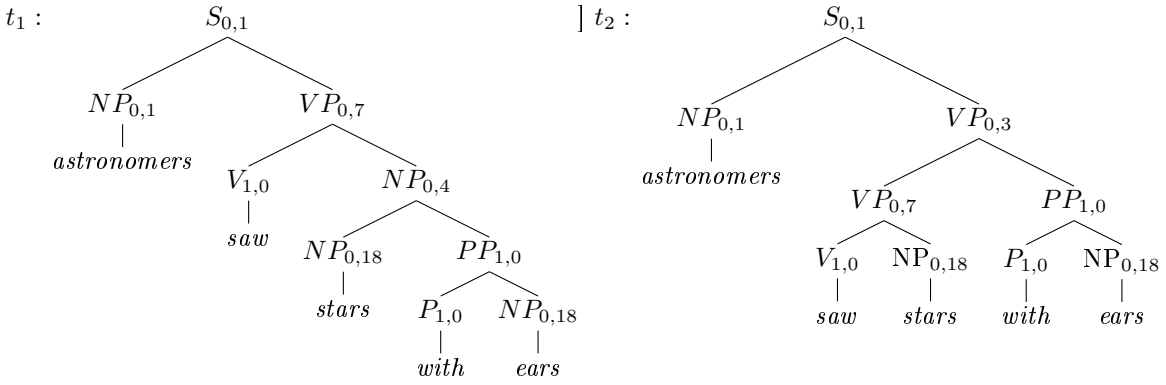
Supongamos que queremos calcular el modelo $P(she \mid person) = \frac{C(person,she)}{C(person)} = \frac{2}{223} \approx 0{,}009$. Por otro lado: $P(to \mid inferior) = \frac{C(inferior,to)}{inferior} = \frac{7}{33} \approx 0{,}212$. Pero: $P(inferior \mid was) = \frac{C(was,inferior)}{was} = \frac{0}{9409} = 0$ porque el bigrama «was inferior» no fue visto en el corpus.

Probabilistic Context Free Grammars. Otra posible aplicación son las gramáticas probabilísticas libres de contexto («Probabilistic Context Free Grammars» [PCFG]). Una PCFG consta de: (i) un conjunto de terminales de vocabulario: $\{w^k\}$ con $k = 1, \ldots, V$ términos; (ii) una conjunto de nodos no terminales: $\{N^i\}$, $i = 1, \ldots, n$, donde N^1 es el símbolo de comienzo; (iii) un conjunto de reglas $\{N^i \to \xi^j\}$, donde ξ^j es una secuencia de terminales y no terminales; (iv) un conjunto de probabilidades para cada regla, tal que todas las expansiones de una regla sumen uno: $\forall i \sum_j P\left(N^i \to \xi^j\right) = 1$.

Un importante supuesto es que la probabilidad de un sub-árbol no depende de las palabras no dominadas por dicho sub-árbol ni de los nodos que se encuentran fuera de este. O sea que hay un supuesto de independencia entre las reglas. Sea entonces la siguiente PCFG.

$N^i \to \xi^j$	$P\left(N^i \to \xi^j\right)$	$N^i \to \xi^j$	$P\left(N^i \to \xi^j\right)$
$S \to NP\ VP$	1,0	$NP \to NP\ PP$	0,4
$PP \to P\ NP$	1,0	$NP \to astronomers$	0,1
$VP \to V\ NP$	0,7	$NP \to ears$	0,18
$VP \to VP\ PP$	0,3	$NP \to saw$	0,04
$P \to with$	1,0	$NP \to stars$	0,18
$V \to saw$	1,0	$NP \to telescopes$	0,1

Cuadro 11: PCFG simple

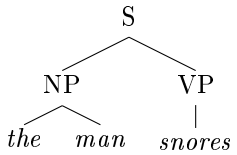


$P(t_1) = 1{,}0 \times 0{,}1 \times 0{,}7 \times 1{,}0 \times 0{,}4 \times 0{,}18 \times 1{,}0 \times 1{,}0 \times 0{,}18 = 0{,}0009072$

$P(t_2) = 1{,}0 \times 0{,}1 \times 0{,}3 \times 0{,}7 \times 1{,}0 \times 0{,}18 \times 1{,}0 \times 1{,}0 \times 0{,}18 = 0{,}0006804$

$P(w_{15}) = P(t_1) + P(t_2) = 0{,}0009072 + 0{,}0006804 = 0{,}0015876$

Los árboles muestran dos posibles *parseos* [parcing] de la secuencia *astronomers saw stars with* ears. La probabilidad de cada *parseo* es simplemente el producto de las probabilidades de todas las reglas. La probabilidad de la oración es la suma de las probabilidades de ambos *parseos*. ¿Por qué es lícito multiplicar las probabilidades? Tomemos el ejemplo más simple *the man snores*, con el parseo que sigue.



Queremos sacar la probabilidad conjunta: $P(S \to NP\ VP \cap NP \to the\ man \cap VP \to snores) = P(A \cap B \cap C)$. Aplicamos entonces la regla de la cadena:

$P(A \cap B \cap C) = P(A)\, P(B \mid A)\, P(C \mid A \cap B) \Rightarrow$

$P(S \to NP\ VP)\, P(NP \to the\ man \mid S \to NP\ VP)\, P(VP \to snores \mid S \to NP\ VP \cap NP \to the\ man)$

Luego aplicamos la suposición de independencia entre las reglas, con lo cual: $P(A \mid B) = P(A)$ y $P(C \mid A, B) = P(C)$; y llegamos al producto de las probabilidades de las reglas:

$P(S \to NP\ VP)\, P(NP \to the\ man)\, P(VP \to snore)$.

Surprisal. La información que conlleva cualquier unidad lingüística se puede medir mediante la métrica de «surprisal» o «sorpresa» [Hale, 2001; Levy, 2008]: $s(x) = log_2 \frac{1}{P(x|contexto)} = -log_2\left(P\left(x \mid contexto\right)\right)$.[3] Cuando x tenga

[3] Recordar la regla del logaritmo que dice que: $log_a\left(\frac{1}{x}\right) = log_a\left(x^{-1}\right) = -log_a(x)$;por ejemplo: $log_2\left(\frac{1}{3}\right) = -log_2(3)$. Otra regla útil es: $log_a\left(\frac{b}{c}\right) = log_a(b) - log_a(c)$

una probabilidad condicional alta, la «sorpresa» será baja; y a la inversa, cuando la probabilidad sea baja, la «sorpresa» será alta. Hale (2001) propuso que el «esfuerzo cognitivo» que se necesita para procesar una palabra es proporcional a su «surprisal»; es decir que las palabras más predecibles por el contexto serán más fáciles de procesar. Hay varias formas de computar dicha métrica. Una de ellas es haciendo uso de modelos de lenguaje a n-gramas: $s(w_{k+1}) = -log_2 P(w_{k+1} \mid w_{k-2}, w_{k-1}, w_k)$. Usando un modelo de bigramas se puede calcular la medida a partir de un corpus. Por ejemplo para la concordancia en *autos negros*, se tendrá que la «sorpresa» de ver el plural *negros* luego de haber visto el plural *autos* será $s(negros) = -log_2 P(negros \mid autos)$. Asimismo es posible definir la métrica en términos de probabilidades derivadas de una gramática de estructura de frase libre de contexto probabilística (PCFG):

$$s(w_{k+1}) = -log_2 P(w_{k+1} \mid w_1, \ldots, w_k) = -log_2 \frac{P(w_1,\ldots,w_k,w_{k+1})}{P(w_1,\ldots,w_k)} =$$
$$= -\{log_2 P(w_1, \ldots, w_k, w_{k+1}) - log_2 P(w_1, \ldots, w_k)\} = log_2 P(w_1, \ldots, w_k) - log_2 P(w_1, \ldots, w_k, w_{k+1}) =$$
$$log_2 \sum_T P(T, w_1, \ldots, w_k) - log_2 \sum_T P(T, w_1, \ldots, w_k, w_{k+1})$$
$$\Rightarrow s(w_{k+1}) = pp_{w_k} - pp_{k+1}$$

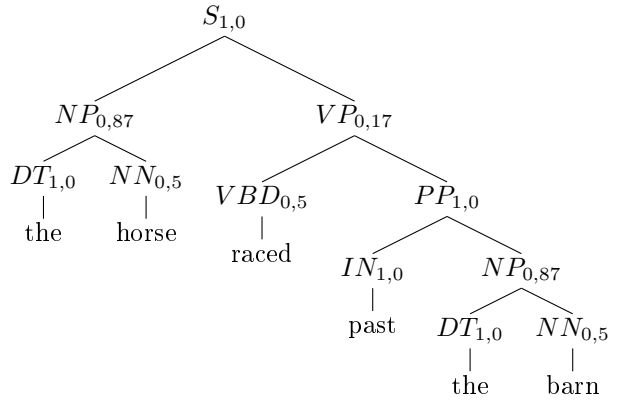
En este caso es necesario sumar las probabilidades de las estructuras sintácticas que llevan a la linearización $w_1, \cdots, w_k$ y también aquellas de las estructuras que llevan a $w_1, \cdots, w_{k+1}$. La «sorpresa» de w_{k+1} es la diferencia de sus logaritmos.

$N^i \to \xi^j$	$P\left(N^i \to \xi^j\right)$	$N^i \to \xi^j$	$P\left(N^i \to \xi^j\right)$
$S \to NP\ VP$	1,0	$DT \to the$	1,0
$NP \to DT\ NN$	0,87	$NN \to horse$	0,5
$NP \to NP\ VP$	0,13	$NN \to barn$	0,5
$PP \to IN\ NP$	1,0	$VBD \to fell$	0,5
$VP \to VBD\ PP$	0,17	$VBD \to raced$	0,5
$VP \to VBN\ PP$	0,75	$VBN \to raced$	1,0
$VP \to VBD$	0,08	$IN \to past$	1,0

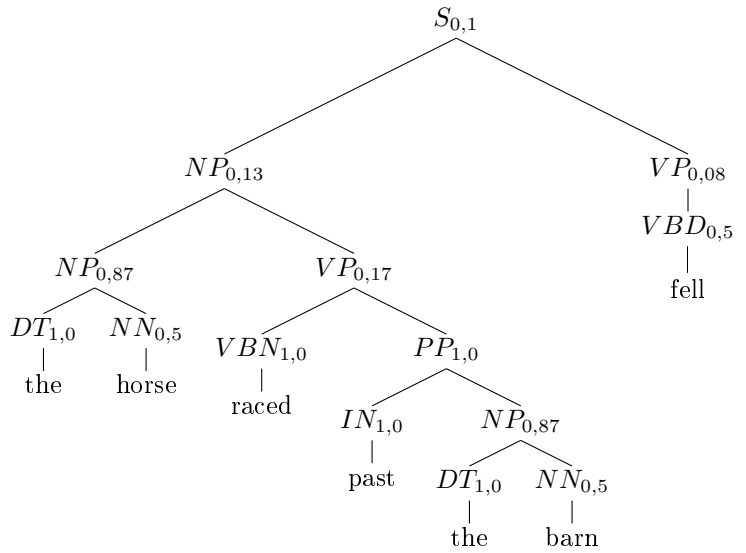
Cuadro 12: PCFG para *the horse raced past the barn fell*

Veamos el fenómeno de *garden-path* en comprensión de oraciones. El ejemplo típico es: *the horse raced past the barn fell*. Leyendo hasta la palabra *barn* los sujetos tienden a interpretar que se trata de una oración cuyo verbo es *raced* [ver t_1]. Sin embargo, al leer la palabra siguiente *fell*, se opta por una interpretación de oración reducida (*the horse (that was) raced past the barn*), que es el sujeto de la oración y cuyo verbo es *fell* [ver t_2]. Al leer *fell* debería haber más esfuerzo cognitivo; o sea, se debería notar un salto de «surprisal».

t_1 :



t_2 :



Se calcula del siguiente modo:

$P(w_1, \ldots, w_k) = 1{,}0 \times 0{,}87 \times 1{,}0 \times 0{,}5 \times 0{,}17 \times 0{,}5 \times 1{,}0 \times 1{,}0 \times 0{,}87 \times 1{,}0 \times 0{,}5 = 0{,}01608413$

$pp_{w_k} = log_2 P(w_1, \ldots, w_k) = log_2(0{,}01608413) = -5{,}958219$

$P(w_1, \ldots, w_{k+1}) = 1{,}0 \times 0{,}13 \times 0{,}87 \times 1{,}0 \times 0{,}5 \times 0{,}17 \times 1{,}0 \times 1{,}0 \times 1{,}0 \times 0{,}87 \times 1{,}0 \times 0{,}5 \times 0{,}08 \times 0{,}5 = 0{,}0001672749$

$pp_{w_{k+1}} = log_2 P(w_1, \ldots, w_{k+1}) = log_2(0{,}0001672749) = -12{,}54549$

$s(w_{k+1}) = pp_{w_k} - pp_{k+1} = -5{,}958219 - (-12{,}54549) = 6{,}587271$

Notar que el «surprisal» hasta la palabra *barn* es: $s(w_{k+1}) = pp_{w_k} - pp_{k+1} = -4{,}757306 - (-5{,}958219) = 1{,}200913$. O sea el «surprisal» aumenta más de cinco veces cuando se produce el *garden-path*.

7. Variables aleatorias y funciones de distribución.

Variable aleatoria [v. a.]. Una variable aleatoria [v. a.] X es una función que conecta a los elementos s del espacio muestral S a los números reales ($\mathbb{R}$). Es decir que: $s \to X(s) = x$. Los valores de $\mathbb{R}$ que se asignan $(x_1, x_2, \ldots)$ definen el rango o conjunto de valores posibles de la variable aleatoria X. Por ejemplo supongamos nuevamente el ejemplo de los binomios posibles de dos sustantivos cualesquiera, por ejemplo, $\{día, noche\}$. El espacio muestral era el conjunto de los posibles órdenes de ambos términos $S = \{pp, pq, qp, qq\}$. Supongamos que la variable aleatoria X es «cantidad de veces que aparece el término p» entonces, por ejemplo, la función X asignará el número 2 a *pp*. o sea, $X(pp) = 2$. La probabilidad será una función de los valores del rango de X al intervalo $[0, 1]$. Por ejemplo, la $P(X = 2) = \frac{1}{4}$ porque *pp* ocurre una vez en los cuatro tipos de órdenes. Los otros son:

s	$X(s) = x$	$P(X = x)$
pp	2	$\frac{1}{4}$
pq	1	$\frac{2}{4} = \frac{1}{2}$
qp	1	
qq	0	$\frac{1}{4}$

Cuadro 13: La v.a. X es la «cantidad de veces que aparece el término p»

Supongamos ahora que el espacio muestral esta constituido por todos las composiciones escritas de los alumnos de nivel intermedio de español como segunda lengua en las escuelas públicas italianas. La variable aleatoria es «cantidad de errores de concordancia». Entonces, la variable aleatoria es una función que recibe como input a la composición *n-ésima* y devuelve como output la cantidad de errores cometidos en cada composición. En principio los posibles valores de X son los números naturales mayores o iguales a *cero*. Denotamos con mayúscula (X) a la variable aleatoria y con minúscula (x) a los posibles valores que puede recibir.

Función de distribución acumulada (f.d.a.). A cada variable aleatoria [v. a.] X se le asocia una función de distribución acumulada: $F_X(x) = P(X \leq x), \forall x$. Dicha función indica la probabilidad de obtener un valor de la

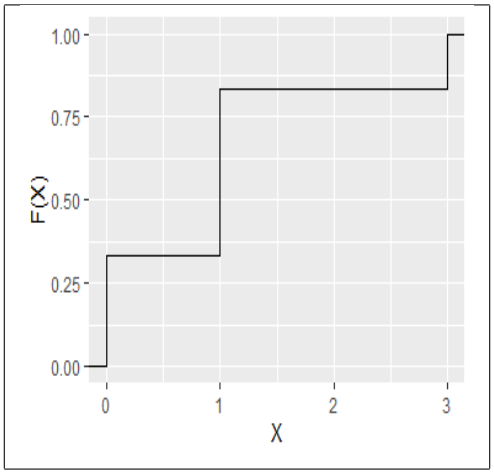


Figura 16: Gráfico de la función de distribución acumulada.

variable aleatoria menor o igual a x. Si $F_X(x)$ es una f.d.a. vale lo siguiente: (1) cuando $x \to -\infty$, $F_X(x) = 0$ y cuando $x \to \infty$, $F_X(x) = 1$; (2) $F_X(x)$ es una función creciente de X; (3) $F_X(x)$ es una función continua a derecha. Si una v.a. X es *continua* si su $F_X(x)$ es continua para todo x. En cambio, una v.a. es *discreta* si su $F_X(x)$ es escalonada de x. Por otro lado dos v.a. X e Y son idénticamente distribuidas ($X \sim Y$) si sus distribuciones acumuladas son iguales: $F_X(x) = F_Y(y)$. Sea la v. aleatoria: $X = \{0,0,1,1,1,3\}$, con $Rg(X) = [0,3]$,$n = 6$. Observar que el primer punto de cada intervalo esta incluido porque la distribución debe ser continua a derecha.

X	$f_X(x) = P(X = x)$	$F_X(x) = P(X \leq x)$
0	$\frac{2}{6} = \frac{1}{3}$	$\frac{2}{6} = \frac{1}{3}$
1	$\frac{3}{6} = \frac{1}{2}$	$\frac{2}{6} + \frac{3}{6} = \frac{5}{6}$
2	0	$\frac{5}{6} + 0 = \frac{5}{6}$
3	$\frac{1}{6}$	$\frac{5}{6} + \frac{1}{6} = \frac{6}{6} = 1$

Cuadro 14: Función de distribución acumulada.

$$F_X(x) \begin{cases} 0 & x < 0 \\ \frac{1}{3} & 0 \leq x < 1 \\ \frac{5}{6} & 1 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Funciones de distribución de masa (f.d.m.) y de densidad (f.d.p.). Si la v. a. es discreta tiene una función $f_X(x)$ de probabilidad *de masa* (f.p.m.) que asocia probabilidad a puntos. Si la v. a. es continua tiene una función $f_X(x)$ *de densidad* de probabilidad (f.d.p.) que asocia probabilidad a intervalos. La Figura 12 muestra una f.m.p. y la Figura 13 es una f.d.p. correspondiente a $X \sim N(0,1)$, junto a su distribución acumulada.

X	$f_X(x) = P(X = x)$	$F_X(x) = P(X \leq x)$
1	0,3	0,3
2	0,2	0,5
3	0,05	0,55
4	0,4	0,95
5	0,05	1

Cuadro 15: Función de distribución de probabilidad de masa.

$$F_X(x)\begin{cases}0 & x<1\\ 0{,}3 & 1\le x<2\\ 0{,}5 & 2\le x<3\\ 0{,}55 & 3\le x<4\\ 0{,}95 & 4\le x<5\\ 1 & x\ge 5\end{cases}$$

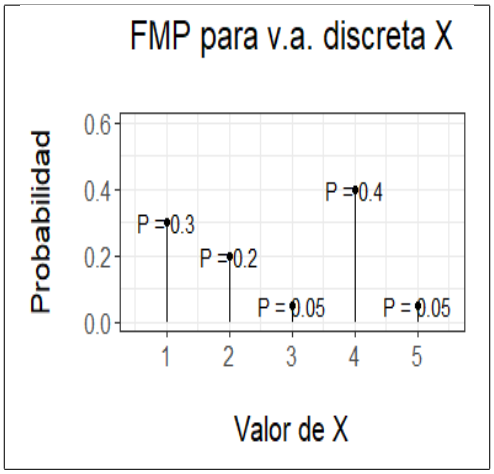


Figura 17: Gráfico de la función de masa de probabilidad.

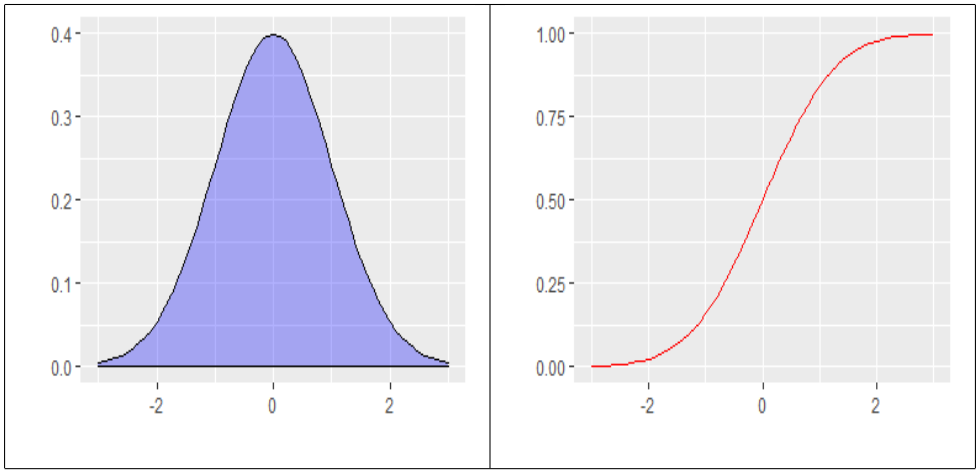


Figura 18: Gráfico de la función de densidad de probabilidad para una normal estándar y su correspondiente función acumulada.

Si una función $f_X(x)$ es una f.d.p. (v.a. continua) o una f.p.m. (v. a. discreta) entonces se verifica que: (1) sus valores son mayores o iguales a cero: $f_X(x)\ge 0, \forall x$; (2) sus valores suman *uno*: $\sum_x f_X(x)=1$ (caso discreto) y $\int_{-\infty}^{\infty} f_X(x)\,dx=1$ (caso continuo: el *área* bajo la curva es *uno*). El *soporte de una distribución* son los valores de la v.a. X donde la función de densidad o de masa es *positiva* ($f_x(X)>0$).

Las correspondientes distribuciones acumuladas se definen como:

$$X\begin{cases}discreta: & F_X(x)=P(X\le x)=\sum_{i\le x} f_X(i)\\ continua: & F_X(x)=P(X\le x)=\int_{-\infty}^{x} f_X(t)\,dt\end{cases}$$

¿Cuanto vale la probabilidad en un punto x de una v. a.? En el caso de una variable discreta es el «salto» en la función de densidad acumulada. Si el rango de una v.a. discreta X son los valores $x_1 < x_2 < \ldots < x_n$ entonces la probabilidad en $X = x_1$ es $f_X(x_1) = F_X(x_1)$ (el valor de la acumulada en $X = x_1$) y para el resto de los valores del rango ($X = x_i$) es: $f_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ (o sea la resta entre los valores sucesivos de probabilidad acumulada). Por ejemplo, en el Cuadro 14:

- $f_X(x_1 = 0) = \frac{2}{6} = \frac{1}{3}$
- $f_X(x_2 = 1) = F_X(x_2) - F_X(x_1) = \frac{5}{6} - \frac{2}{6} = \frac{3}{6} = \frac{1}{2}$;
- $f_x(x_3 = 2) = F_X(x_3) - F_X(x_2) = \frac{5}{6} - \frac{5}{6} = 0$;
- $f_X(x_4 = 3) = F_X(x_4) - F_X(x_3) = \frac{6}{6} - \frac{5}{6} = \frac{1}{6}$.

Por supuesto, se trata de los valores de la columna dos del cuadro.

Ahora bien, en el caso continuo la probabilidad es *un área* bajo la curva de densidad. Dicha área se calcula por medio de una integral definida. La cuestión es que una integral definida desde un punto x hasta el mismo punto es *cero*: $\int_a^a x\,dx = 0$. Por lo tanto, si la variable es continua: $P(X = x_i) = 0$. Entonces, por ejemplo, $P(X = 1) = 0$. Pero al ser un área, siempre debe ser positiva. La probabilidad se saca calculando el área debajo de la curva de densidad desde $-\infty$ hasta el punto x; o sea mediante la distribución acumulada: $F_X(X) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$.

¿Cuánto vale la probabilidad en un intervalo de valores de una v. a.? En el caso discreto basta sumar las probabilidades incluidas la de los extremos del intervalo. Si se usa la distribución acumulada, se resta la probabilidad acumulada en el punto a de aquella acumulada del punto b, pero es necesario sumar la probabilidad puntual el a (porque se trata de un intervalo cerrado a izquierda ([)) y $F_X(a)$ incluye a la probabilidad en a pero, al ser restada, se la debe incluir otra vez).

$$P(a \leq X \leq b) = \sum_{i=a}^{b} f_X(i)$$

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

En donde i indica el i-ésimo nivel de la variable con m niveles. Por ejemplo, para el Cuadro 15: $i = \{1, 2, 3, 4, 5\}$. Entonces:

- $P(2 \leq X \leq 4) = f_X(2) + f_X(3) + f_X(4) = 0{,}2 + 0{,}05 + 0{,}4 = 0{,}65$.
- $P(2 \leq X \leq 4) = F_X(4) - F_X(2) + f_X(2) = 0{,}95 - 0{,}5 + 0{,}2 = 0{,}65$.

En el caso continuo, es necesario tener en cuenta que todos los casos siguientes son equivalentes:

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

.Dado que, para una v.a. continua, la probabilidad en un punto es *cero*, los extremos de los intervalos no aportan. En este caso restamos la probabilidad acumulada de a de la probabilidad acumulada de b. Es decir:

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x)\ dx - \int_{-\infty}^{a} f_X(x)\ dx = \int_{a}^{b} f_X(x)\ dx; a \leq b$$

¿Cuánto vale $P(X \geq a)$ en el caso continuo? Pues en este caso simplemente calculamos la acumulada hasta a y la restamos de 1, ya que una distribución continua tiene área unitaria. Así: $P(X \geq a) = 1 - F_X(a)$.

Por último mencionemos cómo se puede pasar de una función acumulada a una función de densidad y viceversa. Pues bien, la derivada de la función acumulada nos da la función de densidad: $\frac{d}{dX} F_X(x) = f_X(x)$. Por otro lado, la integral de la función de densidad es la función acumulada: $F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt$.

El siguiente cuadro resume los procedimientos para calcular probabilidades, con ejemplos para la variable discreta del cuadro 15.

Discreta	$P(X=x_1)=F_X(x_1)=f_X(x_1)$	$P(1)=F_X(1)=f_X(1)=0{,}3$
	$P\left(X=x_{i>1}\right)=F_X(x_i)-F_X\left(x_{i-1}\right)=f_X\left(x_{i>1}\right)$	$P(4)=F_X(4)-F_X(3)=0{,}95-0{,}55=0{,}4=f_X(4)$
	$P(a\leq X\leq b)=F_X(b)-F_X(a)+f_X(a)=\sum_{i=a}^{b}f_X(i)$	$P(2\leq X\leq 5)=F_X(5)-F_X(2)+f_X(2)=1-0{,}5+0{,}2=0{,}7$ $P(2\leq X\leq 5)=\sum_{i=2}^{5}f_X(i)=0{,}2+0{,}05+0{,}4+0{,}05=0{,}7$
	$P(X\leq a)=F_X(X\leq a)=\sum_{i=0}^{a}f_X(i)$	$P(X\leq 3)=F_X(X\leq 3)=\sum_{i=0}^{3}f_X(i)=0{,}3+0{,}2+0{,}05=0{,}55$
	$P(X<a)=P(X\leq a-1)=F_X(X\leq a-1)=\sum_{i=0}^{a-1}f_X(i)$	$P(X<3)=P(X\leq 2)=F_X(X\leq 2)=\sum_{i=0}^{2}f_X(i)=0{,}3+0{,}2=0{,}5$
	$P(X\geq a)=1-F_X(X\leq a-1)=1-\sum_{i=0}^{a-1}f_X(i)$ $P(X\geq a)=\sum_{i=a}^{m}f_X(i)$ (**X tiene m niveles**)	$P(X\geq 3)=1-F_X(X\leq 2)=1-\sum_{i=0}^{2}f_X(i)=1-(0{,}3+0{,}2)=0{,}5$ $P(X\geq 3)=\sum_{i=3}^{5}f_X(i)=0{,}05+0{,}4+0{,}05=0{,}5$
	$P(X>a)=P(X\geq a+1)=1-F_X(X\leq a)=1-\sum_{i=0}^{a}f_X(i)$ $P(X>a)=P(X\geq a+1)=\sum_{i=a+1}^{m}f_X(i)$ (**X tiene m niveles**)	$P(X>3)=P(X\geq 4)=1-F_X(X\leq 3)=1-0{,}55=0{,}45$ $P(X>3)=P(X\geq 4)=1-\sum_{i=0}^{3}f_X(x_i)=1-(0{,}3+0{,}2+0{,}05)=0{,}45$ $P(X>3)=P(X\geq 4)=\sum_{i=4}^{5}f_X(i)=0{,}4+0{,}05=0{,}45$
Continua	$P(X\leq a)$	$F_X(a)$
	$P(X\geq a)$	$1-F_X(a)$
	$P(a\leq X\leq b)$	$F_X(b)-F_X(a)$

8. Algunas familias de distribuciones de probabilidad usuales.

Las distribuciones son modelos ideales de poblaciones. Pero más que tratar con una sola distribución, tratamos con una familia de distribuciones. Dicha familia se caracteriza por uno o más parámetros, lo cual permite cambiar ciertas características de la distribución manteniendo la misma forma funcional. Por ejemplo, podemos elegir a la distribución normal para modelar cierta población pero no conocer con precisión la media poblacional. En este caso estaríamos frente a una familia de distribuciones normales con media μ, un parámetro no especificado con rango $-\infty<\mu<\infty$. Diremos que la variable aleatoria X tiene una f.p.m, f.d.p., f.d.a. perteneciente a una determinada familia de distribuciones indexada con el vector de parámetros $\theta=(\theta_1,\theta_2,\ldots,\theta_n)$: $X\sim f(x,\theta)$ [f.p.m. o f.d.p.]; $X\sim F(x,\theta)$ [f.d.a.]; con $\theta\in\Theta$. En probabilidades estamos interesados en calcular la probabilidad de un valor determinado de una variable aleatoria dado que conocemos la distribución de la cual procede.

8.1. Familias de distribuciones discretas.

8.1.1. Bernoulli y Binomial.

Un experimento de *Bernoulli* es aquel que tiene dos posibles resultados. Una v.a. X tiene distribución $Be(p)$ si: $X=\begin{cases}1 & \textit{con probabilidad } p\\ 0 & \textit{con probabilidad } 1-p\end{cases}$; a veces $X=1$ es denominado como «éxito» y $X=0$, como «fracaso». Muchos experimentos pueden modelarse de esta manera, por ejemplo $X=1$ puede significar que el n-esimo alumno dice correctamente una concordancia y $X=0$ que la dice incorrectamente; o bien que una persona pronuncia el fonema */s/* en posición final de una palabra o que no lo pronuncia (hay aspiración); etc. La esperanza de la distribución es $E[X]=p$ y la varianza es $V[X]=p(1-p)$.

$$f_X(x;p)=p^x(1-p)^{1-x};X=\{0,1\}$$

Ahora supongamos el evento de que en el i-esimo ensayo se obtiene $X=1$; o sea: $A_i=\{X=1\ \textit{en el } i-\textit{ésimo ensayo}\},i=1,2,\ldots,n$. Y definamos la nueva v.a. $X=$ *número total de éxitos en n ensayos*. El evento $X=x$ ocurre si de todos los eventos $A_1,\ldots,A_n$ suceden exactamente x de estos y $n-x$ no ocurren. Por ejemplo, a cada sujeto de un experimento se le presentan $n=10$ sustantivos y la tarea consiste en determinar si es o no un animal. Aquí $A_i=\{X=1\}$ significa que en la presentación de la i-ésima palabra ha respondido que se trata de un animal. Entonces, la variable $X=6$ indica que ha respondido que es animal seis veces de las diez palabras presentadas (y $n-x=10-6=4$ respondió que no es animal). Genéricamente una posible sucesión de ocurrencias (A_i) o no ocurrencias (A_i^c) del evento en n ensayos Bernoulli sería: $A_1\cap A_2\cap A_3^c\cap\ldots\cap A_{n-1}\cap A_n^c$. ¿Cuál es la probabilidad de esta secuencia? Como los ensayos se suponen independientes entre sí, tenemos que la conjunta es el producto de probabilidades que sigue la distribución Bernoulli. Entonces:

$$P(A_1\cap A_2\cap A_3^c\cap\ldots\cap A_{n-1}\cap A_n^c)=p\times p\times(1-p)\times\ldots\times p\times(1-p)=p^x(1-p)^{n-x}$$

Pero debemos tener en cuenta que hay $\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}$ posible permutaciones posibles (no ordenadas) de dicha secuencia[4]. Entonces llegamos a la f.p.m binomial con parámetros n y p, $X \sim Bi(n;p)$:

$$f_X(x;n,p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}; n \geq x; Rg(X) = \{0,1,\ldots,n\}$$

Alternativamente, si se tiene una serie de v.a. Bernoulli $X_i \in \{0,1\}$; la variable nueva $Y = \sum_{i=1}^{n} X_i$ (suma de Bernoulli) se distribuye como binomial $Y \sim Bi(n;p)$. Además lo siguiente es equivalente: $Be(p) = Bi(1,p)$. La esperanza de la distribución es $E[X] = np$ y la varianza es $V[X] = np(1-p)$.

En una «caja» hay ocho sustantivos, tres de los cuales son animados (A) y cinco son inanimados (I). Entonces: $p(A) = \frac{3}{8}$ y $p(A^c) = p(I) = 1 - p(A) = \frac{8}{8} - \frac{3}{8} = \frac{5}{8}$. A un sujeto se le presentan tres sustantivos (extraídos al azar de la «caja») y debe decidir si son animados o no. Entonces sea la variable aleatoria X = «número de sustantivos animados en tres sustantivos presentados». Tenemos ocho permutaciones de los eventos «animado» e «inanimado» ($2^n = 2^3 = 8$). La primera columna muestra las posibles permutaciones de la secuencia de eventos. La segunda muestra el valor de la variable aleatoria (cantidad de sustantivos animados en tres intentos). La tercera muestra las probabilidad de cada secuencia. La cuarta calcula la probabilidad de $X = x$ bajo una función de masa de probabilidad binomial.

Permutaciones	$X = x$	$p^x (1-p)^{n-x}$	$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{3-x}$
AAA	3	$\left(\frac{3}{8}\right)^3$	$P(X=3) = \left(\begin{array}{c} 3 \\ 3 \end{array} \right) \left(\frac{3}{8}\right)^3 = (1)\left(\frac{3}{8}\right)^3 \approx 0{,}24$
AII	1	$\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2$	$P(X=1) = \left(\begin{array}{c} 3 \\ 1 \end{array} \right) \left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2 = (3)\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2 \approx 0{,}44$
IAI	1	$\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2$	
IIA	1	$\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2$	
AAI	2	$\left(\frac{3}{8}\right)^2\left(\frac{5}{8}\right)$	$P(X=2) = \left(\begin{array}{c} 3 \\ 2 \end{array} \right) \left(\frac{3}{8}\right)^2\left(\frac{5}{8}\right) = (3)\left(\frac{3}{8}\right)^2\left(\frac{5}{8}\right) \approx 0{,}26$
AIA	2	$\left(\frac{3}{8}\right)^2\left(\frac{5}{8}\right)$	
IAA	2	$\left(\frac{3}{8}\right)^2\left(\frac{5}{8}\right)$	
III	0	$\left(\frac{5}{8}\right)^3$	$P(X=0) = \left(\begin{array}{c} 3 \\ 0 \end{array} \right) \left(\frac{5}{8}\right)^3 = (1)\left(\frac{5}{8}\right)^3 \approx 0{,}05$

Cuadro 16: Probabilidades $P(X = x)$ para $X \sim Bi\left(n = 3, p = \frac{3}{8}\right)$; $X = \{0,1,2,3\}$

Hagamos el cálculo exacto: $X \sim Bi\left(n = 3, p = \frac{3}{8}\right)$, con función de probabilidad de masa: $f_X(x) = P(X = x) = \left(\begin{array}{c} 3 \\ x \end{array} \right) p^x (1-p)^{3-x}$. Por ejemplo, $P(X = 1)$ es:

$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!} \Rightarrow \left(\begin{array}{c} 3 \\ 1 \end{array} \right) = \frac{3!}{1!(3-1)!} = \frac{3\times2\times1}{1\times(2\times1)} = \frac{6}{2} = 3$

$P(X = x) = \left(\begin{array}{c} 3 \\ x \end{array} \right) p^x (1-p)^{3-x} = \left(\begin{array}{c} 3 \\ 1 \end{array} \right) \left(\frac{3}{8}\right)^1 \left(1 - \frac{3}{8}\right)^{3-1} = 3\left(\frac{3}{8}\right)\left(\frac{8-3}{8}\right)^2 = 3\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)^2 = 3\left(\frac{3}{8}\right)\left(\frac{25}{64}\right) = 0{,}43945312$

Las probabilidades para los otros valores de $X \in [0,3]$ son:

$X = x$	$P(X = x)$
0	0,24414062
1	0,43945312
2	0,26367188
3	0,05273438

Cuadro 17: Probabilidades para $X = x$ si $X \sim Bi\left(3, \frac{3}{8}\right)$

[4]En realidad este es el número combinatorio pero sucede que las permutaciones concuerdan con el número combinatorio para el caso de dos grupos. Es decir que:
$\frac{n!}{n_1!n_2!\ldots n_k!} = \left(\begin{array}{c} n \\ n_1!n_2!\ldots n_k! \end{array} \right) = \left(\begin{array}{c} n \\ n_1!n_2! \end{array} \right) = \left(\begin{array}{c} n \\ x!(n-x)! \end{array} \right) = \left(\begin{array}{c} n \\ x \end{array} \right)$

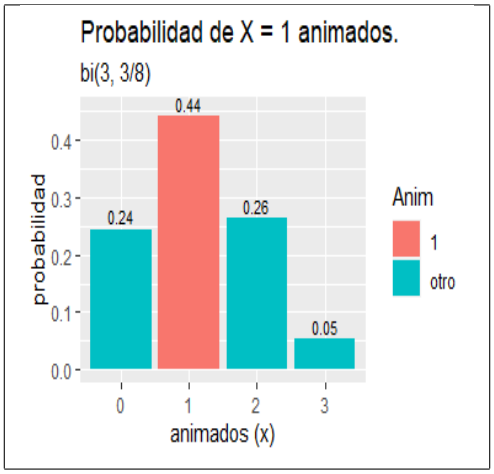


Figura 19: Gráfico de la función de masa de probabilidad.

Estas son las características de un experimento binomial:

- hay n ensayos de prueba (de Bernoulli) con n fijo.
- los n ensayos son idénticos, cada uno con dos resultados posibles: éxito (E) o fracaso (F).
- los n ensayos son independientes entre sí. Es decir que el resultado de uno no influye sobre la probabilidad del resultado de otro.
- la probabilidad de éxito es constante a través de los ensayos.

Las características de la distribución binomial son: (1) el valor de la variable cercano a $E[X]$ es el más probable; (2) la distribución es simétrica con valores de p cercanos a 0,5 (muy parecida a una Normal), es asimétrica a izquierda con valores de p cercanos a 1 y asimétrica a derecha con valores de p cercanos a 0. La Figura que sigue muestra las diferentes situaciones para $X \sim Bi(8,p)$, según $p = \{0{,}1, 0{,}5, 0{,}8\}$ (los números están redondeados al segundo decimal, por eso en algunos valores aparece probabilidad cero; en realidad la masa de probabilidad en estos valores es positiva pero muy cercana a cero).

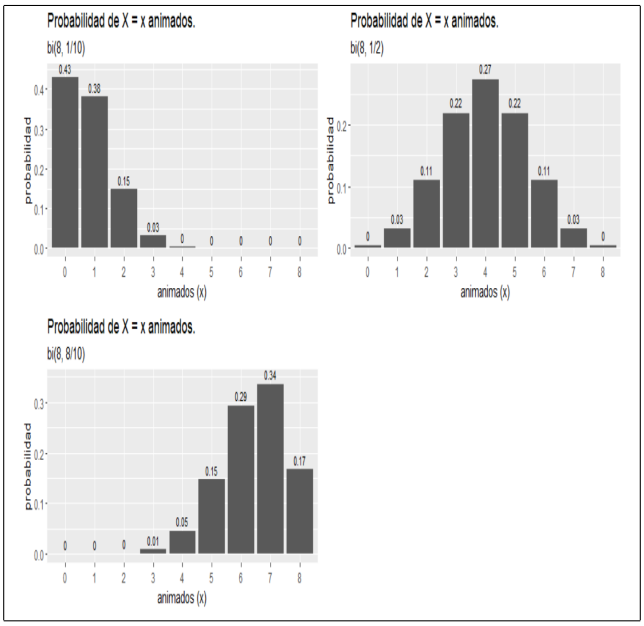


Figura 20: $X \sim Bi(8,p)$, según $p = \{0{,}1, 0{,}5, 0{,}8\}$

Ahora veamos cómo calcular probabilidades a partir de $X \sim Bi(8,0{,}2)$ si la probabilidad decidir que un sustantivo es animado es $p = 0{,}2$ y hay $n = 8$ sustantivos para intentar. En este caso tenemos:

$X = x$	$P(X = x)$
0	0,1678
1	0,3355
2	0,2936
3	0,1468
4	0,0459
5	0,0092
6	0,0011
7	0,0001
8	0,0000

Cuadro 18: Probabilidades para $X = x$ si $X \sim Bi\left(8, \frac{1}{5}\right)$

Calculamos las siguientes probabilidades, según el procedimiento visto para las distribuciones discretas:

- $P(X = 2) = f_X(2) = 0{,}2936$
- $P(X = 2) = F_X(X \leq 2) - F_X(X \leq 1) = \sum_{x=0}^{2} bi(x, 8, 0{,}2) - \sum_{x=0}^{1} bi(x, 8, 0{,}2) = 0{,}7969 - 0{,}5033 = 0{,}2936$
- $P(X < 2) = F_X(X \leq 1) = \sum_{x=0}^{1} bi(x, 8, 0{,}2) = 0{,}1678 + 0{,}3355 = 0{,}5033$
- $P(X \leq 2) = F_X(X \leq 2) = \sum_{x=0}^{2} bi(x, 8, 0{,}2) = 0{,}1678 + 0{,}3355 + 0{,}2936 = 0{,}7969$
- $P(X \geq 2) = 1 - F_X(X \leq 1) = 1 - \sum_{x=0}^{1} bi(x, 8, 0{,}2) = 1 - 0{,}5033 = 0{,}4967$
- $P(X > 2) = P(X \geq 3) = 1 - F_X(X \leq 2) = 1 - \sum_{x=0}^{2} bi(x, 8, 0{,}2) = 1 - 0{,}7969 = 0{,}2031$
- $P(2 \leq X \leq 5) = f_X(2) + f_X(3) + f_X(4) + f_X(5) = 0{,}2936 + 0{,}1468 + 0{,}0459 + 0{,}0092 = 0{,}4955$
- $P(2 \leq X \leq 5) = F_X(X \leq 5) - F_X(X \leq 2) + f_X(2) = \sum_{x=0}^{5} bi(x, 8, 0{,}2) - \sum_{x=0}^{2} bi(x, 8, 0{,}2) + f_X(2) = 0{,}9988 - 0{,}7969 + 0{,}2936 = 0{,}4955$

Por último mencionemos que dadas m variables binomiales independientes de parámetros n_i $(i = 1, 2, \ldots, m)$ y p (igual para todas); la suma de todas ellas también es una binomial de parámetros $n = n_1 + n_2 + \ldots + n_m$ y p. O sea: $Y = \sum_{i=1}^{m} X_i \sim Bi\left(\sum_{i=1}^{m} n_i, p\right)$.

8.1.2. Poisson.

La distribución de Poisson modela una variable aleatoria en la cual se cuenta el número de éxitos en un continuo (tiempo, superficie, volumen). Tiene parámetro λ. Es importante notar que, a diferencia de la Binomial, no hay un n especificado. Se denota: $X \sim P(\lambda)$

$$f_X(x; \lambda) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0, 1, \ldots, \infty.$$

x es el número de ocurrencias del evento o fenómeno; λ es la «intensidad», es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado. e es la base de los logaritmos naturales ($e = 2,71828\ldots$). La esperanza y la varianza de la distribución es el parámetro λ: $E[X] = V[X] = \lambda$. La siguiente figura muestra la forma de la distribución para $\lambda = \{1, 5, 10, 15\}$, por ejemplo, para el conteo de errores gramaticales en un ensayo de español para extranjeros.

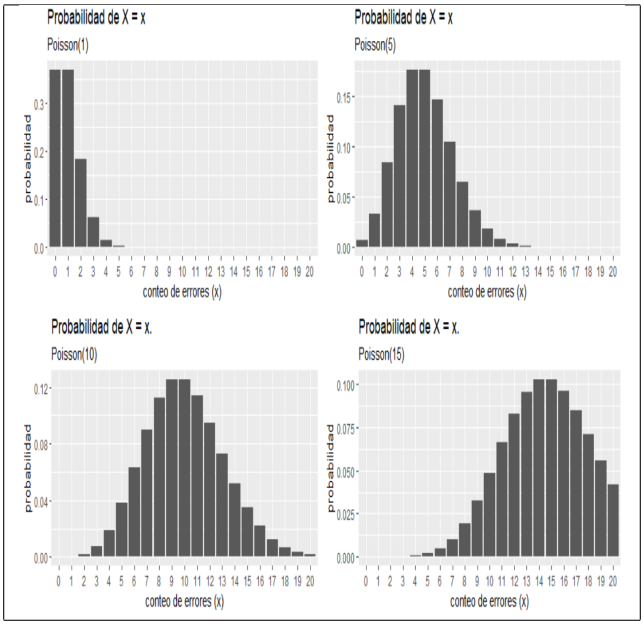


Figura 21: $X \sim P(\lambda)$, según $\lambda = \{1, 5, 10, 15\}$.

Supongamos que durante un examen oral de español para extranjeros un alumno de nivel $B1$ comete errores con una intensidad de $\lambda = \frac{5}{3}$ (5 errores cada tres minutos). ¿Cuál es la probabilidad de no cometer error? ¿y la de cometer al menos 2 errores?

$$P(X = 0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\left(\frac{5}{3}\right)}\left(\frac{5}{3}\right)^0}{0!} = e^{-\left(\frac{5}{3}\right)} = 0{,}189$$

$$P(X = 1) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\left(\frac{5}{3}\right)}\left(\frac{5}{3}\right)^1}{1!} = e^{-\left(\frac{5}{3}\right)}\left(\frac{5}{3}\right) = 0{,}314$$

$$P(X \geq 2) = 1 - F_X(X \leq 1) = 1 - \sum_{x=0}^{1} Pois\left(x, \tfrac{5}{3}\right) = 1 - (0{,}189 + 0{,}314) = 1 - 0{,}503 = 0{,}497$$

Cuando $n \to \infty$, $p \to 0$, $\lambda = np$ constante ($p = \frac{\lambda}{n}$), entonces la distribución Binomial se aproxima a una distribución Poisson: $X \sim Bi(n, p) \to X \approx P(\lambda)$. En la práctica es preferible que $n \geq 100$, $p \approx 0$ y $np \approx 1$.

8.2. Familias de distribuciones continuas.

8.2.1. Uniforme.

Esta distribución asigna igual probabilidad a los valores de un intervalo de la v.a. Su función de densidad de probabilidad [f.d.p.] tiene la forma de un rectángulo donde $b - a$ es la base y $\frac{1}{b-a}$ es la altura. Su esperanza es $E[X] = \frac{a+b}{2}$ y su varianza es $Var[X] = \frac{(b-a)^2}{12}$.

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & c.c. \end{cases}$$

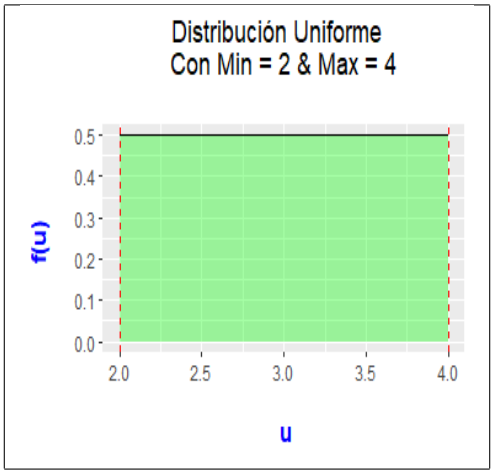


Figura 22: $X \sim U(2,4)$.

8.2.2. Normal.

La distribución normal tiene las siguientes características:

- Es una curva suave, acampanada, simétrica, con única moda.
- Tiene dos parámetros: (1) μ indica la posición y (2) σ indica la escala o dispersión (ancho de la campana).
- Su punto de simetría es μ. Hay dos puntos de inflexión (cambios de concavidad) a distancia $-\sigma$ y $+\sigma$ de μ (eje de simetría).
- Su esperanza es $E[X] = \mu$ y su varianza es $V[X] = \sigma^2$.
- Se denota como $X \sim N(\mu, \sigma^2)$, «la v.a. X tiene distribución normal con esperanza μ y varianza σ^2».

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

La Figura siguiente muestra diferentes distribuciones: (1) con $\mu = -1, 0, 1$; (2) con $\sigma = 0{,}2; 0{,}5; 0{,}9$ ($\mu = 0$).

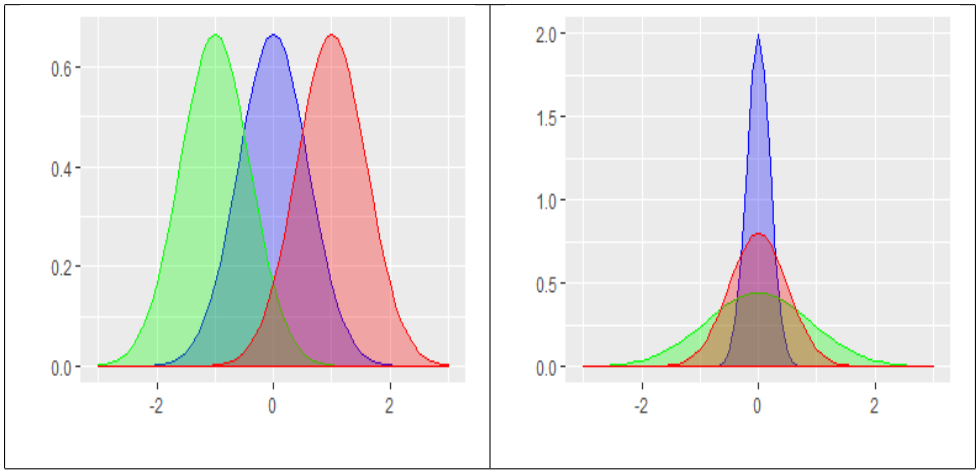


Figura 23: Función de densidad de probabilidad normales con: con $\mu = -1, 0, 1$ [Izq.]; (2) con $\sigma = 0{,}2, 0{,}5, 0{,}9$ [Der.].

Por otro lado, toda la campana queda comprendida entre la media más / menos tres desvíos típicos: (1) entre $\mu \pm 3\sigma$ hay densidad $p = 0{,}997$; (2) entre $\mu \pm 2\sigma$ hay densidad $p = 0{,}954$; (3) entre $\mu \pm \sigma$ hay densidad $p = 0{,}683$.

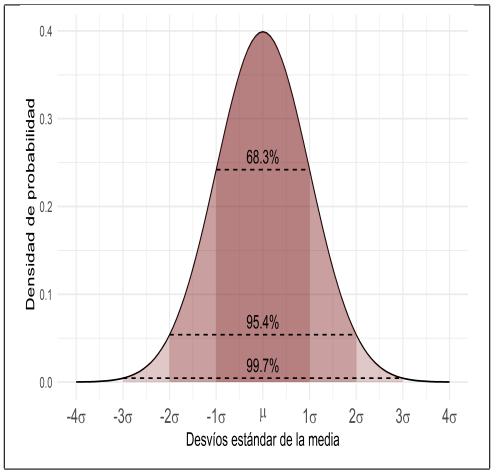


Figura 24: Distribución normal.

Para calcular la probabilidad se usa la distribución normal estándar. Sea $X \sim N\left(\mu, \sigma^2\right)$, la transformación estándar es $Z = \frac{X-\mu}{\sigma}$, $Z \sim N\left(0,1\right)$. Z mide la distancia al centro $(\mu = 0)$ en unidades de desviaciones estándar.

$$f_Z\left(z;1,0\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Las probabilidades se calculan de la siguiente manera; donde $\Phi\left(k\right) = P\left(Z \leq k\right)$ es la función normal estándar acumulada en k.

- $P\left(a \leq X \leq b\right) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(Z_a \leq Z \leq Z_b\right) = \Phi\left(z_b\right) - \Phi\left(z_a\right)$
- $P\left(X \leq b\right) = P\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(Z \leq z_b\right) = \Phi\left(z_b\right)$
- $P\left(X \geq b\right) = P\left(\frac{X-\mu}{\sigma} \geq \frac{b-\mu}{\sigma}\right) = P\left(Z \geq z_b\right) = 1 - \Phi\left(z_b\right)$

Por ejemplo, sea $X \sim N\left(\mu = 80, \sigma^2 = 144\right)$, entonces: $Z = \frac{X-80}{12}$ (¡ojo!: $\sigma^2 = 144 \Rightarrow \sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$); y $Z \sim N\left(0,1\right)$.

a) $P\left(90 \leq X \leq 100\right) = P\left(\frac{90-80}{12} \leq \frac{X-80}{12} \leq \frac{100-80}{12}\right) = P\left(0{,}83 \leq Z \leq 1{,}67\right) = \Phi\left(1{,}67\right) - \Phi\left(0{,}83\right) = 0{,}9525 - 0{,}7967 = 0{,}1558$

b) $P\left(X \leq 60\right) = P\left(\frac{X-80}{12} \leq \frac{60-80}{12}\right) = P\left(Z \leq -1{,}67\right) = \Phi\left(-1{,}67\right) = 0{,}0475$

c) $P\left(X \geq 110\right) = P\left(\frac{X-80}{12} \geq \frac{110-80}{12}\right) = P\left(Z \geq 2{,}5\right) = 1 - \Phi\left(2{,}5\right) = 1 - 0{,}9938 = 0{,}0062$

Por último, veamos cómo aproximar la Binomial a la Normal. Supongamos $X \sim Bi\left(n,p\right)$ con $E\left[X\right] = np$ y $Var\left[X\right] = np\left(1-p\right)$, entonces podemos aproximar una v.a. Binomial X con una v.a. normal $Y \sim N\left(\mu = np, \sigma^2 = np\left(1-p\right)\right)$. En la práctica se recomienda que n sea grande y $p \approx 0{,}5$ (en todo caso no debe ser extremo, o sea cercano a 0 o 1). Téngase en cuanta como regla heurística: $np \geq 5$ y $n\left(1-p\right) \geq 5$ (deben cumplirse ambos requisitos: cantidad de éxitos y fracasos mayor o igual a cinco). Por ejemplo, $X \sim Bi\left(25, 0{,}6\right)$, podemos aproximar X con la v.a. Y con $\mu = 25\left(0{,}6\right) = 15$ y $\sigma = \sqrt{25\left(0{,}6\right)\left(1-0{,}6\right)} = 2{,}45$. Entonces:

$$P\left(X \leq 13\right) \simeq P\left(Y \leq 13\right) = P\left(\frac{Y-15}{2{,}45} \leq \frac{13-15}{2{,}45}\right) = P\left(Z \leq -0{,}82\right) = 0{,}206$$

Si calculamos usando la binomial:

$$P\left(X \leq 13\right) = F_x\left(X \leq 13\right) = \sum_{x=0}^{13} Bi\left(x, 25, 0{,}6\right) = \sum_{x=0}^{13} \left(\begin{array}{c} 25 \\ x \end{array} \right) \left(0{,}6\right)^x \left(1-0{,}6\right)^{25-x} = 0{,}267$$

Es posible aproximar la binomial a la normal mejor haciendo una corrección de continuidad en la Normal:

$$P(X \leq x) = P\left(Y \leq x + \tfrac{1}{2}\right)$$
$$P(X \geq x) = P\left(Y \geq x - \tfrac{1}{2}\right)$$
$$P(X < x) = P\left(Y \leq x - \tfrac{1}{2}\right)$$
$$P(X > x) = P\left(Y \geq x + \tfrac{1}{2}\right)$$

En el ejemplo anterior:

$$P(X \leq 13) \simeq P(Y \leq 13{,}5) = P\left(\frac{Y-15}{2{,}45} \leq \frac{13{,}5-15}{2{,}45}\right) = P(Z \leq -0{,}61) = 0{,}271$$

8.2.3. Gamma.

La familia de distribuciones Gama con parámetros α y β, denotada $gamma(\alpha, \beta)$, es:

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

El parámetro α dicta la "forma" de la distribución (cuanto más grande más "picuda" será) y β es un parámetro de escala (cuanto más grande más dispersión). $\Gamma(n) = (n-1)!$ es la función Gamma. El siguiente gráfico muestra la forma de la distribución según diferentes elecciones para los parámetros $k = \alpha$ y $\theta = \beta$. Su esperanza es $E[X] = \alpha\beta$ y su varianza es $V[X] = \alpha\beta^2$.

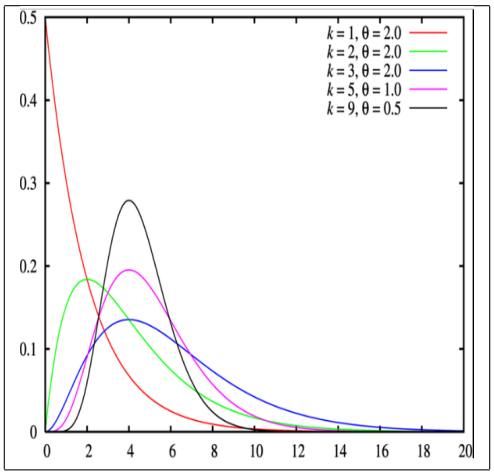


Figura 25: La distribución Gama.

Hay casos especiales importantes de la distribución Gamma. Estableciendo: $\alpha = \frac{v}{2}$ $(v \in \mathbb{N})$ y $\beta = 2$ se obtiene la distribución Chi-cuadrada con v grados de libertad $\chi^2_{(\nu)}$:

$$f_X(x; v) = \frac{1}{\Gamma\left(\frac{v}{2}\right) 2^{\frac{v}{2}}} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}; 0 < x < \infty; v = 1, 2, \ldots$$

$E[X] = \alpha\beta = \left(\frac{v}{2}\right) 2 = v$; $V[X] = \alpha\beta^2 = \frac{v}{2} 2^2 = 2v$.

La siguiente figura muestra la distribución con diferentes valores de $v = k$.

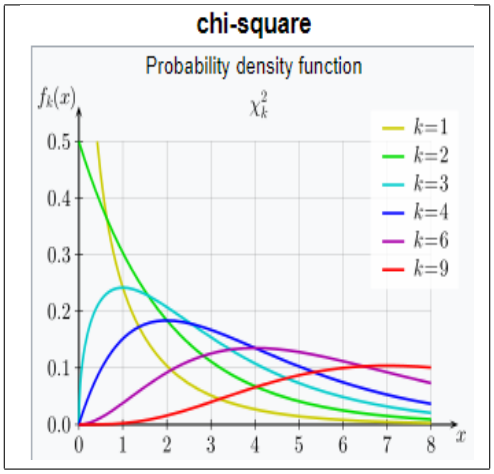


Figura 26: La distribución Chi-cuadrado.

Los siguientes resultados también son de importancia:

a) $Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2_{(1)}$. El cuadrado de una variable normal estándar tiene distribución Chi-cuadrado con un grado de libertad.

b) Si $X_1, \ldots, X_n$ son v.a. independientes cada una distribuida como chi-cuadrado con v_i grados de libertad: $X_i \sim \chi^2_{(v_i)}$; entonces su suma también tiene distribución chi-cuadrado cuyo parámetro es la suma de los grados de libertad de la distribución de cada variable; o sea: $X_1 + X_2 + \ldots + X_n \sim \chi^2_{(v_1+v_2+\ldots+v_n)}$.

Además, si en Gamma, $\alpha = 1$ se obtiene la *fdp* de la Exponencial con parámetro de escala β.

$$f_X(x;\beta) = \frac{1}{\Gamma(1)\,\beta^1} x^{1-1} e^{-\frac{x}{\beta}} = \frac{1}{\beta} e^{-\frac{x}{\beta}}; 0 < x < \infty$$

$E[X] = \beta$; $V[X] = \beta^2$.

La siguiente Figura muestra la distribución con diferentes valores de $\beta = \lambda$.

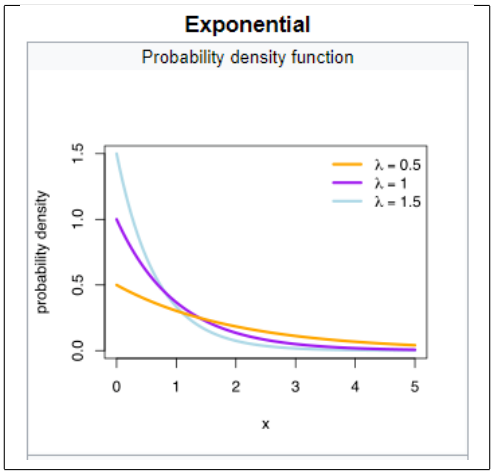


Figura 27: La distribución Exponencial.

En general la distribución Gamma modela una v.a. que indica el tiempo hasta que ocurre un evento (por ejemplo, latencias en experimentos de psicolingüística). De hecho, la Exponencial se utiliza para modelar la función *hazard* en análisis de datos hasta un evento. Además el tiempo entre arribos en un *proceso de Poisson* tiene distribución exponencial.

8.2.4. Ley de Zipf.

Si se cuentan las ocurrencias de las palabras (*types*) en un corpus y luego se listan dichas palabras en orden de frecuencia ocurrencia, es posible ver una relación entre la frecuencia de una palabra f y su posición en la lista, es decir, su rango r. La ley de Zipf establece que: $f \propto \frac{1}{r}$. Esto es, la frecuencia de una palabra en un corpus es inversamente proporcional a su posición (rango) en la lista de frecuencias. Dicho de otro modo, la primera palabra en la lista de frecuencias es dos veces más frecuente que la segunda palabra en la lista; tres veces más frecuente que la tercera, etc. El resultado de esto es que un corpus contiene pocas palabras con frecuencia muy alta y muchas con frecuencia muy baja. Por ello, la mayoría de las palabras de un corpus ocurren solamente una vez (se denominan *hapax legomena*). La Figura que sigue muestra la distribución de frecuencias de cien palabras tomadas de un corpus. Se grafican las frecuencias en orden descendente (de las frecuencias más altas a las más bajas). Se nota que la distribución es fuertemente asimétrica a derecha, con una cola muy larga (Notar que es incluso más larga que en una Exponencial). El rango de las frecuencias está entre $[0, 75080]$.

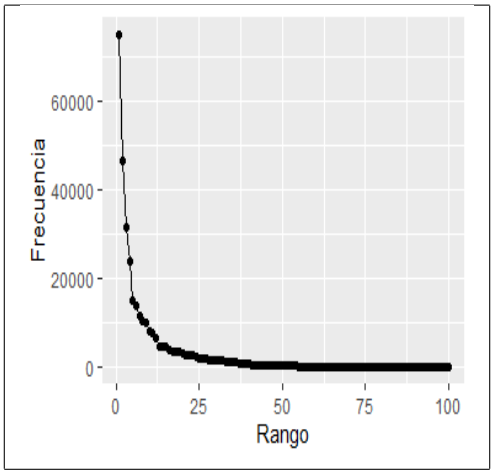


Figura 28: Distribución de frecuencias de *100* palabras de un corpus.

9. Esperanza, varianza y covarianza (optativo).

9.1. Esperanza.

Valor esperado. La esperanza de una variable aleatoria X es un valor promedio ponderado de acuerdo a la distribución de probabilidad de X.

$$E[X] = \begin{cases} \sum_{x=0}^{\infty} x f_X(x) & X \text{ v.a. discreta} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & X \text{ v.a. continua} \end{cases}$$

(1) Ejemplo con X discreta. Calcularemos la esperanza de para la distribución del Cuadro *17*, en donde $X \sim Bi\left(n = 3, p = \frac{3}{8}\right)$. Como es Binomial, sabemos que $E[x] = np = 3\left(\frac{3}{8}\right) = \frac{9}{8} = 1{,}125$; lo cual coincide con:

$$E[X] = \sum_{x=0}^{3} x f_X(x) = 0 \times 0{,}24414062 + 1 \times 0{,}43945312 + 2 \times 0{,}26367188 + 3 \times 0{,}05273438 = 1{,}125$$

(2) Ejemplo con X discreta. Calcularemos la esperanza de una distribución Poisson: $E[X] = \lambda$. Recordar que su f.p.m. es:

$$f_X(x; \lambda) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0, 1, \ldots, \infty.$$

$$E[X] = \sum_{x=0}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \Rightarrow x = 0 \; da \; cero \Rightarrow \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$
$$\sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}(\lambda^1 \lambda^{x-1})}{x(x-1)!} = \lambda \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^{x-1}}{x(x-1)!}$$
$$\lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \Rightarrow sustitución: y = x - 1 \Rightarrow \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda(1) = \lambda$$

Propiedades de la esperanza. Sea las v.a. X, Y y a, b, c constantes.

a) $E[aX + bY + c] = aE[X] + bE[Y] + c$ (la esperanza de una suma es la suma de sus esperanzas).

b) $E[c] = c$ (la esperanza de una constante es la misma constante).

c) Si $X \geq 0 \Rightarrow E[X] \geq 0$ (si X es positiva entonces también lo es su esperanza).

d) Si $X \geq Y \Rightarrow E[X] \geq E[Y]$.

e) Si $a \leq X \leq b \Rightarrow a \leq E[X] \leq b$

f) $E[XY] = E[X]E[Y]$ si X e Y son *independientes.*

9.2. Momentos de una distribución.

Sea $n \in \mathbb{N}$; el $n - ésimo$ momento (alrededor del cero) de X es: $\mu_n = E[X^n]$ (si $E[X^n] < \infty$; o sea, si existe). Por ejemplo, el primer momento es $\mu_1 = E[X]$, o sea, la esperanza; el segundo momento es $\mu_2 = E\left[X^2\right]$. El el $n - ésimo$ momento (alrededor del cero) de X se calcula como (la esperanza es $n = 1$):

$$E[X^n] \begin{cases} \sum_{x=0}^{\infty} x^n f_X(x) & X \; v.a. \; discreta \\ \int_{-\infty}^{\infty} x^n f_X(x)\, dx & X \; v.a. \; continua \end{cases}$$

Por otro lado, el $n - ésimo$ momento *central* (alrededor de $\mu = \mu_1 = E[X]$) de X es: $\mu_n = E[(X - \mu)^n]$, con $\mu = \mu_1 = E[X]$. Se calcula como[5]:

$$E[(X-\mu)^n] \begin{cases} \sum_{x=0}^{\infty} (x-\mu)^n f_X(x) & X \; v.a. \; discreta \\ \int_{-\infty}^{\infty} (x-\mu)^n f_X(x)\, dx & X \; v.a. \; continua \end{cases}$$

9.3. Varianza.

La varianza de una v.a. X es el *segundo momento central.*

$$V[X] = E\left[(X - E[X])^2\right]$$

Se calcula simplemente reemplazando $n = 2$ en la fórmulas para los momentos centrales:

$$V[X] = E\left[(X-\mu)^2\right] \begin{cases} \sum_{x=0}^{\infty} (x-\mu)^2 f_X(x) & X \; v.a. \; discreta \\ \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx & X \; v.a. \; continua \end{cases}$$

Otra fórmula para la varianza es: $V[X] = E\left[X^2\right] - (E[X])^2$.

[5] Notar que el primer momento central ($n = 1$) es cero, ya que, si $\mu = E[X]$:

$$\mu_1 = E\left[(X-\mu)^1\right] = E[X] - E[\mu] = E[X] - \mu = E[X] - E[X] = 0$$

(1) Ejemplo con X discreta. Calcularemos la varianza de X para la distribución del Cuadro *17*, en donde $X \sim Bi\left(n=3, p=\frac{3}{8}\right)$. Como es Binomial, sabemos que $V[X] = np(1-p) = 3\left(\frac{3}{8}\right)\left(1-\frac{3}{8}\right) = \frac{9}{8}\frac{5}{8} = \frac{45}{64} = 0{,}703125$; y recordar que $\mu = E[X] = 1{,}125$; entonces:

$$V[X] = E\left[(X-\mu)^2\right] = \sum_{x=0}^{3}(x-\mu)^2 f_X(x) =$$
$$(0-1{,}125)^2 \times 0{,}24414062 + (1-1{,}125)^2 \times 0{,}43945312 + (2-1{,}125)^2 \times 0{,}26367188 + (3-1{,}125)^2 \times 0{,}05273438 =$$
$$(0{,}24414062 \times 0{,}24414062) + (0{,}43945312 \times 0{,}43945312) + (0{,}26367188 \times 0{,}26367188) + (0{,}05273438 \times 0{,}05273438) =$$
$$= 0{,}703125$$

O bien, usando $E\left[X^2\right]$:

$$E\left[X^2\right] = \sum_{x=0}^{3} x^2 f_X(x) = 0^2 \times 0{,}24414062 + 1^2 \times 0{,}43945312 + 2^2 \times 0{,}26367188 + 3^2 \times 0{,}05273438 = 1{,}96875$$

$$V[X] = E\left[X^2\right] - (E[X])^2 = 1{,}96875 - (1{,}125)^2 = 0{,}703125$$

(2) Ejemplo con X discreta. Calcularemos la varianza de una distribución Poisson: $V[X] = \lambda$.

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) f_X(x) = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} \Rightarrow x=0, x=1 \; da\; cero \Rightarrow \sum_{x=2}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!}$$
$$\sum_{x=2}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=2}^{\infty} x(x-1)\frac{e^{-\lambda}\left(\lambda^2\lambda^{(x-2)}\right)}{x(x-1)(x-2)!} = \lambda^2 \sum_{x=2}^{\infty}\frac{e^{-\lambda}\lambda^{(x-2)}}{(x-2)!} \Rightarrow sustitución: y = x-2$$
$$\Rightarrow \lambda^2 \sum_{y=0}^{\infty}\frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2$$

Notar que: $E\left[X^2\right] = E\left[X^2 - X + X\right] = E[X(X-1) + X] = E[X(X-1)] + E[X] = \lambda^2 + \lambda$

Entonces: $V[X] = E\left[X^2\right] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Propiedades de la varianza. Sea la v.a. X, con $V[X] < \infty$ y a, b constantes.

a) $V[X] \geq 0$ (no puede ser negativa).

b) $V[b] = 0$ (la varianza de una constante es *cero*).

c) $V[aX+b] = a^2 V[X]$

d) $V[X \pm Y] = V[X] + X[Y]$ si X e Y son independientes[6].

e) $V[X+Y] = V[X] + V[Y] + 2Cov[XY]$ si X e Y no son independientes.

f) $V[X-Y] = V[X] + V[Y] - 2Cov[XY]$ si X e Y no son independientes.

Las tres últimas propiedades se pueden generalizar como: $V[aX \pm bY] = a^2 V[X] + b^2 V[Y] \pm 2abCov[X,Y]$. Si son independientes $Cov[XY] = 0$.

9.4. Covarianza.

La covarianza de dos variables aleatorias es una propiedad poblacional de la distribución conjunta. Cuando los valores altos de una de las variables aleatorias suelen mayoritariamente corresponderse con los valores altos de la otra, y lo mismo se verifica para los pequeños valores de una con los de la otra, se corrobora que tienden a mostrar comportamiento similar lo que se refleja en un valor *positivo* de la covarianza. Por el contrario, cuando los valores altos de una variable aleatoria suelen corresponder mayoritariamente a los menores valores de la otra, expresando un comportamiento opuesto, la covarianza es negativa.

El signo de la covarianza, por lo tanto, expresa la tendencia en la relación *lineal* entre las variables.

Sean: $E[X] = \mu_X$, $E[Y] = \mu_Y$, $V[X] = \sigma_X^2$, $V[Y] = \sigma_Y^2$; con $\sigma_X^2 > 0$ y $\sigma_Y^2 > 0$. Entonces:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

[6] Observar que aquí no importa el signo. En efecto: $V[X-Y] = V[X+(-Y)] = V[X] + V[(-1)Y] = V[X] + (-1)^2 V[Y] = V[X] + V[Y]$

La versión estandarizada de las variables aleatorias es el coeficiente de correlación, que es independiente de la escala de medida de las variables[7].

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Tienen el siguiente comportamiento:

- $-\infty < Cov(X,Y) < \infty$ y $-1 < \rho < 1$.
- Si $Cov(X,Y) > 0$ hay dependencia lineal positiva, es decir que altos (bajos) valores de X se corresponden con altos (bajos) valores de Y. Si $Cov(X,Y) < 0$ hay dependencia lineal negativa; es decir que altos valores de X se corresponden con bajos de Y, y bajos de X con altos de Y. Si $Cov(X,Y) = 0$ no hay relación lineal entre las dos v.a.
- Es fundamental que la relación entre las variables sea lineal. Si no lo fuera (por ejemplo, una relación cuadrática) entonces $Cov(X,Y) = 0$ no implica que no haya relación.

La covarianza también se puede calcular mediante: $Cov[X,Y] = E[XY] - \mu_X \mu_Y$.

Propiedades de la Covarianza.

a) $Cov(X,a) = 0$ (la covarianza de una v.a. con una constante es cero).

b) $Cov(X,X) = V[X]$ (la covarianza de una v.a. consigo misma es la varianza de esa v.a.).

c) $Cov(X,Y) = Cov(Y,X)$ (simetría).

d) $Cov(aX + c, bY + d) = abCov(X,Y)$.

e) Si Xe Y son v.a. independientes entonces $cov(X,Y) = 0$; pero al revés no es cierto: si $cov(X,Y) = 0$ no necesariamente son independientes (excepto en la distribución normal bivariada).

10. Distribuciones muestrales.

Como dijimos, en probabilidades calculamos la probabilidad del valor de una variable aleatoria dado que conocemos los parámetros de la distribución que genera dicho valor. Ahora suponemos una cierta distribución cuyos parámetros son, en general, desconocidos y extraemos una muestra aleatoria de dicha distribución. El conjunto de variables aleatorias $X_1 = x_1$, $X_2 = x_2$,..., $X_n = x_n$ (la v.a. X_i toma el valor x_i) se llama «muestra aleatoria» [m. a.] de tamaño n de una población si: (1) las variables aleatorias son mutuamente independientes; (2) la f.d.p. (f.p.m.) de cada X_i son las mismas. Es decir que las v.a. son *independientes e idénticamente distribuidas* (i.i.d.). Si la población modelada por la f.d.p. o la f.p.m. es miembro de una familia paramétrica $f_X(x,\theta)$, entonces:

$$f_X(x_1, x_2, \ldots, x_n; \theta) = f_{X_1}(x_1) \times f_{X_2}(x_2) \times \ldots \times f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Denotaremos una m.a. como conjunto de letras en mayúsculas: $m.a. = \{X_1, X_2, \ldots, X_n\}$ y a los valores que toman en una muestra como letras minúsculas: $datos = \{x_1, x_2, \ldots, x_n\}$. Ahora bien, un estadístico estima un determinado parámetro desconocido de una distribución (conocida o no). Se define como una función de la muestra: $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$ es estimador del parámetro poblacional θ (un número desconocido). Por ejemplo: $media = \frac{X_1 + X_2 + \ldots + X_n}{n}$, $mínimo = min\left(X^{(1)} \leq X^{(2)} \leq \ldots \leq X^{(n)}\right)$; etc. El estadístico es función de los datos de la muestra y no puede incluir ningún *parámetro poblacional*; por ejemplo, lo siguiente no es un estadístico: $\frac{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_n-\mu)^2}{n}$. Pero lo siguiente sí lo es: $\frac{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+\ldots+\left(X_n-\bar{X}\right)^2}{n}$ (varianza sesgada). Los estadísticos son variables aleatorias porque si cambian los datos de la muestra, la función da un valor diferente. Por lo tanto, los estadísticos, al ser v. a., tendrán una distribución asociada y también esperanza y varianza (si existen). La distribución muestral de un estadístico es la distribución de probabilidad asociada a los valores que puede tomar el estadístico de interés (cuando cambian los datos de una muestra de tamaño n). Definamos:

$$media\ muestral:\ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

[7] Notar que es la esperanza de las variables estandarizadas: $\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = E\left[\frac{X-\mu_X}{\sigma_X}.\frac{X-\mu_Y}{\sigma_Y}\right] = E[Z_X Z_Y]$

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

$$proporción\ muestral:\ \hat{p} = \frac{x}{n}; x = total\ aciertos$$

En la varianza, $n-1$ es una corrección para poblaciones finitas; si la población tiene n grande no importa mucho poner n o $n-1$. Por otro lado, notemos que: $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$.

Es posible verificar esto haciendo:

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) = \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \frac{n}{n}\sum_{i=1}^{n} 2X_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \sum_{i=1}^{n}\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}\bar{X} + \sum_{i=1}^{n}\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2\end{aligned}$$

Por lo tanto, lo siguiente es equivalente:

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

Veamos el siguiente Teorema. Sea $X_1, \ldots, X_n$ una muestra aleatoria (o sea i.i.d.) de una distribución *cualquiera* con esperanza $E[X_i] = \mu$ y varianza $V[X_i] = \sigma^2 < \infty$. Entonces:

a) $E\left[\bar{X}\right] = \mu$ (la esperanza (de la distribución) de la media muestral es la esperanza poblacional μ).

b) $V\left[\bar{X}\right] = \frac{\sigma^2}{n}$ (la varianza (de la distribución) de la media muestral es $\frac{\sigma^2}{n}$). Notar que cuando $n \to \infty$, $V\left[\bar{X}\right] \to 0$: al aumentar el tamaño de la muestra la varianza de la distribución muestral tiende a cero. Notar que $V\left[\bar{X}\right]$ indica la precisión del estimador media muestral. Cuando aumenta n dicho estimador es cada vez más preciso.

c) $E\left[S^2\right] = \sigma^2$(la esperanza (de la distribución) de la varianza muestral es la varianza poblacional σ^2)

Vamos a ejemplificar mediante una simulación de la distribución muestral de $\bar{X}$. Por ejemplo supongamos una v. a. que sigue una distribución Exponencial con parámetro $\lambda = 2$: $X \sim Exp(\lambda = 2)$. Supongamos que dicha distribución modela el tiempo (en años luego de la explosión léxica) hasta la primera ocurrencia del plural en sustantivos en la adquisición de L1. Esta es la distribución poblacional. Una variable aleatoria (distribuida como) exponencial tiene esperanza $E[X] = \mu = \frac{1}{\lambda} = \frac{1}{2}$ y varianza $V[X] = \sigma^2 = \frac{1}{\lambda^2} = \frac{1}{2^2} = \frac{1}{4}$. Su forma es la siguiente:

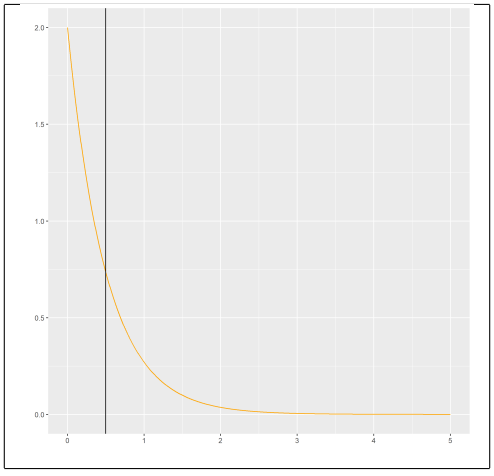


Figura 29: $X \sim Exp(\lambda = 2)$

Estamos interesados en conocer el valor de la esperanza poblacional: ¿Cuál es el tiempo promedio de primera producción de un sustantivo plural en los niños? El valor verdadero es (en este ejercicio) $\mu = \frac{1}{2}$ (los niños de la

población tardan en promedio seis meses en producir el primer sustantivo plural). Pero en general este número lo desconocemos porque no tenemos acceso a la población sino a la muestra que nos tocó de tamaño diez, digamos $m_0 = \{X_1, X_2, \ldots, X_{10}\}$. Entonces estimamos μ mediante la media muestral $\bar{X}$. Pero $\bar{X}$ es un estadístico que cambia de valor con la muestra, es una variable aleatoria con una distribución muestral. Para construir la distribución de la media muestral tomamos 1000 muestras de tamaño $n = 10$ de la distribución exponencial $Exp\,(\lambda = 2)$ y a cada muestra le calculamos $\bar{X}_k$ $(k = 1, 2, 3, \ldots, 1000)$; es decir:

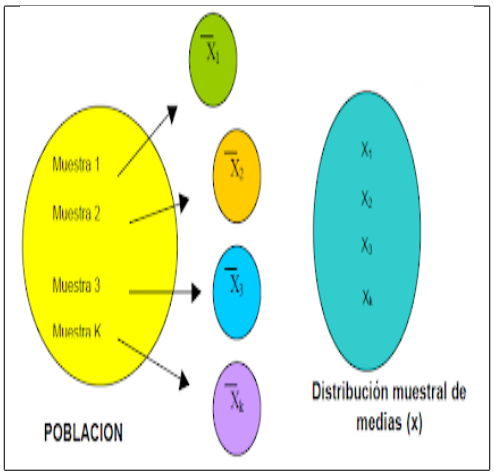


Figura 30: Creación de la distribución muestral de $\bar{X}$

Luego a partir de los valores $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{1000}\}$ construimos la siguiente distribución.

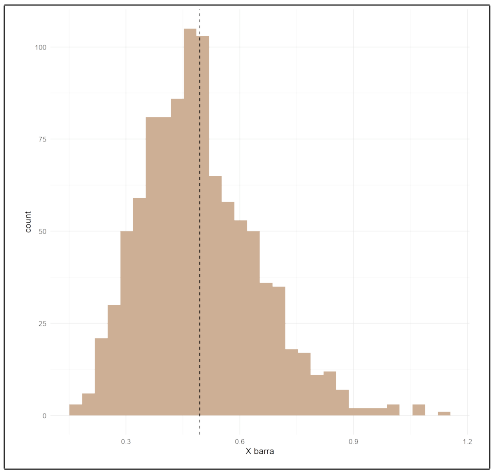


Figura 31: Distribución muestral de $\bar{X}$

La esperanza de la distribución muestral es $E\left[\bar{X}\right] = 0{,}4960523 \approx 0{,}5 = \mu$ y la varianza es $V\left[\bar{X}\right] = 0{,}02613567 \approx 0{,}025 = \frac{0{,}25}{10} = \frac{\sigma^2}{n}$.

estimador insesgado. Un estadístico $\hat{\theta} = T\,(X_1, X_2, \ldots, X_n)$ es un estimador insesgado de un parámetro poblacional θ si sucede que la esperanza de la distribución muestral del estimador (o sea del estadístico) es igual al valor verdadero del parámetro poblacional.

$$E\left[T\,(X_1, X_2, \ldots, X_n)\right] = E\left[\hat{\theta}\right] = \theta$$

Dicho en otras palabras. Un estadístico, al ser una v.a., tiene una distribución con una esperanza y una varianza. El soporte del estadístico está formado por todos los posibles valores que puede tener el estadístico al cambiar la muestra. Si el promedio de todos esos valores posibles (la esperanza del estadístico) coincide con el valor verdadero del parámetro, entonces diremos que el estimador es insesgado. Es decir que, en promedio, el estimador está estimando

bien. En los resultados enunciados más arriba, se puede ver que la media ($\bar{X}$) y la varianza (S^2) muestrales son estimadores insesgados de los parámetros poblacionales respectivos μ y σ^2.

(Optativo) A continuación se demuestran los incisos de los resultados enunciados.

(a) $E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \Rightarrow i.d. \Rightarrow \frac{1}{n}\sum_{i=1}^{n} \mu_i = \frac{1}{n}n\mu = \mu$ ($\bar{X}$ es insesgado de μ)

(b) $V\left[\bar{X}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] \Rightarrow indep. \Rightarrow \frac{1}{n^2}\sum_{i=1}^{n} V\left[X_i\right] \Rightarrow i.d. \Rightarrow \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$

(c) Probar que $E\left[S^2\right] = \sigma^2$, o sea que s^2 es insesgado de σ^2 requiere un poco más de paciencia. Se comienza aplicando que: $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$.

$$E\left[S^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$
$$= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n} X_i^2\right] - E\left[n\bar{X}^2\right]\right\} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\bar{X}^2\right]\right\}$$

ahora recordemos lo siguiente:

$$V\left[X_i\right] = \sigma^2 = E\left[X_i^2\right] - \left(E\left[X_i\right]\right)^2 = E\left[X_i^2\right] - \mu^2 \Rightarrow E\left[X_i^2\right] = \sigma^2 + \mu^2$$

$$V\left[\bar{X}\right] = E\left[\bar{X}^2\right] - \left(E\left[\bar{X}\right]\right)^2 \Rightarrow E\left[\bar{X}\right] = \mu \Rightarrow V\left[\bar{X}\right] = E\left[\bar{X}^2\right] - \mu^2 \Rightarrow E\left[\bar{X}^2\right] = V\left[\bar{X}\right] + \mu^2 = \frac{\sigma^2}{n} + \mu^2$$

Luego, reemplazando en $E\left[X_i^2\right]$ y en $E\left[\bar{X}^2\right]$, nos queda:

$$\frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\bar{X}^2\right]\right\} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\}$$
$$= \frac{1}{n-1}\left\{n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} = \frac{1}{n-1}\left\{n\left[\sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)\right]\right\}$$
$$= \frac{1}{n-1}\left\{n\left[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right]\right\} = \frac{1}{n-1}\left\{n\left[\sigma^2 - \frac{\sigma^2}{n}\right]\right\} = \frac{1}{n-1}\left\{n\left[\sigma^2\left(1 - \frac{1}{n}\right)\right]\right\}$$
$$= \frac{1}{n-1}\left\{n\left[\sigma^2\left(\frac{n-1}{n}\right)\right]\right\} = \sigma^2$$

Si la varianza hubiera sido dividiendo por $\frac{1}{n}$ en lugar de $\frac{1}{n-1}$, hubiéramos tenido: $\frac{1}{n}\left\{n\left[\sigma^2\left(\frac{n-1}{n}\right)\right]\right\} = \left(\frac{n-1}{n}\right)\sigma^2$. Es decir que la varianza está sesgada por un factor de $\left(\frac{n-1}{n}\right)$. Notar que el sesgo del estimador $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ se puede corregir dividiendo por $\left(\frac{n}{n-1}\right)$ obteniendo la varianza muestral insesgada: $S^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.

error estándar. El error estándar de la distribución de la media muestral se llama «error estándar de la media» y se calcula mediante el desvío: $ES\left[\bar{X}\right] = \sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, cuyo estimador es: $\widehat{ES} = \frac{S}{\sqrt{n}}$. Indica cuán preciso es el estimador; o sea cuán cerca se encuentran los valores del estimador del centro, o sea, de su promedio; el cual, al ser insesgado, coincide con el verdadero parámetro. Cuando menos error estándar más precisa es la estimación de μ hecha por $\bar{X}$. Es importante distinguir S de $\widehat{ES}$. El estadístico S nos dice cuál es la dispersión de los datos. El segundo, $\widehat{ES}$, nos dice cuál es la dispersión en la distribución muestral de la media; es decir, es una medida de la precisión de la estimación que hace la media muestral de la media poblacional. Muchas veces se resume la información de los datos como $\bar{X} \pm desvío$ pero no se aclara de cuál de los dos desvíos se trata. Si queremos dar información sobre la dispersión de los datos alrededor de su media muestral, usaremos $\bar{X} \pm S$. Por otro lado, si queremos dar información sobre la precisión con la que $\bar{X}$ estima a μ, usaremos: $\bar{X} \pm \widehat{ES}$.

muestreo de una distribución normal. Veamos ahora algunos resultados relacionados con el hecho de muestrear específicamente a partir de la Normal. Sea $X_1, \ldots, X_n$ una muestra aleatoria (o sea i.i.d.) de una población Normal, o sea, $N\left(\mu, \sigma^2\right)$. Entonces:

a) $\bar{X}$ y S^2 son dos estadísticos independientes entre sí. Esto es bastante sorprendente ya que la media muestral se usa para definir la varianza muestral.

b) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ o bien estandarizando la v.a. normal $\bar{X}$, se tiene: $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. O sea que si X_i es normal entonces también $\bar{X}$ es normal, sin importar el tamaño n de la muestra. Notar que las expresiones siguientes son equivalentes: $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$.

c) $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(v=n-1)}$, la varianza muestral se distribuye como chi-cuadrado con $n-1$ grados de libertad.

Por ejemplo, sea $X_1, X_2, \ldots, X_n$ una m.a. con $E[X_i] = \mu = 4{,}8$, $\sigma_{X_i} = 0{,}5$, $X_i \sim N\left(\mu, \sigma^2\right)$, y el tamaño muestral es $n = 16$. ¿cuál es la probabilidad de que la media de la muestra sea mayor a cinco?. Como cada X_i tiene distribución normal, entonces $\bar{X} \sim N\left(4{,}8, \frac{(0{,}5)^2}{16}\right)$; $EE[\bar{X}] = \sigma_{\bar{X}} = \sqrt{\frac{(0{,}5)^2}{16}} = \frac{0{,}5}{\sqrt{16}} = \frac{0{,}5}{4} = 0{,}125$. Sacamos las probabilidades a partir de la normal estándar.

$$P\left(\bar{X} > 5\right) = P\left(\frac{\bar{X} - 4{,}8}{0{,}125} > \frac{5 - 4{,}8}{0{,}125}\right) = P\left(Z_{\bar{x}} > 1{,}6\right) = 1 - \Phi\left(1{,}6\right) = 1 - 0{,}9452 = 0{,}0548$$

Notar que σ debe ser un parámetro conocido. Si no lo fuera es necesario aproximarlo por su estimador $\widehat{\sigma} = S$, pero dicha aproximación solamente es válida si *el tamaño muestral es grande* ($n \geq 30$).

Observar que a partir de (3) se pueden calcular la esperanza y la varianza de S^2. Si una v.a. X sigue una distribución χ^2_v entonces $E[X] = v$ y su varianza es $V[X] = 2v$. Sabemos que si X_i es normal, $(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(v=n-1)}$. Entonces la esperanza de $(n-1)\frac{S^2}{\sigma^2}$ es $v = n-1$ y su varianza es $2v = 2(n-1)$. Por lo tanto:

$$E\left[(n-1)\frac{S^2}{\sigma^2}\right] = n-1 \Rightarrow E\left[(n-1)\frac{S^2}{\sigma^2}\right] = \frac{(n-1)}{\sigma^2}E\left[S^2\right] = n-1 \Rightarrow E\left[S^2\right] = \frac{n-1}{n-1}\sigma^2 = \sigma^2$$

$$V\left[(n-1)\frac{S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow V\left[(n-1)\frac{S^2}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4}V\left[S^2\right] = 2(n-1) \Rightarrow V\left[S^2\right] = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}$$

distribución t de Student. Una v.a. T tiene distribución t de Student de parámetro v (que es el número de grados de libertad) si su f.d.p. es:

$$f_T(t) = \frac{\Gamma\left(\frac{v-1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \cdot \frac{1}{(v\pi)^{\frac{1}{2}}} \cdot \frac{1}{\left(1 + \frac{t^2}{v}\right)^{\frac{v+1}{2}}}; -\infty < t < \infty$$

La esperanza es $E[T] = 0$ ($v > 1$) y la varianza es $V[T] = \frac{v}{v-2}$ ($v > 2$). La distribución posee $p-1$ momentos porque en $t_{(v=1)}$ no existe la media ni la varianza (de hecho con $v = 1$ la distribución es Cauchy, que tiene esperanza y varianza infinitas); y en $t_{(v=2)}$ no existe la varianza. La distribución es unimodal, simétrica, acampanada y centrada en cero. Cuando aumentan los grados de libertad la distribución se «empina» y se parece cada vez más a la normal estándar. Esto es porque cuando $v \to \infty$, $V[T] \to 1$, como en la normal estándar. La figura que sigue muestra $t_{(1)}$ (Cauchy) en negro, $t_{(5)}$ en naranja, $t_{(20)}$ en verde y una normal estándar $N(0,1)$ en azul. Observar que la distribución t tiene más densidad de probabilidad en las colas respecto de la Normal.

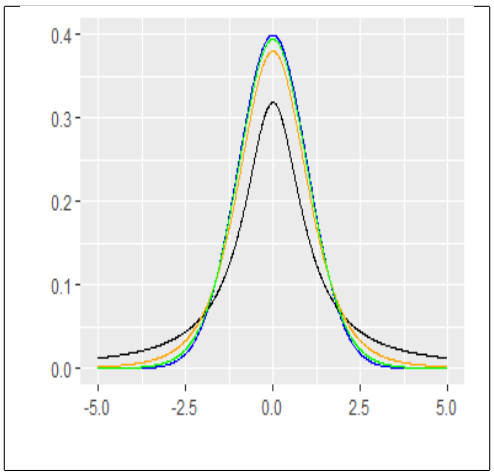


Figura 32: Comparación: Normal estándar (azul); $t_{(5)}$ (naranja), $t_{(20)}$ (verde), Cauchy (negro)

Sean $X \sim N(0,1)$ y $Y \sim \chi^2_{(v)}$ variables aleatorias independientes, entonces el cociente entre X y la raíz de Y dividida por sus grados de libertad tiene distribución t de Student con v grados de libertad:

$$X \sim N(0,1), Y \sim \chi^2_{(v)} \text{ v.a. independientes} \Rightarrow T = \frac{X}{\sqrt{\frac{Y}{v}}} \sim t_{(v)}$$

Definamos: $X = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ e $Y = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)}$, X e Y independientes. Construyamos la variable T como sigue.

$$T = \frac{X}{\sqrt{\frac{Y}{v}}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)\frac{S^2}{\sigma^2}}{(n-1)}}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{(n-1)\frac{S^2}{\sigma^2}\cdot\frac{1}{(n-1)}}} = \frac{\bar{X}-\mu}{\frac{\sqrt{\sigma^2}}{\sqrt{n}}}\cdot\frac{1}{\frac{\sqrt{S^2}}{\sqrt{\sigma^2}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$$

Si $X_1, \ldots, X_n$ es una m.a. de una $N(\mu, \sigma^2)$, sabemos que su media muestral estandarizada se distribuye como $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Reemplazando σ por su estimador $\hat{\sigma} = S$ llegamos a que la media muestral estandarizada se distribuye como t de Student con $v = n-1$ grados de libertad.

$$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$$

Entonces cuando la variable aleatoria X_i tiene distribución normal, el tamaño de muestra n es pequeño ($n < 30$) y σ^2 es desconocido, la media muestral tiene distribución t de Student con $v = n-1$ grados de libertad. Notar que los grados de libertad dependen del tamaño muestral n; y al aumentar este aumentan los grados de libertad, haciendo que, con n grande ($n \geq 30$) la distribución t se casi indistinguible de una Normal estándar (como se vio en la Figura 24). Es decir que con n grande recuperamos el caso anterior.

Por ejemplo, sea $X_1, X_2, \ldots, X_n$ una m.a. con $E[X_i] = \mu = 4$, $S_{X_i} = 2$, $X_i \sim N(\mu, \sigma^2)$, y el tamaño muestral es $n = 9$. ¿cuál es la probabilidad de que la media de la muestra sea mayor a cinco?. Como cada X_i tiene distribución normal, entonces: $T = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)} \Rightarrow T = \frac{\bar{X}-4}{\sqrt{\frac{(2)^2}{9}}} = \frac{\bar{X}-4}{\frac{\sqrt{4}}{\sqrt{9}}} = \frac{\bar{X}-4}{\frac{2}{3}} \sim t_{(v=n-1=9-1=8)}$. Sacamos las probabilidades a partir de una t con $v = 8$ grados de libertad.

$$P(\bar{X} > 5) = P\left(\frac{\bar{X}-4}{\frac{2}{3}} > \frac{5-4}{\frac{2}{3}}\right) = P(T_{(8)} > 1{,}5) = 1 - P(T_{(8)} \leq 1{,}5) = 1 - 0{,}914 = 0{,}086$$

distribución F de Snedecor. Si $X_1, X_2, \ldots, X_{n_1}$es una m.a. (i.i.d.) de una normal $N(\mu_1, \sigma_1^2)$ y $Y_1, Y_2, \ldots, Y_{n_1}$es una m.a. (i.i.d.) de una normal $N(\mu_2, \sigma_2^2)$ estamos interesados en comparar las varianzas de las dos poblaciones, o sea: $\frac{\sigma_1^2}{\sigma_2^2}$. Información de la población está contenida en la muestra, o sea en el ratio $\frac{S_1^2}{S_2^2}$. La distribución F nos permite comparar ambas a través del ratio: $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$, como se explica a continuación.

Sea $V \sim \chi^2_{(p)}$; $W \sim \chi^2_{(q)}$; V y W son independientes. Entonces el cociente entre estas variables divididas por sus grados de libertad sigue una distribución F de Snedecor con p y q grados de libertad.

$$F = \frac{\frac{V}{p}}{\frac{W}{q}} \sim F_{(p,q)}$$

Más específicamente consideremos que $V = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{(p=n_1-1)}$ y $W \sim \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{(q=n_2-1)}$; V y W independientes. Entonces:

$$F = \frac{\frac{V}{p}}{\frac{W}{q}} = \frac{\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}}{(n_1-1)}}{\frac{\frac{(n_2-1)S_2^2}{\sigma_2^2}}{(n_2-1)}} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}\cdot\frac{1}{(n_1-1)}}{\frac{(n_2-1)S_2^2}{\sigma_2^2}\cdot\frac{1}{(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{\sigma_1^2}\cdot\frac{\sigma_2^2}{S_2^2} \sim F_{(p=n_1-1, q=n_2-1)}$$

Notar que si $\sigma_1^2 = \sigma_2^2$, entonces $\frac{S_1^2}{\sigma_1^2}.\frac{\sigma_2^2}{S_2^2} = \frac{S_1^2}{S_2^2}$. En suma un cociente de varianzas tiene distribución F con $p = n_1 - 1$ (numerador) y $q = n_2 - 1$ (denominador) grados de libertad.

Una v.a. tiene distribución $F_{(p,q)}$ si su f.d.p. es:

$$f_X(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{\frac{p}{2}} \frac{x^{\frac{p}{2}-1}}{\left(1+\frac{p}{q}x\right)^{\frac{p+q}{2}}}; 0 < x < \infty$$

La Figura que sigue ilustra la distribución.

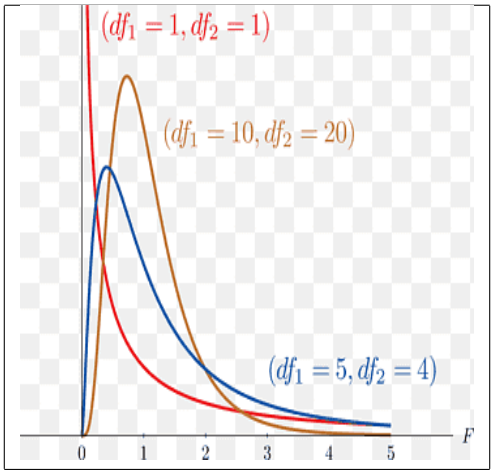


Figura 33: Distribución F de Snedecor.

Algunas relaciones que involucra la distribución F son las siguientes.

a) Si $X \sim F_{p,q}$, entonces $\frac{1}{X} \sim F_{q,p}$

b) Si $X \sim t_{(q)}$, entonces $X^2 \sim F_{1,q}$

Teorema central del límite (T.C.L.). Sea $X_1, X_2, \ldots, X_n$ una m.a. (i.i.d.) de una distribución *cualquiera* con $E[X_i] = \mu$ y $V[X_i] = \sigma^2 > 0$ y sea el estadístico $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (media muestral depende del tamaño de la muestra). El teorema estable que a medida que crece el tamaño muestral la distribución de la media muestral (suponiendo X_i con cualquier distribución) se acerca a una Normal $N\left(\mu, \frac{\sigma^2}{n}\right)$ o bien que la media muestral estandarizada se acerca a una normal estándar $N(0,1)$. La aproximación es válida con $n \geq 30$.

$$\bar{X}_n \xrightarrow[n\to\infty]{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1)$$

Algunas consideraciones sobre el teorema.

- Partiendo de la independencia entre las v.a. x_i y de varianza finita, ¡se llega a la distribución normal! No importa cual sea la distribución de X_i.
- Permite saber la distribución de la media muestral
- Es necesario que σ sea conocida. Sin embargo, via la aplicación del teorema de Slutsky es posible reemplazar σ por S_n, obteniendo que: $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow[n\to\infty]{d} N(0,1)$.
- La aproximación a la Normal que da el T.C.L. depende de la distribución original de las variables aleatorias. Si se parte de una distribución muy asimétrica la tasa de convergencia a la Normal va a ser mayor.
- a medida que crece el tamaño muestral $\bar{X} \to \mu$ (Ley de los grandes números) y $\frac{\sigma^2}{n} \to 0$.

- el teorema se refiere a la media muestral. NO dice que con n grande la distribución de la v.a. X_i se aproxima a una Normal.

Por ejemplo, supongamos que $X_1, X_2, \ldots, X_n$ es una m.a. (i.i.d.) de una distribución *cualquiera* con $E[X_i] = 210$ y $\sigma_{X_i} = 40$ y $n = 64$. Entonces según el T.C.L. $\bar{X} \approx N\left(\mu = 210, \frac{\sigma^2}{n} = \frac{40^2}{64} = 25\right)$ o bien $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} = \sqrt{64}\frac{\bar{X}_{64} - 210}{40} = \frac{\bar{X}_{64} - 210}{\sqrt{\frac{40^2}{64}} = \sqrt{25} = 5} \approx N(0,1)$. Notar el signo $\approx$ que significa que la distribución es *aproximadamente* Normal con n grande.
¿Cuál es la probabilidad de que la media muestral sea al menos de 230? Simplemente calculamos la probabilidad bajo la Normal estándar.

$$P\left(\bar{X} \geq 230\right) = P\left(\frac{\bar{X} - 210}{5} \geq \frac{230 - 210}{5}\right) = P\left(Z_{\bar{x}} \geq 4\right) = 1 - \Phi(4) = 1 - 0{,}9999683 = 3{,}17 \times 10^{-5}$$

Si hubiéramos tenido $n = 16$, en este caso no habríamos podido calcular la probabilidad porque no podemos aplicar el T.C.L.

distribución muestral para la proporción. Sea $X_1, X_2, \ldots, X_n$ m.a. con $X_i \sim Bi(1, p)$; o sea cada X_i es igual a 1 (éxito) o 0 (fracaso). El estimador del la proporción poblacional p es $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Además sabemos que $Y = \sum_{i=1}^{n} X_i \sim Bi(n, p)$, con $E[Y] = np$ y $V[Y] = np(1-p)$. Como es un estadístico, tiene una distribución dada por los valores que toman las muestras de tamaño n extraídas de la población. La esperanza del estadístico es $E[\hat{p}] = p$ y la varianza es $V[\hat{p}] = \frac{p(1-p)}{n}$; ya que:

$$E[\hat{p}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}np = p$$

$$V[\hat{p}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Como la proporción es una media, es decir, $\hat{p} = \bar{X}$; entonces se puede aplicar el teorema central del límite a dicha media (estandarizada) y obtener que:

$$\bar{X}_n \xrightarrow{d} N\left(p, \frac{p(1-p)}{n}\right)$$

$$\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1)$$

Por medio de la aplicación del teorema de de Slutsky es posible reemplazar el parámetro p por su estimador $\hat{p}$, obteniendo que:

$$\bar{X}_n \xrightarrow{d} N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

$$\frac{\bar{X}_n - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{d} N(0,1)$$

Dicha aproximación a la normal es válida si $np \geq 5$ y $n(1-p) \geq 5$ (deben cumplirse ambos requisitos). Por ejemplo, para $p = 0{,}6$ bastarían al menos $n = 15$ datos:
$n = 10 \Rightarrow np = 10 \times 0{,}6 = 6 > 5; n(1-p) = 10 \times 0{,}4 = 4 < 5$
$n = 15 \Rightarrow np = 15 \times 0{,}6 = 9 > 5; n(1-p) = 15 \times 0{,}4 = 6 > 5$
Por ejemplo, sea $X_1, X_2, \ldots, X_n$ m.a. con $X_i \sim Bi(1, p)$; $p = 0{,}85$ y $n = 200$. Calcular la probabilidad de que la proporción muestral sea como mucho 0.8. Como $np = 200(0{,}85) = 170 > 5$ y $n(1-p) = 200(0{,}15) = 30 > 5$, podemos usar la aproximación a la Normal via el T.C.L. Entonces: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}85(1-0{,}85)}{200}} \simeq 0{,}025$

$$P\left(\bar{X} \leq 0{,}8\right) = P\left(\frac{\bar{X} - 0{,}85}{0{,}025} \leq \frac{0{,}8 - 0{,}85}{0{,}025}\right) = P\left(Z_{\bar{x}} \leq -2\right) = \Phi(-2) = 0{,}022$$

corrección de varianzas por población finita. Hasta ahora se ha supuesto que se está muestreando *con reemplazo* de una población infinita (muy grande). Cuando pasa esto estamos muestreando una fracción pequeña de la población. Pero a medida que el tamaño poblacional disminuye y muestreamos *sin reemplazo* una fracción cada vez más grande de observaciones, las observaciones de la muestra no son independientes entre sí. Para corregir esto es necesario corregir las varianzas de los estimadores $\bar{X}$ y $\hat{p}$ con el factor $\frac{N-n}{N-1}$. Si el tamaño n de las muestras es pequeño con respecto al tamaño N de la población, el efecto de esta corrección es despreciable, pues el factor es aproximadamente uno. En la práctica se considera esta situación cuando el tamaño de las muestras es a lo sumo del orden del 5 % del de la población ($n \leq 0{,}05N$). Por ende, es apropiado usar esta corrección cuando se muestrea más del 5 % de la población y esta tiene un tamaño poblacional conocido.

$$V\left[\bar{X}\right] = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$$

$$V\left[\hat{p}\right] = \frac{p\left(1-p\right)}{n}.\frac{N-n}{N-1}$$

resumen. Resumamos la información sobre los estimadores vistos en esta sección en el cuadro que sigue.

θ	$\hat{\theta}$	$E\left[\hat{\theta}\right]$	$V\left[\hat{\theta}\right]$	$Z_{\hat{\theta}} \sim f_{\hat{\theta}}$
μ	$\bar{X}$	$E\left[\bar{X}\right] = \mu$	$V\left[\bar{X}\right] = \frac{\sigma^2}{n}$	$Z_{\bar{x}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \begin{cases} \sim N\left(0,1\right) & ; X_i \sim N\left(\mu,\sigma^2\right); \textit{cualquier } n; \sigma \textit{ conocida} \\ \sim t_{(n-1)} & ; X_i \sim N\left(\mu,\sigma^2\right); n < 30; \sigma \textit{ desconocida} \\ \approx N\left(0,1\right) & ; X_i \sim f_X\left(x\right); n \geq 30; \sigma \textit{ (des)conocida} \end{cases}$
σ^2	S^2	$E\left[S^2\right] = \sigma^2$	$V\left[S^2\right] = \frac{2\sigma^4}{n-1}$	$\left(n-1\right)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad ; X_i \sim N\left(\mu,\sigma^2\right)$
p	$\hat{p}$	$E\left[\hat{p}\right] = p$	$V\left[\hat{p}\right] = \frac{p(1-p)}{n}$	$\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N\left(0,1\right); np \geq 5 \; y \; n\left(1-p\right) \geq 5; p \textit{ (des)conocida}$

Cuadro 19: Estimadores y distribuciones muestrales.

11. Estimación.

En estadística estamos interesados en calcular parámetros poblacionales θ desconocidos; o sea, números desconocidos que caracterizan la distribución de la población. Una muestra contiene información sobre el parámetro. Un estadístico resume la información de la muestra en un determinado valor, a partir de aplicar una función a los datos de la muestra. El valor que da el estadístico estima el valor del parámetro; o sea que el estadístico es un estimador *puntual* $\hat{\theta}$ del parámetro poblacional θ. Desearíamos que dicho valor se acerque al verdadero valor del parámetro lo más posible. Como vimos, los estadísticos (estimadores) son variables aleatorias; porque al cambiar los datos de la muestra, cambia el valor que arroja la función. Por tanto el estimador varía de muestra en muestra, y por ende, es una variable aleatoria que tiene una distribución. En el apartado anterior se mostraron las distribuciones para la media, varianza y proporción muestrales. Así como sus respectivas varianzas y esperanzas. Para que un estimador sea «bueno» debe cumplir ciertas propiedades. Mencionaremos tres de ellas.

11.1. Propiedades de los «buenos» estimadores. (optativo)

11.1.1. Estimador insesgado.

Se dice que un estimador es insesgado cuando la esperanza de la distribución muestral de dicho estimador coincide (aproximadamente) con el verdadero valor del parámetro que se desea estimar; es decir: $E\left(\hat{\theta}\right) = \theta$. Esto significa que, en promedio, los valores posibles que puede asumir el estimador se aproximan al valor que se desea estimar. Vimos en el apartado anterior que $E\left[\bar{X}\right] = \mu$, $E\left[S^2_{n-1}\right] = \sigma^2$ ($S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$) y $E\left[\hat{p}\right] = p$; es decir, que los tres estimadores son insesgados. Por otro lado, también vimos que si usamos el estadístico $S^2_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ para estimar a σ^2, entonces $E\left[S^2_n\right] = \left(\frac{n-1}{n}\right)\sigma^2$; con lo cual no obtenemos un estimador insesgado. Sin embargo los estimadores también pueden ser *asintóticamente* insesgados. Esto quiere decir que, a medida que aumenta el tamaño de la muestra n, la esperanza del estimador se acerca al parámetro que se desea estimar. En el caso de

$E\left[S_n^2\right] = \left(\frac{n-1}{n}\right)\sigma^2$ se puede verificar que cuando $n \to \infty$, $\frac{n-1}{n} \to 1$, y por lo tanto, con n grande el efecto del factor $\frac{n-1}{n}$ es despreciable, con lo cual: $E\left[S_n^2\right] \to \sigma^2$ cuando $n \to \infty$. Entonces decimos que S_n^2 es *asintóticamente insesgado*. El sesgo de un estimador $\hat{\theta}$ es la media o valor esperado de su distribución muestral menos el valor paramétrico θ; o sea: $Sesgo\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta$. De aquí se deduce que si el estimador es insesgado ($E\left(\hat{\theta}\right) = \theta$) su sesgo es cero: $Sesgo\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta = \theta - \theta = 0$.

11.1.2. Mínima varianza.

Si el estimador es insesgado entonces la esperanza (centro) de su distribución muestral coincide con el verdadero parámetro. Por ende, la varianza (o desvío) de dicha distribución muestral es una medida de la dispersión de los valores que puede tomar el estimador alrededor del «verdadero parámetro» (media). Lo que quisiéramos es que los valores que puede tomar el estimador se hallaran alrededor del parámetro; o sea lo más cerca posible del verdadero valor a estimar. Es suma,la varianza es una medida de la concentración de la distribución muestral alrededor del parámetro mismo. Por lo tanto un buen estimador será aquel que tenga varianza de la distribución muestral pequeña. Luego, de entre todos los estimadores insesgados se debe elegir el de menor varianza. Si tenemos que $V\left[\hat{\theta}_1\right] < V\left[\hat{\theta}_2\right]$, decimos que $\hat{\theta}_1$es más «eficiente» que $\hat{\theta}_2$ (si ambos son insesgados). Por ejemplo, si tenemos una muestra aleatoria de tamaño $2n+1$ (impar) proveniente de una distribución normal con n grande, ¿Es mejor estimar el centro de la distribución poblacional con la media muestral $\bar{X}$ o con la mediana muestral $\tilde{X}$? Pues bien, la varianza de $\bar{X}$ es $V\left[\bar{X}\right] = \frac{\sigma^2}{n}$ y la varianza de la mediana es $V\left[\tilde{X}\right] = \frac{\pi\sigma^2}{2n} \approx 1{,}57\frac{\sigma^2}{n}$; entonces $\bar{X}$ es mejor estimador de μ que $\tilde{X}$ (ambos estimadores son insesgados: $E\left[\tilde{X}\right] = \tilde{\mu}$ y $E\left[\bar{X}\right] = \mu$).

11.1.3. Error cuadrático medio.

Un buen estimador es también aquel que minimiza el error cuadrático medio. Se trata de una medida utiliza tanto la información sobre la posición como el de la dispersión de la distribución muestral. Se define como:

$$ECM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right] = V\left[\hat{\theta}\right] + \left(E\left[\hat{\theta}\right] - \theta\right)^2 = V\left[\hat{\theta}\right] + \left[sesgo\left(\hat{\theta}\right)\right]^2$$

En la fórmula, $\left(\hat{\theta} - \theta\right)^2$ es el cuadrado de la distancia entre el parámetro y su estimador; o sea es el error cuadrático. El error cuadrático medio es el promedio de los cuadrados de las distancias entre el estimador y el parámetro. Indica cuánto se pierde en promedio al estimar el parámetro θ con su estimador $\hat{\theta}$. Vemos que dicho promedio es igual a la suma de la varianza del estimador y su sesgo (ambos sumandos ≥ 0). Si un estimador es insesgado, el error cuadrático medio coincide con la varianza del estimador: $ECM\left[\hat{\theta}\right] = V\left[\hat{\theta}\right]$.

Una forma gráfica de ilustrar los conceptos de sesgo y precisión es mediante el famoso ejemplo de el tiro al blanco por parte de arqueros. En la figura siguiente a la izquierda tenemos un arquero «insesgado» porque en promedio le da al blanco; pero es poco preciso porque los tiros tienen alta variabilidad. En el centro tenemos una arquero preciso porque los tiros tienen poca dispersión, al estar concentrados; pero por otra parte hay sesgo porque el promedio de los tiros no coincide con el blanco. Por último, a la derecha tenemos un arquero insesgado y presico: los tiros se concentran alrededor del blanco y en promedio le dan al blanco.

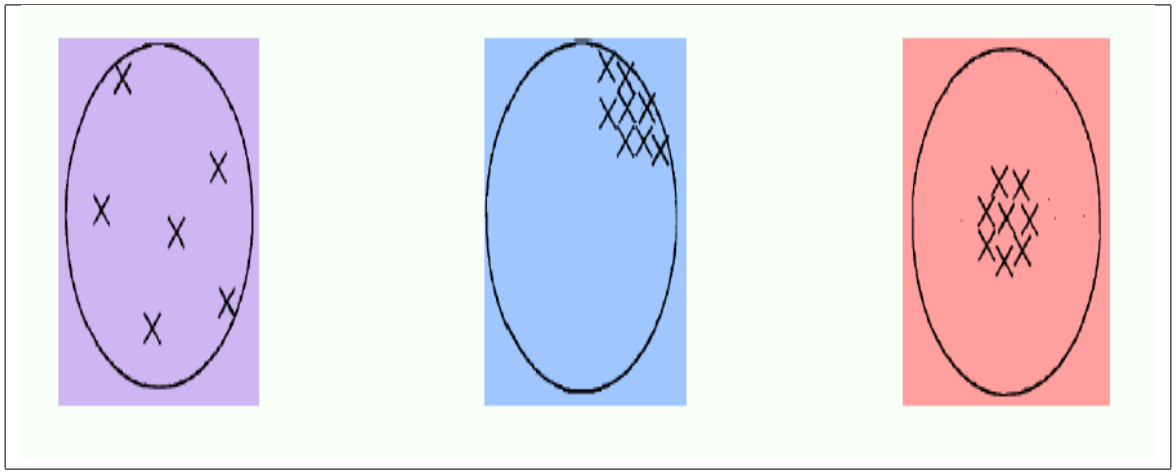


Figura 34: Ejemplo de los arqueros.

11.1.4. Consistencia.

Un estimador es consistente si: (1) es (asintóticamente) insesgado; (2) su varianza tiende a *cero* cuando el tamaño muestral n tiende a infinito.

$\hat{\theta}$ es consistente de θ si $\begin{cases} E\left[\hat{\theta}\right] = \theta \\ V\left[\hat{\theta}\right] \longrightarrow 0 \quad \textit{cuando } n \to \infty \end{cases}$

Por ejemplo, el estimador $\bar{X}$ es consistente pues: (1) $E\left[\bar{X}\right] = \mu$ y (2) $V\left[\bar{X}\right] = \frac{\sigma^2}{n} \to 0$ cuando $n \to 0$.

Vamos a ejemplificar el concepto de consistencia con una pequeña simulación. Supongamos que obtenemos *1000* muestras aleatorias de tamaño n con $n \in \{2,3,5,10,15,20,30\}$ a partir de una $Bi\,(1,0{,}2)$. A cada conjunto de las *1000* muestras de tamaño n le calculamos la media $\hat{p}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ $(X_i \in \{0,1\})$. Así obtenemos una distribución de medias para cada tamaño muestral n. El siguiente cuadro muestra la media (esperanza) y la varianza $(V\left[\bar{X}_n\right] = \frac{\hat{p}(1-\hat{p})}{n})$ de dicha distribución para cada tamaño muestral. Aquí el parámetro poblacional es $p = 0{,}2$. Vemos que $E\left[\bar{X}\right] \simeq 0{,}2$ siempre, entonces $\hat{p}$ es insesgado de p. Por otro lado, se observa que $V\left[\bar{X}\right]$ tiende a cero cuando n crece (la estimación se vuelve cada vez más precisa). Por ende $\hat{p}$ es consistente de p.

n	$E\left[\bar{X}\right]$	$V\left[\bar{X}\right]$
2	0.2	0.08
3	0.206	0.055
5	0.199	0.032
10	0.197	0.016
15	0.198	0.011
20	0.198	0.008
30	0.2	0.005

Cuadro 20: $\hat{p}$ es insesgado de p y su varianza tiende a cero cuando n crece.

11.2. Estimación puntual por máxima verosimilitud. (optativo)

Antes de realizar el experimento el resultado es desconocido. Cuando los parámetros son conocidos, las probabilidades nos permiten predecir un resultado desconocido, por ejemplo, la probabilidad de un determinado valor de una v.a. si esta sigue una distribución binomial. Ahora invirtamos el proceso: al realizar el experimento el resultado se hace conocido: obtenemos datos. Pero ahora no conocemos el valor del parámetro que generó los datos. El método de estimación de parámetros por máxima verosimilitud consiste en determinar cuán verosímil es que un determinado valor de un parámetro haya generado los datos que se observan. Se asignan probabilidades a cada valor posible del parámetro y de éstas se elige el valor del parámetro con mayor probabilidad de haber generado los datos. Dicho de otra forma, en una función de (masa o densidad de) probabilidad los valores de los datos varían y se asignan probabilidades a cada valor posible de los datos, según parámetros que son fijos. En cambio, en una función de

verosimilitud, los datos son fijos y se asignan probabilidades a cada valor posible de un parámetro (o combinación de valores de varios parámetros). El valor del parámetro que se elige es aquel que maximiza la función de verosimilitud.

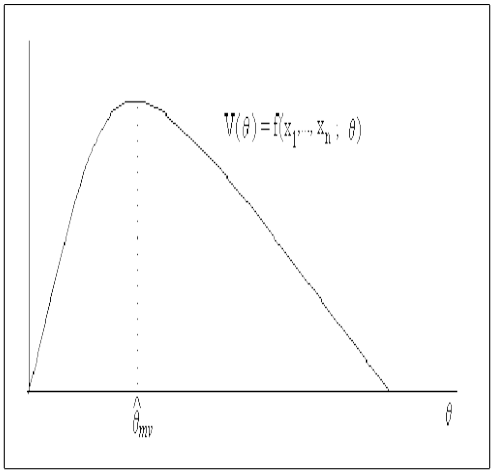


Figura 35: La función de verosimilitud.

Supongamos que tenemos datos que provienen de una distribución Bernoulli $Ber(p)$. Por supuesto, los datos son un conjunto de realizaciones de variables aleatorias que pueden tomar los valores 1 o 0. Supongamos que «1» significa que un sujeto dice correctamente una concordancia. Una vez vistos los datos, nos preguntamos si dicho sujeto está contestando al azar, o sea si la $p(Y=1) = 0{,}5$ o si por otro lado, tiene alta precisión en su respuesta, digamos $p(Y=1) = 0{,}8$. Sean los siguientes datos:

$$y = \underset{12veces}{(1,\ldots,1)} \quad \underset{5veces}{(0,\ldots,0)} \quad \underset{23veces}{(1,\ldots,1)} \quad \underset{8veces}{(0,\ldots,0)} \quad \underset{15veces}{(1,\ldots,1)} \quad \underset{3veces}{(0,\ldots,0)}$$

Queremos comparar $L(0{,}5; y)$, la probabilidad de que los datos vengan de una Bernoulli con parámetro $p = 0{,}5$; con $L(0{,}8; y)$, la probabilidad de que los datos vengan de una Bernoulli con parámetro $p = 0{,}8$. En el primer caso: $p(Y=1) = 0{,}5$ y $p(Y=0) = 0{,}5$. En el segundo caso: $p(Y=1) = 0{,}8$ y $p(Y=0) = 0{,}2$. Como se trata de una muestra aleatoria, cada variable aleatoria es independiente de cualquier otra. Por lo tanto, para calcular la probabilidad de la muestra, o sea la probabilidad conjunta, simplemente multiplicamos las probabilidades de cada realización.

$L(0{,}8; y) = \underset{12veces}{(0{,}8,\ldots,0{,}8)} \quad \underset{5veces}{(0{,}2,\ldots,0{,}2)} \quad \underset{23veces}{(0{,}8,\ldots,0{,}8)} \quad \underset{8veces}{(0{,}2,\ldots,0{,}2)} \quad \underset{15veces}{(0{,}8,\ldots,0{,}8)} \quad \underset{3veces}{(0{,}2,\ldots,0{,}2)} = (0{,}8)^{50}(0{,}2)^{16} = 9{,}35 \times 10^{-17}$

$L(0{,}5; y) = \left(\frac{1}{2}\right)^{50}\left(\frac{1}{2}\right)^{16} = 1{,}35 \times 10^{-20}$

Elegimos el valor del parámetro que maximiza la probabilidad de obtener los datos observados; aquí: $L(0{,}8; y) > L(0{,}5; y)$. Por lo tanto, los datos provienen de una Bernoulli con parámetro $p = 0{,}8$: el sujeto tiene alta precisión en las concordancias.

(optativo). Ahora presentaremos formalmente el método de estimación. El estimador de máxima verosimilitud es el valor $\hat{\theta}$ que maximiza $L(\theta, \boldsymbol{x})$, donde la función L depende de θ y $\boldsymbol{x}$ es la muestra fija. Como la muestra es aleatoria, $L(\theta, \boldsymbol{x})$ se puede factorizar. Un requisito importante es que el soporte de la f.d.p. o f.p.m. no dependa de θ (o sea que θ no sea parte de los valores de la variable aleatoria con masa o densidad positiva).

$$L(\theta; \boldsymbol{x}) = \begin{cases} P(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} P(\theta, x_i) & caso\ discreto \\ f(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} f(\theta, x_i) & caso\ continuo \end{cases}$$

El estimador de máxima verosimilitud es el valor de θ que maximiza dicha función:

$$\hat{\theta}_{MV} = \underset{\theta}{argmax} L(\theta, \boldsymbol{x})$$

Ahora bien, como el logaritmo es una función estrictamente creciente maximizar $L(\theta;\boldsymbol{x})$ es equivalente a maximizar $\mathcal{L}(\theta;\boldsymbol{x}) = logL(\theta;\boldsymbol{x})$. Para maximizar el logaritmo de la función de verosimilitud se deben tomar derivadas parciales e igualar a cero: $\frac{\partial \mathcal{L}(\theta;\boldsymbol{x})}{\partial \theta} = 0$ y obtener el valor de $\theta = \hat{\theta}$. Por último hay que comprobar que realmente se trata de un máximo viendo que el signo de la derivada segunda evaluada en el estimador sea negativo: $\frac{\partial^2 \mathcal{L}(\theta;\boldsymbol{x})}{\partial^2 \theta} \mid_{\theta=\hat{\theta}} < 0$.

Sea $\boldsymbol{x} = (X_1, X_2, \ldots, X_n)$ una m.a. de una Bernoulli $X_i \sim Bi(1,p)$, $P(\theta,\boldsymbol{x}) = \theta^X (1-\theta)^{1-X}$. Entonces planteamos la función de máxima verosimilitud:

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} P(\theta, x_i) = \prod_{i=1}^{n} \theta^{X_i} (1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^{n} X_i} (1-\theta)^{n-\sum_{i=1}^{n} X_i}$$

Luego, tomamos el logaritmo $\mathcal{L}(\theta;\boldsymbol{x}) = logL(\theta;\boldsymbol{x})$:

$$\mathcal{L}(\theta;\boldsymbol{x}) = \sum_{i=1}^{n} x_i log(\theta) + \left(n - \sum_{i=1}^{n} X_i\right) log(1-\theta)$$

Derivamos respecto de θ e igualamos a cero:

$$\frac{\partial \mathcal{L}(\theta;\boldsymbol{x})}{\partial \theta} = \frac{\sum_{i=1}^{n} X_i}{\theta} - \frac{\left(n - \sum_{i=1}^{n} X_i\right)}{1-\theta} = 0$$

multiplicar ambos términos por $\frac{1}{n}$ para obtener $\bar{x}$:

$$\Rightarrow \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\theta} - \frac{\frac{1}{n}\left(n - \sum_{i=1}^{n} X_i\right)}{1-\theta} = 0$$

$$\Rightarrow \frac{\bar{X}}{\theta} - \frac{1-\bar{X}}{1-\theta} = 0 \Rightarrow \frac{(1-\theta)\bar{X} - \theta\left(1-\bar{X}\right)}{\theta(1-\theta)} = 0$$

$$\Rightarrow \bar{X} - \bar{X}\theta - \theta + \bar{X}\theta = 0 \Rightarrow \bar{X} - \theta = 0 \Rightarrow \hat{\theta}_{MV} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Como ya habíamos visto para la proporción muestral[8].

11.3. Estimación por intervalo.

La estimación por intervalo nos habla de la precisión o exactitud de un estimador. La distribución de muestreo nos indica la probabilidad de que un estimador caiga a cierta distancia del parámetro (en la versión estandarizada). Sería interesante obtener un intervalo de valores estimados con una probabilidad conocida de cubrir el parámetro buscado; lo que llamaremos intervalo de confianza. Entonces *el intervalo de confianza de un parámetro* es el rango de valores dentro del cual se cree que se encuentra el parámetro. Por otra parte, *el nivel de confianza* es la probabilidad de que el intervalo de confianza contenga al parámetro. Es un número cercano a 1, por ej., 0,95 o 0,99.

Si $\hat{\theta}$ es estimador del parámetro θ se desea determinar un intervalo de la forma $[a,b]$ tal que este contenga al parámetro θ con probabilidad $1-\alpha$, siendo α el margen de error (un número pequeño, como 0,05 o 0,01). Entonces:

$$P\left(a = \hat{\theta}_1 \leq \theta \leq b = \hat{\theta}_2\right) = 1-\alpha$$

$$P(\theta \notin [a,b]) = \alpha$$

Donde a y b son los límites del intervalo, α es el margen de error y $1-\alpha$ es el nivel de confianza, es decir la probabilidad de que el intervalo de confianza (IC) contenga al verdadero valor (o sea, al parámetro). Por ejemplo,

[8] Para verificar que se trata de un máximo:

$$\frac{\partial^2 \mathcal{L}(\theta;\boldsymbol{x})}{\partial^2 \theta} \mid_{\theta=\bar{x}} = -\frac{\bar{X}}{\theta^2} - \frac{(1-\bar{X})}{(1-\theta)^2} \mid_{\theta=\bar{x}} \Rightarrow -\frac{\bar{X}}{\bar{x}^2} - \frac{(1-\bar{X})}{\left(1-\bar{X}\right)^2} = -\frac{1}{\bar{X}} - \frac{1}{(1-\bar{X})} < 0$$

un nivel de confianza de $1-\alpha = 1-0{,}05 = 0{,}95$ (95 %) nos indica que si repitiéramos el experimento y tomáramos muestras para estimar el parámetro poblacional, en 95 de cada 100 veces el intervalo contendrá a θ y esto no sucederá en cinco veces (o sea 5 % de la veces cometeremos un error de estimación). Lo que NO significa es que hay un 95 % de probabilidad de que θ esté en el intervalo porque θ es un número desconocido, no una variable aleatoria y, por ende, no se le puede asociar una probabilidad. 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 $1-\alpha$". Sin embargo, una vez que se construye el intervalo ya no tiene sentido hablar de probabilidad porque el intervalo es un segmento fijo; entonces, hablamos de la «confianza» de que el intervalo contenga a θ. Dicha confianza se sigue del método de construcción de los intervalos, que asegura que el $(1-\alpha)\times 100$ por ciento de las muestras producirán intervalos de valores de $\hat{\theta}$ que contienen a θ.

11.3.1. IC para μ cuando σ^2 es conocido.

Los intervalos de confianza se construyen siguiendo el método del pivote, que se ilustra a continuación. Sea $X_1, \ldots, X_n$ una muestra aleatoria (o sea i.i.d.) de una población Normal $N\left(\mu, \sigma^2\right)$, con σ^2 conocido. Sabemos que el estimador de $\theta = \mu$ es $\hat{\theta} = \bar{X}$ y que su distribución de muestreo es Normal $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. El primer paso es construir un «pivote»; esto es, una función de una variable aleatoria que depende de la muestra y del parámetro, pero cuya función de distribución no depende del parámetro a estimar. Esto lo logramos simplemente estandarizando la variable aleatoria :

$$Z_{\bar{x}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Notar que Z es una función de la muestra n y del parámetro μ pero la normal estándar no tiene esperanza μ, sino cero, por lo cual, ya no depende del parámetro a estimar μ. El segundo paso es elegir dos constantes a y b tales que $P(a \leq Z_{\bar{x}} \leq b) = 1-\alpha$. Si $Z_{\bar{x}} \sim N(0,1)$ podemos elegir el percentil z_α como el valor que deja un área bajo la curva α hacia la derecha, o sea: $P(Z \geq z_\alpha) = \alpha$. Como la distribución Normal estándar es simétrica elegimos los extremos $a = -z_{\frac{\alpha}{2}}$ y $b = z_{\frac{\alpha}{2}}$. Los valores de $z_{\frac{\alpha}{2}}$ más usados son:

$(1-\alpha)\times 100$	α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
90 %	0.1	0.05	1.645
95 %	0.05	0.025	1.96
99 %	0.01	0.005	2.576

Cuadro 21: Los valores de $z_{\frac{\alpha}{2}}$ más usados.

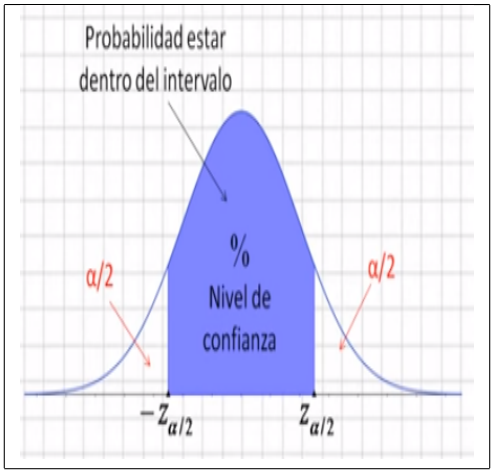


Figura 36: Probabilidad de que el I.C. contenga a θ.

Por último despejamos al parámetro μ.

$$P\left(-z_{\frac{\alpha}{2}} \leq Z_{\bar{x}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

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

Ahora restamos $\bar{X}$ a cada término y multiplicamos por -1, con lo cual:

$$P\left(\bar{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Entonces: $IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$.

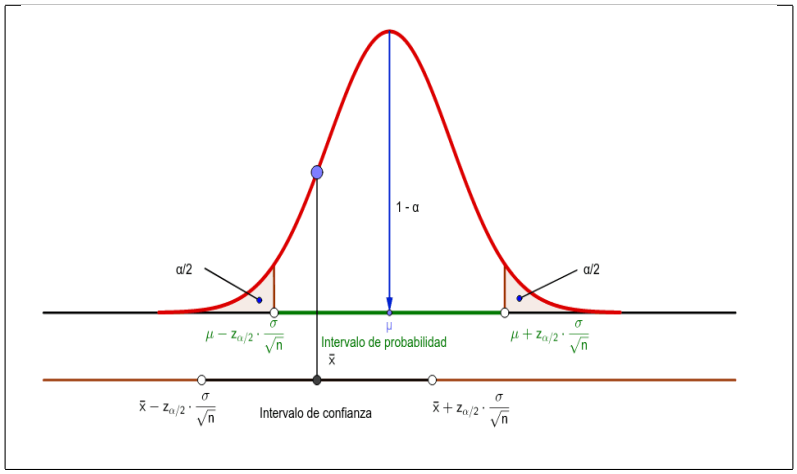


Figura 37: IC para μ.

Hagamos una simulación para visualizar lo dicho. Supongamos una m.a. de una Normal de parámetros $\mu = 0$ y $\sigma = 2$. Tomamos una muestra de $n = 5$, lo cual nos da la m.a.: $\{-1{,}12, -0{,}46, 3{,}11, 0{,}14, 0{,}25\}$, con $\bar{X} = 0{,}38$. Y calculamos el IC para μ:

$$IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = \left[0{,}38 \pm 1{,}96 \times \frac{2}{\sqrt{5}}\right] = [-1{,}36, 2{,}14]$$

Observamos que el IC contiene a $\mu = 0$. Ahora repetimos la operación tomando *100* muestras de $n = 5$ y calculamos para cada muestra el IC. Podemos visualizar los intervalos en el gráfico que sigue. Si se presta atención se verá que solamente tres intervalos no contienen al valor $\mu = 0$.

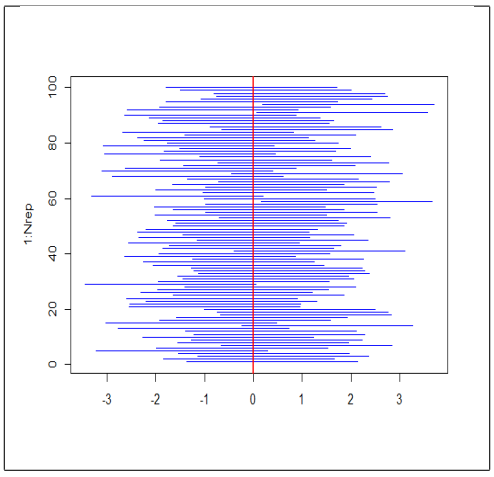


Figura 38: IC en 100 muestras de $N(0, 4)$ con σ conocida.

La precisión de la estimación depende de la longitud del IC, que es:

$$L_o = 2z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Notar que la longitud aumenta (estimación menos precisa) si aumentan $z_{\frac{\alpha}{2}}$ o σ y esta disminuye (estimación más precisa) si aumenta n. Por otra parte, observar que a un determinado n el IC presentado antes tiene siempre la misma longitud porque el percentil $z_{\frac{\alpha}{2}}$ siempre es el mismo.

Ahora bien σ es una característica de la población, pero el coeficiente de confianza $1-\alpha$ y el tamaño muestral lo define el investigador. Podemos calcular el tamaño muestral necesario para obtener un intervalo de nivel $1-\alpha$ para μ cuya longitud no supere una determinada longitud L_o.

$$n \geq \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{L_o}\right)^2$$

Si σ no se conoce se debe tomar una muestra piloto para estimarlo mediante S. Por ejemplo si $S = 14$, $\alpha = 0{,}05$ y $L_o = 4$, obtendríamos: $n \geq \left(\frac{2(1{,}96)(14)}{4}\right)^2 = 188$.

Por otra parte sabemos que si la v.a. no sigue una distribución normal y si n es grande ($n \geq 30$), el T.C.L. nos habilita a aproximar: $Z_{\bar{x}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \approx N(0,1)$; por lo tanto llegaríamos a la misma expresión para el IC para μ. Sin embargo en este caso el IC es de *nivel asintótico* $1-\alpha$ porque dicha confianza se alcanza con n grande: $IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$.

11.3.2. IC para μ cuando σ^2 es desconocido.

Hemos visto que cuando la m.a. proviene de una Normal y n es chico entonces $Z_{\bar{x}} = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$. Bajo estas condiciones el IC para μ de nivel $1-\alpha$ es: $IC(\mu) = \left[\bar{X} \pm t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$.

Pero ahora la longitud del intervalo cambia según el n muestral porque el percentil $t_{n-1,\frac{\alpha}{2}}$ depende de n; por ejemplo si $n = 10$, $t_{9,0,025} = 2{,}26$ pero $z_{0,025} = 1{,}96$.

Supongamos como ejemplo que tenemos una m.a. proveniente de una $N(\mu, \sigma^2)$. Se toma una muestra de tamaño $n = 5$ con $\bar{X} = 99$ y $S = 15$. Para un IC de nivel $1-\alpha$, $\alpha = 0{,}05$ ($\frac{\alpha}{2} = 0{,}025$) el percentil de la distribución muestral es $t_{4,0,025} = 2{,}776$. Entonces:

$$IC(\mu) = \left[\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] = \left[99 \pm 2{,}776 \times \frac{15}{\sqrt{5}}\right] = [80{,}37, 117{,}62]$$

Hagamos la misma simulación de antes pero sin suponer varianza conocida. La figura que sigue ilustra los intervalos sobre *100* muestras de tamaño 5. Si se presta atención se verá que solamente cuatro intervalos no contienen al valor $\mu = 0$. Observar también que los intervalos no tienen la misma longitud (porque S cambia de valor con cada muestra).

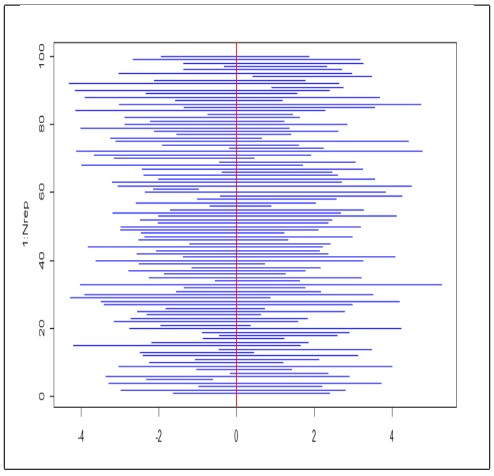


Figura 39: IC en 100 muestras de $N(0,4)$ con σ desconocida.

Si la distribución no es Normal y para n grande, el T.C.L. (y aplicando el teorema de Slutsky) nos autoriza a reemplazar σ por S, y poder aproximar: $Z_{\bar{x}} = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \approx N(0,1)$. Ahora el IC de nivel asintótico $1-\alpha$ queda: $IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$.

El cuadro que sigue ilustra las consideraciones hechas hasta ahora sobre el IC para μ.

σ conocido	$X_i \sim N(\mu, \sigma^2)$	$IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$	nivel exacto $1-\alpha$	cualquier n
	$X_i \sim f_X(x)$	$IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$	nivel asintótico $1-\alpha$	n grande
σ desconocido	$X_i \sim N(\mu, \sigma^2)$	$IC(\mu) = \left[\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$	nivel exacto $1-\alpha$	n pequeño
	$X_i \sim f_X(x)$	$IC(\mu) = \left[\bar{X} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$	nivel asintótico $1-\alpha$	n grande

Cuadro 22: IC para μ de nivel $1-\alpha$

11.4. IC para la proporción.

Sabemos ya que $\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$ si $np \geq 5$ y $n(1-p) \geq 5$. Además el teorema de Slutsky nos permite reemplazar p por $\hat{p}$. El intervalo de confianza para la proporción poblacional p es de nivel asintótico $1-\alpha$ y es el que sigue (recordar que $\hat{p} = \bar{X}_n$):

$$IC(p) = \left[\bar{X}_n \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right]$$

Por ejemplo si $\hat{p} = \frac{26}{200} = 0{,}13$; entonces:

$IC(p) = \left[\bar{X}_n \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right] = \left[0{,}13 \pm 1{,}96\sqrt{\frac{0{,}13(1-0{,}13)}{200}}\right] = [0{,}083; 0{,}177]$

La longitud del intervalo es: $L_o = 2z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}$. Como antes, podemos calcular el tamaño muestral necesario para obtener un intervalo de nivel $1-\alpha$ para p cuya longitud no supere una determinada longitud L_o, haciendo:

$$n \geq \left(\frac{2z_{\frac{\alpha}{2}}}{L_o}\right)^2 \bar{X}_n\left(1-\bar{X}_n\right)$$

Si no se conoce $\hat{p}$ es mejor tomar una muestra piloto para determinarlo. En cualquier caso, siempre es posible el valor más alto de la función $p(1-p)$ que se alcanza en $p = \frac{1}{2}$. Esto producirá el intervalo de amplitud máxima. En dicho caso: $n \geq \left(\frac{z_{\frac{\alpha}{2}}}{L_o}\right)^2$.

Por ejemplo supongamos que en una muestra piloto de $n = 26$ se obtiene $\hat{p} = 0{,}32$ y queremos $L_o = 0{,}12$ (12 %); entonces:

$$n \geq \left(\frac{2(1{,}96)}{0{,}12}\right)^2 0{,}32(0{,}68) = 232$$

Aumentar el nivel de confianza casi duplica n: $n \geq \left(\frac{2(2{,}58)}{0{,}12}\right)^2 0{,}32(0{,}68) = 402$

Por otro lado, disminuir la longitud a la mitad ¡cuadruplica el tamaño muestral necesario!.

$$n \geq \left(\frac{2(1{,}96)}{0{,}06}\right)^2 0{,}32(0{,}68) = 928$$

Por ultimo, si no conociéramos $\hat{p}$ y usamos en su lugar $\hat{p} = \frac{1}{2}$, obtenemos: $n \geq \left(\frac{1{,}96}{0{,}12}\right)^2 = 266$. Notar que no aumenta mucho el tamaño muestral respecto del primer caso porque $\hat{p} = 0{,}32$ ya está cerca de $\frac{1}{2}$.

11.5. IC para la varianza.

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población Normal $N\left(\mu, \sigma^2\right)$ y σ es desconocida.

a) Si μ es conocida entonces un IC de nivel $1-\alpha$ para σ^2 es: $\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right]$.

b) Si μ es desconocida entonces un IC de nivel $1-\alpha$ para σ^2 es: $\left[\frac{\sum_{i=1}^{n}(X_i-\bar{x}_n)^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n}(X_i-\bar{x}_n)^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]$.

Observar que, como la distribución χ^2 no es simétrica ahora los percentiles deben ser $\frac{\alpha}{2}$ a izquierda y $1-\frac{\alpha}{2}$ a derecha.

12. Muestreo simple al azar. (optativo)

Todos los métodos que veremos en el curso suponen que los individuos de la muestra han sido muestreados a partir de poblaciones infinitas, pero mediante muestreo simple al azar. Este tipo de muestreo garantiza la independencia en la muestra, lo cual es requisito fundamental en las muestras aleatorias. Esta sección introduce el método de muestreo simple al azar para poblaciones finitas.

12.1. Conceptos básicos.

Esta sección introduce nociones básicas de muestreo para poblaciones finitas. Definamos primero algunos conceptos básicos.

Población (inferencial). Es el conjunto de todas las unidades de interés hacia donde los resultados del estudio se deberán extrapolar. Es la población inferencial para la que idealmente se quiere sacar conclusiones. Por ejemplo: la población de los ensayos de todos los estudiantes de español de argentina.

Muestra. Subconjunto de la población que da información sobre toda la población. Debería ser un subconjunto que sea representativo de la población, permita proyectar resultados a la población y tenga poco error.

Población objetivo. Es la población resultante luego de excluir elementos de la población inferencial. Por ejemplo, la población de los ensayos de estudiantes de español del instituto de idiomas de la facultad de FyL de la UBA mayores de edad.

Población del marco. El marco muestral es el conjunto de los elementos que sirven para para localizar y acceder a cada uno de los elementos de la población objetivo. No debería incluir omisiones ni repeticiones. De aquí se va a seleccionar la muestra. Por ejemplo, una lista de todos los alumnos estudiantes de español mayores de edad del instituto de idiomas de la facultad de FyL de la UBA, con información auxiliar (domicilio, número de teléfono, curso, etc.). Idealmente la población del marco debería coincidir con la población objetivo.

Muestra actual. Es aquella formada por los individuos que responden la encuesta (o participan del estudio).

Problemas con el marco muestral. Muchas veces una parte de la población objetivo no forma parte dela marco muestral, o sea hay sub-cobertura. O bien el marco muestral puede sufrir de sobre-cobertura: hay más elementos en el marco que en la población objetivo (por ejemplo el marco está formado el listado de todos los alumnos pero se está interesado en la población de alumnos orientales). Puede haber también duplicación de elementos o bien información (auxiliar) incorrecta o desactualizada.

Tipos de muestreo. El muestreo pueden ser ser:

a) Probabilístico: Todos los elementos de la población tienen una probabilidad positiva de ser seleccionados para formar parte de la muestra. Las estimaciones que se obtienen tienen un margen de error. El error de muestreo es cuantificable. Utiliza procedimientos de selección aleatorias para las unidades de la población que reducen al máximo la arbitrariedad y diferentes tipos de sesgos. Permite emplear la inferencia estadística para proyectar los resultados a la población objetivo. Algunos tipos de muestreo probabilístico son: muestreo aleatorio simple, muestreo sistemático, muestreo estratificado, muestreo por conglomerados, muestreo multi-etápico, muestreo proporcional al tamaño, etc.

b) No Probabilístico: El método de selección no usa la teoría de probabilidades. No se conoce la probabilidad de que una unidad sea seleccionada en la muestra. Los mecanismos de selección no usan (o u san mal) la probabilidad. No se puede cuantificar la precisión en términos probabilístico. No obliga a tener un marco de muestreo para la selección. No garantiza muestras representativas. Algunos tipos de muestreo no probabilístico son: muestreo por cuotas (encuestar hasta alcanzar una cota especificada de cantidad de gente); muestreo de voluntarios (encuestar a los que se ofrezcan a participar); muestreo por conveniencia (encuestar a los que me parece más fácil llegar: amigos, conocidos, etc.); muestreo de «bola de nieve» (encuestar a una persona y que esta me indique a otra persona a quien encuestar), etc.

Error de muestreo. Se produce por trabajar con una muestra y no con toda la población. Si se desea conocer una cantidad poblacional θ, a partir de la muestra se obtendrá un valor aproximado $\hat{\theta}$, por lo general diferente del valor poblacional. Esa diferencia es el error de muestreo, y en el muestreo probabilístico, se podrá dar una medida del mismo. Este tipo de error disminuye al aumentar el tamaño muestral. Cada tipo de diseño probabilístico tiene una fórmula para calcularlo.

Errores ajenos al muestreo. Estos incluyen errores por: no respuesta, encuestas mal diseñadas (preguntas que inducen a una determinada respuesta); errores de medición (respuestas falsas, preguntas no comprensibles, las personas no recuerdan, fallas en el instrumento de medición), errores de cobertura (el marco muestral no coincide con la población objetivo).

12.2. Probabilidades de selección e inclusión.

Definamos la población finita $U = \{y_1, y_2, \ldots, y_N\}$ con N elementos. Sea S el conjunto de todas las muestras posibles (de tamaño muestral n) de la población U de tamaño poblacional N; o sea $S = \{s_1, s_2, s_3, \ldots\}$. Para conocer cuantas muestras posibles de tamaño n podemos extraer de la población U de tamaño N aplicamos el combinatorio:

$$C_{N,n} = \left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!\,(N-n)!}$$

Por ejemplo, sea la población de cuatro elementos $U = \{a, b, c, d\}$; $N = 4$. Podemos extraer seis muestras de tamaño dos:

$$C_{4,2} = \left(\begin{array}{c} 4 \\ 2 \end{array} \right) = \frac{4!}{2!\,(4-2)!} = \frac{4!}{2!\,(2)!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)\,(2 \times 1)} = \frac{24}{4} = 6$$

Son las siguientes: $s_1 = \{a, b\}, s_2 = \{a, c\}, s_3 = \{a, d\}, s_4 = \{b, c\}, s_5 = \{b, d\}, s_6 = \{c, d\}$.

En el muestreo probabilístico cada muestra $s_i \in S$ tiene una probabilidad $P(s_i)$ de ser seleccionada; tal que : $P(s_i) \geq 0$ y $\sum_{k=1}^{K} P(s_i) = 1$. Sin embargo las probabilidades de selección pueden seguir un diseño uniforme, en el cual $P(s_i) = \frac{1}{\#muestras}$; o no uniforme, como se muestra a continuación:

s_k	tipo	s_1	s_2	s_3	s_4	s_5	s_6
$P(s_k)$	uniforme	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
	no uniforme	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$

Como cada muestra posible tiene una probabilidad $P(s_i)$ de ser elegida, podemos calcular la probabilidad de cada unidad de la población de aparecer en la muestra seleccionada como:

$$\pi_k = P\left(\textit{probabilidad de la unidad } k \textit{ en una muestra } s\right) = \sum_{s \in S, k \in s} P\left(s\right)$$

O sea que sumamos las probabilidades de las muestras en las cuales la unidad es miembro. Por ejemplo:

$$\textit{diseño uniforme} \rightarrow \pi_a = P\left(s_1\right) + P\left(s_2\right) + P\left(s_3\right) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

$$\textit{diseño no uniforme} \rightarrow \pi_a = P\left(s_1\right) + P\left(s_2\right) + P\left(s_3\right) = \frac{1}{6} + \frac{1}{4} + \frac{1}{6} = \frac{7}{12}$$

Veamos otro ejemplo. Sea $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Y sea la variable y_k con índice k-ésimo:

k	1	2	3	4	5	6	7	8
y_k	1	2	4	4	7	7	7	8

Por ejemplo una posible muestra de tamaño 4, con los *índices* $\{1, 2, 3, 4\}$ sería aquella con los valores $\{1, 2, 4, 4\}$. La cantidad de muestras de tamaño 4 que podemos extraer de la población de tamaño 8 es:

$$C_{8,4} = \left(\begin{array}{c} 8 \\ 2 \end{array} \right) = \frac{8!}{4!\left(8-4\right)!} = 70$$

Bajo un diseño *uniforme* la probabilidad de selección de cada muestra es $P\left(s_i\right) = \frac{1}{70}$ y la probabilidad de inclusión del elemento k-ésimo en una muestra es $\pi_k = \frac{n}{N} = \frac{4}{8} = \frac{1}{2}$ (porque cada elemento de la población está presente en la mitad de las muestras).

12.3. Muestreo simple al azar.

Se trata del tipo de muestreo más simple de todos. Todas las muestras tienen la misma probabilidad de ser seleccionadas:

$$P\left(s_i\right) = \begin{cases} \frac{1}{\left(\begin{array}{c} N \\ n \end{array} \right)} & \textit{si s es de tamaño n} \\ 0 & \textit{en otro caso} \end{cases}$$

Cada elemento de la población tiene la misma probabilidad π_k de ser incluido en la muestra.

$$\pi_k = f = \frac{n}{N}$$

f es la «fracción de muestreo»: nos indica la proporción de la muestra respecto del tamaño poblacional. Por otra parte la inversa de la probabilidad de inclusión indica el «peso» de cada individuo en la población y permite expandir valores a totales poblacionales:

$$w_k = \frac{1}{\pi_k} = \frac{N}{n}$$

Notar que en este tipo de diseño estamos diciendo que todos los individuos «pesan» igual en la población. No serviría para poblaciones heterogéneas; por ejemplo una población de escuelas con cantidades muy diferentes de alumnos. Por otra parte este diseño garantiza la independencia de la muestra obtenida, requisito fundamental para las muestras aleatorias.

Tenemos los siguientes parámetros en la población finita U:

- Total poblacional: $t = \sum_{i=1}^{N} y_i$.
- Media poblacional: $\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i$.

- Varianza poblacional: $S^2 = \frac{1}{N-1} \sum_{i=1}^{n} (y_i - \bar{y}_U)^2$.
- Proporción poblacional: $P = \frac{1}{N} \sum_{i=1}^{N} y_i; y_i \in \{0,1\}$

El siguiente cuadro muestra los estimadores *insesgados* bajo el presente diseño muestral. Los estimadores de la media y la proporción son los que ya conocemos. Los estimadores de las varianzas de las distribuciones muestrales de los estimadores están corregidas por población finita según el factor: $1 - f = 1 - \frac{n}{N}$. Notar que cuando $f \to 1$ entonces $(1 - f) \to 0$ y entonces disminuye el error en la estimación. Cuando $f = \frac{n}{N} < 0{,}05$ el efecto de la corrección por población finita es despreciable.

estimador del total poblacional	$\hat{t} = N \sum_{i=1}^{n} \frac{y_i}{n} = N\bar{y}$
estimador de la varianza del total	$\hat{V}(\hat{t}) = N^2 \left(\frac{1-f}{n}\right) s^2$
estimador de la media poblacional	$\bar{y}_S = \frac{1}{n} \sum_{i=1}^{n} y_i$
estimador de la varianza de la media	$\hat{V}(\bar{y}_S) = \left(\frac{1-f}{n}\right) s^2$
estimador de la proporción poblacional	$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i; y_i \in \{0,1\}$
estimador de la varianza de la proporción	$\hat{V}(\hat{p}) = (1-f) \frac{\hat{p}(1-\hat{p})}{n}$

Cuadro 23: Estimadores insesgados bajo muestreo aleatorio simple.

En lo que atañe al tamaño muestral, se usa la siguiente fórmula cuando $\frac{n_o}{N}$ no es despreciable.

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

para la estimación de la media poblacional se usa: $n_o = \frac{z^2_{\frac{\alpha}{2}} s^2}{c^2}$; y para la estimación de la proporción poblacional se usa: $n_o = \frac{z^2_{\frac{\alpha}{2}} p(1-p)}{c^2}$. En ambas expresiones c es el margen de error máximo que el investigador está dispuesto a cometer. Por ejemplo, en el caso de la proporción si se usa: $z_{\frac{\alpha}{2}} = 1{,}96$, $p = 0{,}5$ y $c = 0{,}05$ (5 % de error), $N = 180$, entonces:

$$n_o = \frac{(1{,}96)^2 (0{,}5)(0{,}5)}{(0{,}05)^2} = 384 \Rightarrow n = \frac{n_o}{1 + \frac{n_o}{N}} = \frac{384}{1 + \frac{384}{180}} = 122$$

13. Comentarios bibliográficos.

Para la parte de estadística descriptiva, gráficos y preprocesamiento se siguió a Chan & col. (2018); también ver: Tan & col. (2005). Para la parte de probabilidades, consúltese Levy (2012) y Casella & Berguer (2002); también consultar a este último para la parte de distribuciones muestrales y estimación, así como a Freund (2014). Para la parte de muestreo ver Lohr (1999). Para las aplicaciones al lenguaje se usaron: Levshina (2015; caps. 1 - 4); Baayen (2008; caps. 1 - 3); Manning & Schütze (1999; caps. 6 y 11).

Referencias

Baayen, R. Harald (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Bresnan, Joan, Tatiana Nikitina Anna Cueni y R Harald Baayen (2007). "Predicting the Dative Alternation". En: *Cognitive Foundations of Interpretation*. Ed. por G. Bouma qnd I. Kraemer y J. Zwarts. Royal Netherlands Academy of Science, págs. 69-94.

Casella, George y Roger L. Berger (2002). *Statistical Inference*. Duxbury Resource Center.

Chan, Debora, Cristina Badano y Andrea Rey (2018). *Análisis inteligente de datos con lenguaje R*. Editorial de la Universidad Tecnológica Nacional.

Davis, Colin J. y Manuel Perea (2005). "BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish". En: *Behavior Research Methods* 37.4, págs. 665-671.

Enders, Craig K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

Freund, John E., Irwin Miller y Marylees Miller (2014). *Mathematical statistics with applications*. Pearson.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model." En: *Proceedings of the second meeting of the north American chapter of the association for computational linguistics and language technologies*, págs. 1-8.

Levshina, Natalia (2015). *How to do Linguistics with R*. John Benjamins.

Levy, Roger (2008). "Expectation-based syntactic comprehension". En: *Cognition* 106.3, págs. 1126-1177.

— (2012). "Probabilistic Models in the Study of Language".

Lohr, Sharon L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Manning, Chris e Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Tan, Pang-Ning, Michael Steinbach y Vipin Kumar (2005). *Introduction to Data Mining*. Pearson/Addison-Wesley.

Van Buuren, Stef y Karin Groothuis-Oudshoorn (2011). "mice: Multivariate imputation by chained equations in R". En: *Journal of Statistical Software* 45.3, págs. 1-67.