Estadística

Ingeniería Técnica en Informática de Sistemas

Manuel Febrero Bande
Pedro Galeano San Miguel
Julio González Díaz
Beatriz Pateiro López

Estadística
Ingeniería Técnica en Informática de Sistemas

Manuel Febrero Bande
Pedro Galeano San Miguel
Julio González Díaz
Beatriz Pateiro López

Estadística

ISBN-13: 978-84-691-0974-8
DL: C 351-2008

Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela

Prólogo

Esta publicación que tienes entre manos no es más que una guía rápida de los conocimientos que se explican en la materia Estatística de la titulación de Ingeniería Informática de Sistemas que se imparte en la Universidad de Santiago de Compostela. Como tal guía rápida no pretende ser exhaustiva sino más bien concreta y ha sido el fruto conjunto de varios miembros del departamento de Estadística e Investigación Operativa, alguno de los cuales se estrenaron en la docencia con estos contenidos. Estos han sido sobre todo compañeros y todos ellos tienen mi agradecimiento.

La Estadística debe desarrollar en el alumno el pensamiento estocástico y la modelización de problemas reales. En muchos campos de la ciencia, y la informática no es una excepción, se deben tomar decisiones en muchos casos en contextos de incertidumbre. Estas decisiones involucran procesos previos como obtención de la máxima información posible, determinación de los focos de error o incertidumbre y modelización de las situaciones estocásticas. La Estadística pretende sentar los cimientos para un análisis pormenorizado de la información disponible, para separar el grano (información) de la paja (ruido) para obtener conclusiones interesantes. Un informático será capaz de almacenar un montón de información pero esta información no será más que basura en el disco si no se le encuentra un sentido. Para ayudarnos en esta tarea, disponemos de una herramienta magnífica y gratuita: el entorno R (www.r-project.org). Esta herramienta democratiza el acceso al cálculo estadístico permitiendo con un bajo consumo de recursos e independientemente de la plataforma obtener imponentes resultados científicos antes sólo al alcance de caras licencias de software. Los ejemplos se han desarrollado en este entorno.

Alguna vez he comparado el proceso estadístico con el proceso de obtener una foto que sirva de titular de un periódico dado que el resultado del proceso estadístico es resumir de manera efectiva una situación como una fotografía resume un instante. Para obtener una buena foto son necesarios tres elementos fundamentales: un motivo que deba ser fotografiado, una cámara de fotos y un fotógrafo. El motivo que debe ser fotografiado es para el estadístico su objeto de estudio y como en el caso de la fotografía, el fotógrafo no tiene el control sobre la escena que quiere resumir pero si debe dedicarle un instante a analizarla, comprenderla y descubrir que quiere obtener de ella. El segundo elemento es la cámara. El fotógrafo debe ser capaz de manejar apropiadamente la cámara para obtener la foto que desea. Por ejemplo, no dominar el foco de la cámara o usar una configuración de estático para fotografiar a un atleta en movimiento sólo provocará la

obtención de una imagen borrosa. En el proceso estadístico la cámara es la técnica que se debe dominar para saber cuáles son sus limitaciones y cuáles sus ventajas. Esta técnica involucra al aparataje matemático que es necesario conocer y dominar. Finalmente, el tercer elemento es el fotógrafo. Éste debe decidir, por ejemplo, sobre el encuadre de la foto o el nivel de detalle que desea así como un estadístico debe decidir cuál va a ser su marco de estudio y la fiabilidad de sus inferencias.

Siguiendo con el símil, esta publicación no es más que la guía rápida a tu primera cámara estadística. La cámara aquí descrita no es muy compleja, sino más bien una de esas cámaras de un solo uso que compramos cuando estamos de vacaciones y nos hemos olvidado la nuestra. Pero el fundamento de esta cámara de un solo uso es similar al de una cámara profesional del fotógrafo más elitista. Espero que esta publicación sirva como puente al campo de la "fotografía estadística" y estimule al lector a seguir analizando otros manuales de cámaras con las que seguir fotografiando la vida.

Santiago de Compostela, 15 de noviembre de 2007
Manuel Febrero Bande

Índice general

1. Estadística descriptiva **1**

1.1. Introducción 1

1.2. Descripción estadística unidimensional 1

1.2.1. Conceptos básicos 1

1.2.2. Frecuencias 2

1.2.3. Representaciones gráficas 3

1.2.4. Medidas de centralización 4

1.2.5. Medidas de dispersión 6

1.2.6. Medidas de forma 7

1.2.7. Otras medidas características 8

1.2.8. Transformaciones en los datos y su efecto en el análisis descriptivo 8

1.3. Descripción estadística de varias variables 9

1.3.1. Representaciones gráficas 10

1.3.2. Momentos 11

1.3.3. Covarianza y correlación 11

1.3.4. Dependencia lineal 12

1.4. Anexo 14

1.5. Ejercicio resuelto 16

2. Modelos de distribución de probabilidad **19**

2.1. Introducción 19

2.2. Espacio probabilístico 19

2.2.1. Experimentos y sucesos 19

2.2.2. Definiciones de probabilidad 21

2.2.3. Probabilidad condicionada 22

2.2.4. Independencia de sucesos 22

2.2.5. Regla del producto 22

2.2.6. Teorema de las probabilidades totales 22

2.2.7. Regla de Bayes 23

2.3. Variables aleatorias unidimensionales 24

2.3.1. Función de distribución de una variable aleatoria 25

2.3.2. Variables aleatorias discretas 26

2.3.3. Variables aleatorias continuas 26

2.3.4. Cambio de variable 27
2.4. Medidas características de una variable aleatoria 28
2.4.1. Media o esperanza matemática de una variable aleatoria 28
2.4.2. Varianza de una variable aleatoria 29
2.4.3. Coeficiente de variación 29
2.4.4. Momentos 29
2.4.5. Mediana 30
2.4.6. Cuantiles 30
2.4.7. Recorrido semi-intercuartílico 30
2.4.8. Moda 30
2.4.9. Coeficientes de asimetría 31
2.4.10. Coeficiciente de apuntamiento o curtosis 31
2.4.11. Desigualdad de Markov 31
2.4.12. Desigualdad de Tchebychev 31
2.4.13. Tipificación de una variable aleatoria 32
2.5. Principales distribuciones unidimensionales discretas 32
2.5.1. Distribución de Bernoulli 32
2.5.2. Distribución binomial 32
2.5.3. Distribución geométrica 33
2.5.4. Distribución binomial negativa 34
2.5.5. Distribución de Poisson 35
2.5.6. Distribución uniforme discreta 36
2.5.7. Distribución hipergeométrica 36
2.6. Principales distribuciones unidimensionales continuas 38
2.6.1. Distribución uniforme 38
2.6.2. Distribución normal 38
2.6.3. Distribución lognormal 40
2.6.4. Distribución exponencial 41
2.6.5. Distribución gamma 41
2.6.6. Distribución de Erlang 42
2.6.7. Distribución de Weibull 43
2.6.8. Distribución de tiempo de fatiga 43
2.6.9. Distribución beta 44
2.6.10. Distribuciones asociadas a la normal 45
2.7. Variables aleatorias multidimensionales 47
2.7.1. Función de distribución de una variable aleatoria bidimensional . . 47
2.7.2. Distribuciones marginales 48
2.7.3. Distribuciones condicionadas 49
2.7.4. Independencia de variables aleatorias 50
2.7.5. Medidas características de una variable aleatoria bidimensional . . 51
2.7.6. Transformaciones de variables bidimensionales 54
2.7.7. Caso n-dimensional 54
2.8. Modelos multidimensionales de distribución de probabilidad 55

2.8.1. Distribución multinomial . . . 55
2.8.2. Distribución normal multidimensional . . . 55
2.9. Sucesiones de variables aleatorias . . . 57
2.9.1. Leyes de los Grandes Números . . . 58
2.9.2. Teorema Central del Límite . . . 59
2.10. Anexo: repaso de combinatoria . . . 61
2.10.1. Combinaciones . . . 61
2.10.2. Combinaciones con repetición . . . 61
2.10.3. Variaciones . . . 62
2.10.4. Variaciones con repetición . . . 62
2.10.5. Permutaciones . . . 62
2.10.6. Permutaciones con repetición . . . 63
2.11. Ejercicios resueltos . . . 63

3. Inferencia paramétrica 73
3.1. Introducción a la Inferencia Estadística . . . 73
3.2. Conceptos . . . 73
3.3. Distribución muestral y función de verosimilitud . . . 75
3.4. Distribuciones en el muestreo de poblaciones normales . . . 77
3.4.1. Estimación de la media de una población . . . 77
3.4.2. Estimación de la varianza de una población . . . 78
3.4.3. Estimación de una proporción . . . 78
3.5. Intervalos de confianza . . . 79
3.5.1. IC para la media de una población normal . . . 79
3.5.2. IC para la varianza de una población normal . . . 80
3.5.3. IC para la diferencia de medias de poblaciones normales . . . 80
3.5.4. Muestras independientes, varianzas poblacionales conocidas . . . 81
3.5.5. Muestras independientes, varianzas desconocidas e iguales . . . 81
3.5.6. Muestras independientes, varianzas desconocidas y desiguales . . . 81
3.5.7. Muestras apareadas, varianzas poblacionales conocidas . . . 82
3.5.8. IC para la razón de varianzas de poblaciones normales . . . 82
3.6. Contrastes de hipótesis . . . 83
3.6.1. Hipótesis estadística . . . 83
3.6.2. Contraste para la media de una población normal . . . 85
3.6.3. Contraste para la varianza de una población normal . . . 87
3.6.4. Contraste para la diferencia de medias de poblaciones normales . . . 88
3.6.5. Contraste para la razón de varianzas de poblaciones normales . . . 91
3.6.6. Relación entre intervalos de confianza y contrastes de hipótesis. . . . 93
3.7. Ejercicio resuelto . . . 93

4. Inferencia no paramétrica 95
4.1. Introducción . . . 95
4.2. Hipótesis sobre la distribución . . . 95
4.2.1. El contraste χ^2 de Pearson . . . 96

4.2.2. El test de Kolmogorov-Smirnov 98
4.2.3. El contraste de Shapiro-Wilks 99
4.2.4. Contrastes de asimetría y curtosis 100
4.2.5. Transformaciones para conseguir normalidad 100
4.3. Contrastes de posición 101
4.3.1. Test de los signos y rangos para muestras apareadas 101
4.3.2. Test de Mann-Whitney-Wilcoxon para muestras independientes . . 101
4.3.3. Test de Kruskal-Wallis para múltiples muestras independientes . . 102
4.4. Hipótesis de independencia 102
4.4.1. Contraste de rachas 103
4.4.2. Contraste de autocorrelación 104
4.4.3. Test de Durbin-Watson 105
4.5. Hipótesis sobre la homogeneidad 105
4.5.1. Test de homogeneidad en tablas de contingencia 106
4.5.2. Test de valores atípicos 106
4.6. Ejercicio resuelto 107

5. Modelos de regresión **109**
5.1. Introducción 109
5.2. Planteamiento e hipótesis básicas 110
5.2.1. Hipótesis básicas iniciales 110
5.3. Estimación 111
5.3.1. Propiedades de los estimadores 113
5.4. Contrastes de regresión y de las hipótesis 116
5.5. Predicción 119
5.5.1. Predicción de la media condicionada a x 119
5.5.2. Predicción de una nueva observación condicionada a x 120
5.6. Ejercicio resuelto 121

Capítulo 1

Estadística descriptiva

1.1. Introducción

El objetivo de la Estadística descriptiva es estudiar procedimientos para sintetizar la información contenida en un conjunto de datos ofreciendo un resumen numérico o gráfico del estado de las cosas. Precisamente, de este concepto viene el nombre de "Estadística" que procede del latín "status" y parte de la necesidad de conocer el entorno en que nos movemos midiendo elementos individuales para obtener conclusiones generales aplicables a todo el conjunto.

1.2. Descripción estadística unidimensional

En este apartado estudiaremos procedimientos para resumir la información de una característica que se pueda observar en los elementos individuales.

1.2.1. Conceptos básicos

- **Población**: Conjunto de personas, objetos o acontecimientos sobre los que queremos obtener una conclusión.
- **Individuo**: Cada uno de los elementos de la población.
- **Muestra**: Subconjunto de la población (que representa adecuadamente a la misma).
- **Variables (o atributos)**: Son las características que se pueden observar o estudiar en los individuos de la población. Según el tipo de característica a medir se pueden clasificar en:
    - **Cualitativas nominales**: Miden características que no toman valores numéricos (color del pelo, raza, *etc.*). A estas características se les llama modalidades.

- **Cualitativas ordinales**: Miden características que no toman valores numéricos pero sí presentan entre sus posibles valores una relación de orden (nivel de estudios: sin estudios, primaria, secundaria, *etc.*).
- **Cuantitativas discretas**: Toman un número discreto de valores (en el conjunto de números naturales) (n^o de hijos de una familia, goles en un partido de fútbol, *etc.*).
- **Cuantitativas continuas**: Toman valores numéricos dentro de un intervalo real (altura, peso, concentración de un elemento, *etc.*).

1.2.2. Frecuencias

El primer método para resumir una muestra de tamaño n $\{x_1,\ldots,x_n\}$ de una variable estadística X, que presenta las modalidades c_1,..., c_m, es calcular la tabla de frecuencias. Como su nombre indica es una tabla donde se presentan las modalidades observadas y sus frecuencias de aparición:

- **Frecuencia Absoluta**: Número de veces que aparece la modalidad. Se denotará por n_i, $0 \leq n_i \leq n$.
- **Frecuencia Absoluta Acumulada**: Número de veces que aparece la modalidad o valores inferiores. Se denotará por N_i, $0 \leq N_i \leq n$, $N_{i-1} \leq N_i$, $N_m = n$.
- **Frecuencia Relativa**: Tanto por uno de las veces que aparece la modalidad. $f_i = n_i/n$, $0 \leq f_i \leq 1$.
- **Frecuencia Relativa Acumulada**: Tanto por uno de las veces que aparece la modalidad o valores inferiores. $F_i = N_i/n$, $0 \leq F_i \leq 1$, $F_{i-1} \leq F_i$, $F_m = 1$.

La siguiente tabla muestra la frecuencias para el conjunto de datos *Titanic* que contiene 4 variables cualitativas nominales de los 2201 pasajeros y tripulantes que se corresponden a: la clase del pasajero (1^a, 2^a, 3^a y tripulación), edad (niño/adulto), supervivencia (si/no) y el sexo (hombre/mujer). Véase el anexo para su implementación en R.

Clase	1st	2nd	3rd	Crew
Frec. Absoluta	325	285	706	885
Frec. Relativa	0,1477	0,1295	0,3208	0,4021
Frec. Absoluta acumulada	325	610	1316	2201
Frec. Relativa acumulada	0,1477	0,2771	0,5979	1,00

Si tenemos una variable continua también podemos crear una tabla de frecuencias agrupando estos datos numéricos en clases. Para ello podemos seguir las siguientes recomendaciones:

- Utilizar los datos limitando el número de cifras significativas.
- Decidir el número de clases a utilizar (k) que debe estar entre 5 y 20. Una regla muy utilizada es hacer $k = \sqrt{n}$.
- Seleccionar los límites de cada clase (LI_i, LS_i) sin ambigüedad y procurar que las clases sean de igual longitud (salvo información que aconseje de distinta longitud).
- Tomar como marca de clase el valor medio del intervalo creado.

Las frecuencias acumuladas tienen sentido con variables que presenten orden (cuantitativas o cualitativas ordinales).

El conjunto de datos *airquality* dispone de medidas de calidad del aire en Nueva York con las variables cuantitativas *Ozone* (ozono en ppb), *Solar.R* (radiación solar en langleys), *Wind* (viento en mph), *Temp* (temperatura en ^{o}F). En la tabla siguiente se muestra la tabla de frecuencias agrupada en 5 clases para la variable *Temp*.

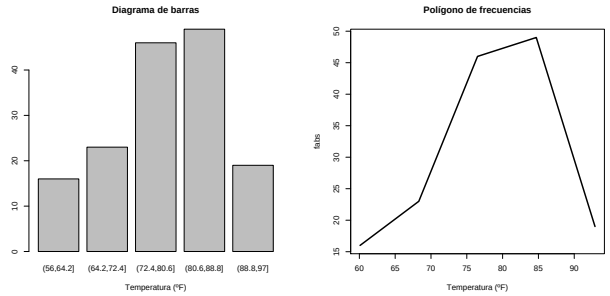
Clase Temp	(56,64.2]	(64.2,72.4]	(72.4,80.6]	(80.6,88.8]	(88.8, 97]
Frec. Abs.	16	23	46	49	19
Marca clase	60,1	68,3	76,5	84,7	92,9

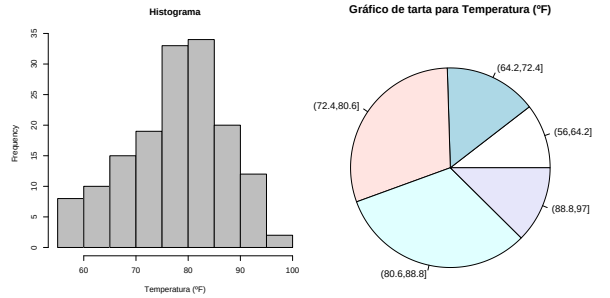
1.2.3. Representaciones gráficas

Si la variable toma pocos valores diferentes o es cualitativa, entonces para representar la distribución de frecuencias no acumuladas se utiliza:

- **Diagrama de barras**: Consiste en un gráfico cartesiano en el que se dibuja x_i en abscisas y n_i (o f_i) en ordenadas dibujando barras verticales en cada punto x_i de longitud n_i (o f_i).
- **Polígono de frecuencias**: Igual que el diagrama de barras pero lo que se unen son los puntos (x_i, n_i) consecutivos.
- **Diagrama acumulativo de frecuencias**: Se usa para representar frecuencias acumulativas. Es como el diagrama de barras pero representando la frecuencia acumulada N_i en vez de la frecuencia absoluta.
- **Histograma**: Es la representación gráfica utilizada para las variables continuas. Es básicamente un diagrama de barras donde la altura de la barra es $h_i = f_i/l_i$, siendo l_i es la longitud del intervalo o clase. La función en R para obtenerlos es *hist* y además de poder dibujar el histograma, calcula las marcas de clase y las frecuencias.
- **Diagrama de sectores (gráfico de tarta)**: Se representa la frecuencia de cada modalidad proporcionalmente al ángulo del sector que lo representa.

- **Pictograma:** Se representa cada modalidad asociando un dibujo cuyo volumen (anchura/altura) es proporcional a la frecuencia.
- **Diagrama de tallo y hojas**: Los datos se redondean a dos o tres cifras significativas, tomándose como tallo la primera o dos primeras cifras y como hojas las ultimas cifras. El tallo se separa de las hojas por una línea vertical. Así, cada tallo se representa una sola vez y el número de hojas representa la frecuencia. La impresión resultante es la de "acostar" un histograma.






```
56 | 0000
58 | 0000
60 | 000
62 | 000
64 | 0000
66 | 0000000
68 | 0000000
70 | 0000
72 | 00000000
74 | 00000000
76 | 0000000000000000
78 | 000000000000
80 | 0000000000000000
82 | 00000000000000
84 | 0000000000
86 | 000000000000
88 | 00000
90 | 00000
92 | 00000000
94 | 00
96 | 00
```

1.2.4. Medidas de centralización

Introducimos a continuación un primer conjunto de medidas cuyo objetivo es obtener un representante del conjunto de los datos.

Media aritmética

Se define la media aritmética (o simplemente media) como: $\bar{x} = \sum_{i=1}^{n} x_i/n$; $\bar{x} = \sum_{i=1}^{k} c_i f_i$ donde la primera expresión corresponde a tener todos los datos cuantitativos y la segunda corresponde a datos agrupados. La media aritmética tiene interesantes propiedades:

1. $\text{mín}(x_i) \leq \bar{x} \leq \text{máx}(x_i)$ y tiene las mismas unidades que los datos originales.

2. Es el centro de gravedad de los datos:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0; \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^{n}(x_i - a)^2.$$

3. Si $y_i = a + bx_i \Rightarrow \bar{y} = a + b\bar{x}$. (las transformaciones lineales se comportan bien con la media).

Media truncada o recortada

Un inconveniente de la media aritmética es que un dato anómalo puede hacerla variar mucho. La contribución de cada dato a la media es x_i/n. Si yo me equivoco al medir o anotar el dato x_i y le sumo 1000 unidades más, el efecto que se produce en la media es que se desplaza $1000/n$ unidades.

Para evitar este efecto se utiliza la media truncada que consiste en calcular la media aritmética de un porcentaje central de los datos (esto es, eliminando un porcentaje de los datos más bajos y de los más altos). Así una media truncada al 10 % calcularía la media aritmética del 90 % de los valores centrales despreciando el 5 % de los valores más bajos y el 5 % de los más altos.

La media recortada es un concepto parecido al anterior salvo que en vez de despreciar un porcentaje de los valores más bajos y más altos lo que se hace es modificar estos valores. Se sustituyen los valores más bajos por el más bajo de los valores centrales y los valores más altos por el más alto de los valores centrales.

Si en la muestra que hemos recogido no hay datos anómalos, la diferencia entre la media truncada (o recortada) y la media aritmética debe ser pequeña. Estas medidas no suelen utilizarse con valores agrupados.

Mediana

Se define la mediana (M_e) como aquel valor que, teniendo los datos ordenados de menor a mayor, deja igual número de valores a su izquierda que a su derecha. Si el número de datos es par se calcula como la media de los dos valores centrales. Si el número de datos es impar se toma como mediana el valor central. Si los datos se han agrupado se determina primero el intervalo mediano (aquel intervalo donde la frecuencia relativa acumulada es menor o igual que 0,5 en su extremo inferior y mayor que 0,5 en su extremo superior) para a continuación elegir un representante de este intervalo como mediana (la marca de clase, $LI_i + l_i(0{,}5\text{-}F_{i-1})/f_i$, *etc.*).

La mediana sería la medida de posición central más robusta (*i.e.* más insensible a datos anómalos) y coincidiría con la media truncada al 100 %. Además la mediana verifica que $\sum_{i=1}^{n}|x_i - M_e| = \min_{a\in\mathrm{R}} \sum_{i=1}^{n}|x_i - a|$.

Moda

La moda de una variable cuantitativa discreta o cualitativa es el valor más frecuente. En el caso de variables cuantitativas agrupadas se define el intervalo modal como aquel con mayor frecuencia relativa. La moda puede no ser única si tenemos varios intervalos con la misma frecuencia relativa máxima.

Otras medias

- Media cuadrática: $C = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$.
- Media geométrica: $G = \sqrt[n]{\prod_{i=1}^{n} x_i}$. Usada para medias de índices o razones.
- Media armónica: $H = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$. Usada para medias de porcentajes y promedios.

Otras medidas de posición

Cuantiles: Son una generalización del concepto de mediana. Teniendo ordenados los datos se define el cuantil de orden p ($0 \leq p \leq 1$) como el valor (q_p) que deja a lo sumo np observaciones a su izquierda y a lo sumo $n(1-p)$ observaciones a su derecha. La mediana es por tanto el cuantil de orden 0.5. Algunos órdenes de estos cuantiles tienen nombres específicos. Así los cuartiles son los cuantiles de orden (0.25, 0.5, 0.75) y se representan por Q_1, Q_2, Q_3. Los deciles son los cuantiles de orden (0.1, 0.2,..., 0.9). Los percentiles son los cuantiles de orden $j/100$ donde j=1,2,...,99. El procedimiento de cálculo de los cuantiles es similar al empleado para la mediana.

1.2.5. Medidas de dispersión

Tratan de medir la concentración o dispersión de las observaciones muestrales.

Varianza y desviación típica

Se define la varianza como $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, es decir, como la media aritmética de los cuadrados de las desviaciones respecto a la media. Se define la desviación típica como la raíz positiva de la varianza (s). Se suele utilizar más la desviación típica porque presenta las mismas unidades que la variable original. Al estar definidas como promedio de cuadrados son siempre no negativas. Respecto a las transformaciones lineales sucede que si $y_i = a + bx_i \Rightarrow s_y^2 = b^2 s_x^2$ y por tanto $s_y = |b|\, s_x$.

Otras medidas de dispersión

- Desviación absoluta respecto a la media: $D_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$.

- Desviación absoluta respecto a la mediana: $D_{Q_2} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Q_2|$.
- Mediana de las desviaciones absolutas: MEDA $= Q_2\{|x_i - Q_2(x)| : i = 1, \ldots, n\}$.
- Recorrido o rango: $R = \text{máx}(x_i) - \text{mín}(x_i)$.
- Rango intercuartílico: $RI = Q_3(x) - Q_1(x)$.
- Recorrido relativo: $RR = (\text{máx}(x_i) - \text{mín}(x_i))/\bar{x}$.
- Coeficiente de variación: $CV = s/\bar{x}$.

Las medidas relativas como el recorrido relativo o el coeficiente de variación sólo tienen sentido cuando la media de la variable es mayor que cero.

1.2.6. Medidas de forma

Las medidas de forma tratan de medir el grado de simetría y apuntamiento en los datos.

Medidas de asimetría

- Coeficiente de asimetría de Pearson: $As_P = (\bar{x} - Q_2)/s$.
- Coeficiente de asimetría de Fisher: $As_F = \sum_{i=1}^{n}(x_i - \bar{x})^3/ns^3$.

El coeficiente de asimetría de Pearson originalmente medía la diferencia entre media y moda. En distribuciones unimodales y aproximadamente simétricas la diferencia entre media y moda es aproximadamente tres veces la diferencia entre media y mediana. Por tanto, se utiliza este último porque el primero no puede calcularse propiamente en distribuciones multimodales. En cualquier caso, la interpretación de estos coeficientes es la siguiente: Si son prácticamente cero se dice que los datos son simétricos. Si toman valores significativamente mayores que cero diremos que los datos son asimétricos a la derecha y si toman valores significativamente menores que cero diremos que son asimétricos a la izquierda.

Medidas de apuntamiento o curtosis

Miden el grado de concentración de una variable respecto a su medida de centralización usual (media). El más usual es el coeficiente de apuntamiento de Fisher que se define como: $Sk_F = \sum_{i=1}^{n}(x_i - \bar{x})^4/ns^4$. Puesto que en Estadística el modelo de distribución habitual de referencia es el gausiano o normal y este presenta teóricamente un coeficiente de apuntamiento de 3, se suele tomar este valor como referencia. Así, si este coeficiente es menor que 3 diremos que los datos presentan una forma platicúrtica, si es mayor que 3 diremos que son leptocúrticos y si son aproximadamente 3 diremos que son mesocúrticos.

1.2.7. Otras medidas características

Varias de las medidas vistas anteriormente utilizan desviaciones de los datos respecto a la media elevadas a distintos órdenes. Este tipo de coeficientes se denominan momentos.

Se define el **momento respecto al origen de orden** r ($r \geq 0$) como: $a_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$.

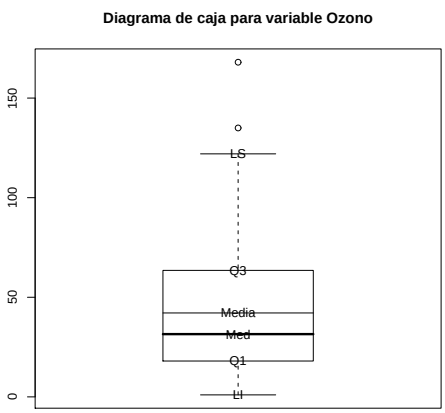
Se define el **momento central de orden** r ($r \geq 0$) como: $m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r$.

La relación entre los dos tipos de momentos viene dada a partir del binomio de Newton: $m_r = \sum_{k=0}^{r} (-1)^k \begin{pmatrix} r \\ k \end{pmatrix} a_{r-k} a_1^k$.

Casos particulares de los momentos son: $a_1 = \bar{x}$, $m_2 = s^2$, $m_3 = s^3 As_F$ y $m_4 = s^4 Sk_F$.

Diagramas de caja

La información obtenida a partir de las medidas de centralización, dispersión y forma se puede usar para realizar diagramas de caja (boxplots) que visualmente nos proporcionen la información de cómo están distribuidos los datos. El diagrama de caja consta de una caja central que está delimitada por la posición de los cuartiles Q_3 y Q_1. Dentro de esa caja se dibuja la línea que representa la mediana. También ocasionalmente se puede representar la media dentro de la caja. De los extremos de la caja salen unas líneas que se extienden hasta los puntos $LI = \text{máx}\{\text{mín}(x_i), Q_1 - 1{,}5\,(RI)\}$ y $LS = \text{mín}\{\text{máx}(x_i), Q_3 + 1{,}5\,(RI)\}$ que representarían el rango razonable hasta el cual se pueden encontrar datos. Los datos que caen fuera del intervalo (LI, LS) se consideran datos atípicos y se representan individualmente.



1.2.8. Transformaciones en los datos y su efecto en el análisis descriptivo

Cuando se desea realizar comparaciones entre valores particulares de variables medidas en distintas escalas conviene tener una referencia común para que la comparación resulte efectiva. Esto se puede conseguir mediante la tipificación. Se define la variable tipificada de una variable estadística X como la variable Z que resulta de restarle su media aritmética y dividir por su desviación típica, esto es, $Z = \frac{X - \bar{x}}{s}$. De esta manera, la nueva variable tendrá media cero y desviación típica unidad pudiéndose comparar sus valores individuales con los de cualquier otra variable tipificada.

Esta transformación no cambia las medidas adimensionales como son el coeficiente de asimetría de Fisher o la curtosis pero por supuesto sí cambia las medidas que presentan unidades. En general, las transformaciones lineales no alteran las medidas de forma adimensionales.

Otro tipo de transformaciones habituales en Estadística sería la familia de transformaciones Box-Cox.

$$X^{(\lambda)} = \begin{cases} \frac{(X+m)^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln(X+m) & \text{si } \lambda = 0, \end{cases} \qquad \text{siendo } X+m>0.$$

Este tipo de transformaciones permiten corregir la asimetría de los datos. Así, para valores de λ mayores que la unidad se corrigen asimetría a la izquierda y para valores λ menores que la unidad se corrigen asimetría a la derecha.

En general, si calculamos una nueva variable y como la transformación h de una variable x, podemos aproximar el efecto de la transformación en la media y varianza mediante las siguientes fórmulas: $\bar{y} \simeq h(\bar{x}) + \frac{1}{2}h''(\bar{x})s_x^2;\ s_y^2 \simeq s_x^2\,[h'(\bar{x})]^2$.

1.3. Descripción estadística de varias variables

Hasta ahora describíamos a cada individuo de la población mediante una única característica, sin embargo lo habitual es que tengamos varias características para un mismo individuo y que estas características puedan presentar relación entre ellas. Empezaremos con el estudio de variables estadísticas bidimensionales, es decir, tenemos dos características por cada individuo.

Variable estadística bidimensional

$\boldsymbol{X \setminus Y}$	$\boldsymbol{d}_1$	$\ldots$	$\boldsymbol{d}_j$	$\ldots$	$\boldsymbol{d}_l$	***Marg. X***
$\boldsymbol{c}_1$	$n_{11}(f_{11})$	$\ldots$	$n_{1j}(f_{1j})$	$\ldots$	$n_{1l}(f_{1l})$	$\sum_{j=1,\ldots,l} n_{1j} \left(\sum_{j=1,\ldots,l} f_{1j} \right)$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	
$\boldsymbol{c}_i$	$n_{i1}(f_{i1})$	$\ldots$	$n_{ij}(f_{ij})$	$\ldots$	$n_{il}(f_{il})$	$\sum_{j=1,\ldots,l} n_{ij} \left(\sum_{j=1,\ldots,l} f_{ij} \right)$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	
$\boldsymbol{c}_k$	$n_{k1}(f_{k1})$	$\ldots$	$n_{kj}(f_{kj})$	$\ldots$	$n_{kl}(f_{kl})$	$\sum_{j=1,\ldots,l} n_{kj} \left(\sum_{j=1,\ldots,l} f_{kj} \right)$
Marg. Y	$\sum_{i=1,\ldots,k} n_{i1} \left(\sum_{i=1,\ldots,k} f_{i1} \right)$		$\sum_{i=1,\ldots,k} n_{ij} \left(\sum_{i=1,\ldots,k} f_{ij} \right)$		$\sum_{i=1,\ldots,k} n_{il} \left(\sum_{i=1,\ldots,k} f_{il} \right)$	$n(1)$

Estudiaremos las características (X,Y) de una población de la cual obtenemos una muestra $(x_1,y_1),\ldots,(x_n,y_n)$. Igual que hemos hecho con una sola variable, cada una de

estas variables se puede agrupar en modalidades. Supongamos que las modalidades (o datos agrupados) de X son $c_1,\ldots,c_k$ y las de Y son $d_1,\ldots,d_l$. Sea además n_{ij} el número de individuos de la muestra que presentan la modalidad c_i de x y la d_j de y. Este número se conoce como la frecuencia absoluta del par (c_i,d_j). Al igual que para variables unidimensionales a $f_{ij} = n_{ij}/n$ se le conoce como frecuencia relativa. Las propiedades de estos números son idénticas al caso unidimensional. La distribución de frecuencias conjunta de la variable bidimensional (X,Y) es el resultado de organizar en una tabla de doble entrada las modalidades de las variables unidimensionales junto con las correspondientes frecuencias absolutas (relativas). Llamaremos **distribuciones marginales** a las distribuciones de frecuencias unidimensionales que resultan de agregar todas las frecuencias que incluyen una determinada modalidad de la variable unidimensional.

Normalmente se denotaran por

$$n_{i\cdot} = \sum_{j=1,\ldots,l} n_{ij} \left(f_{i\cdot} = \sum_{j=1,\ldots,l} f_{ij} \right)$$

cuando correspondan a frecuencias marginales de la primera variable y por

$$n_{\cdot j} = \sum_{i=1,\ldots,k} n_{ij} \left(f_{\cdot j} = \sum_{i=1,\ldots,k} f_{ij} \right)$$

cuando corresponda a la segunda.

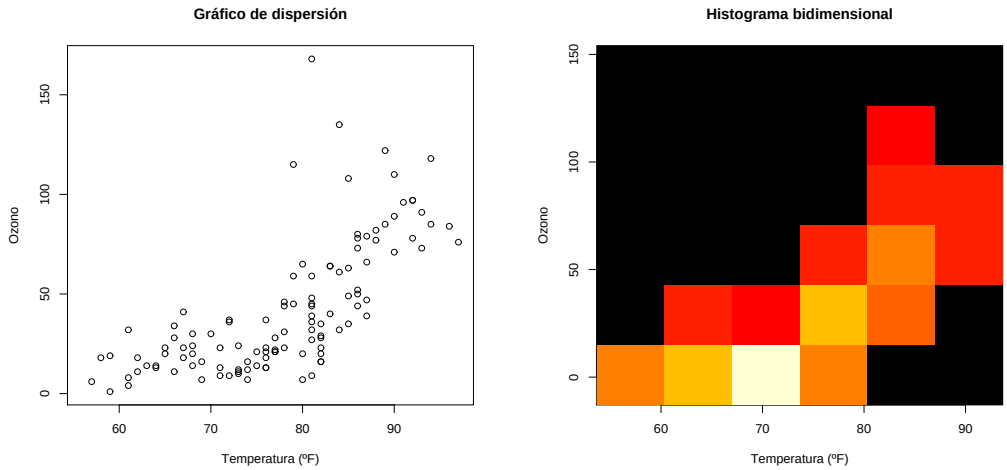
La **distribución de X condicionada a $Y=d_j$** es la distribución unidimensional de X sabiendo que Y ha tomado la modalidad d_j. Esto corresponde a dividir la columna de frecuencias absolutas (relativas) de la modalidad d_j por la suma de todos los valores de la columna. Análogamente se define la distribución de Y condicionada a $X = c_i$. La frecuencia relativa por tanto será $f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$.

1.3.1. Representaciones gráficas

La representación gráfica de las frecuencias se hace ahora en un diagrama de barras con dos dimensiones (una para cada variable) y calculando la altura de la barra de forma que la suma de los volúmenes sea la unidad (histograma bidimensional).

El **diagrama de dispersión** es una representación gráfica específica para variables bidimensionales cuantitativas que trata de medir la relación que existe entre ellas. Consiste en representar en un eje de coordenadas los pares de observaciones (x_i,y_i). La nube así dibujada (a este gráfico también se le llama nube de puntos) refleja la posible relación entre las variables. A mayor relación entre las variables más estrecha y alargada será la nube.

Cuando una de las variables sea categórica y la otra cuantitativa la representación gráfica apropiada incluye todos los gráficos vistos para variables unidimensionales pero clasificados por los valores de la variable categórica.



1.3.2. Momentos

Como ya vimos en el caso unidimensional muchas medidas se pueden escribir en función de los momentos de la variable.

Se define el **momento respecto al origen de orden (r,s)** $(r, s \geq 0)$ como:

$$a_{rs} = \frac{1}{n} \sum_{i=1}^{n} x_i^r y_i^s.$$

Se define el **momento central de orden (r,s)** $(r, s \geq 0)$ como:

$$m_{rs} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^r (y_i - \bar{y})^s.$$

Así, las medias marginales son, respectivamente, $a_{10} = \bar{x}$ y $a_{01} = \bar{y}$. Las varianzas marginales son, respectivamente, $m_{20} = s_x^2$ y $m_{02} = s_y^2$.

1.3.3. Covarianza y correlación

El caso particular de momento de orden (1,1) se conoce con el nombre de covarianza y puede interpretarse como una medida de relación lineal entre las variables X e Y.

$$\mathrm{Cov}(X,Y) = s_{xy} = m_{11} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = a_{11} - a_{10} a_{01}.$$

Esta fórmula es independiente del orden de las variables, es decir, $\mathrm{Cov}(X,Y)$=$\mathrm{Cov}(Y,X)$ y además en el caso de que $X = Y$ tendríamos la definición de varianza de X.

Se define la **correlación lineal** como

$$r(X,Y) = r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

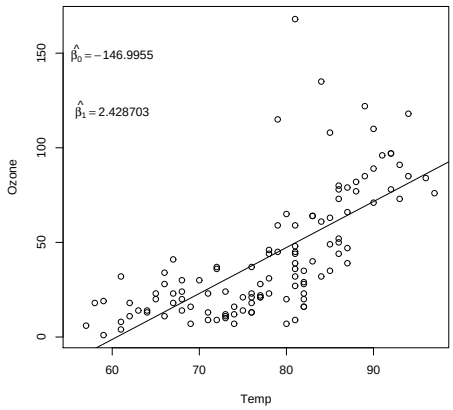
La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables. Así, si toma valores cercanos a -1 diremos que tenemos una relación inversa entre X e Y (esto es, cuando una variable toma valores altos la otra toma valores bajos). Si toma valores cercanos a $+1$ diremos que tenemos una relación directa (valores altos de una variable en un individuo, asegura valores altos de la otra variable). Si toma valores cercanos a cero diremos que no existe relación lineal entre las variables. Cuando el valor de la correlación lineal sea exactamente 1 o -1 diremos que existe una dependencia exacta entre las variables mientras que si toma el valor cero diremos que son incorreladas.

1.3.4. Dependencia lineal

En el estudio de variables bidimensionales tiene mucho interés buscar posibles relaciones entre las variables. La más sencilla de estas relaciones es la dependencia lineal donde se supone que la relación entre la variable dependiente (Y) y la variable regresora (X) se articula mediante una recta de regresión: $Y{=}\beta_0 + \beta_1 X + \varepsilon$ donde ε representa el error cometido que se comete al predecir Y mediante la fórmula lineal de X. El objetivo ahora es buscar los valores de los parámetros desconocidos (β_0, β_1) de la mejor manera posible. Aunque existen muchos métodos, el más clásico es el conocido como método de mínimos cuadrados que consiste en encontrar los valores de los parámetros que, dada la muestra de partida, minimizan la suma de los errores al cuadrado. Dada una muestra (x_1,y_1),...,(x_n,y_n) se trata de encontrar aquellos valores de (β_0, β_1) tal que $\sum_{i=1,\ldots,n} (y_i - \beta_0 - \beta_1 x_i)^2$ sea mínimo.

Los valores de los parámetros se obtienen, por tanto, derivando e igualando a cero obteniéndose la solución $\hat{\beta}_1 = s_{xy}/s_x^2$ y $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ que serán llamados coeficientes de la regresión. De esta manera obtendremos la ecuación de una recta: $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ que llamaremos **recta de regresión de *Y* sobre *X*** para resaltar que se ha obtenido suponiendo que Y es la variable respuesta y que X es la variable explicativa. Intercambiando los papeles de X e Y obtendremos una recta de regresión llamada **recta de regresión de *X* sobre *Y*** que representada en el mismo eje de coordenadas será en general distinta de la anterior. Solamente coincidirán en el caso de que la relación entre X e Y sea exacta.

Una vez resuelto el problema de estimar los parámetros surge la pregunta de si la recta estimada es o no representativa para los datos. Esto se resuelve mediante el **coeficiente de determinación** (R^2) que se define como el cuadrado del coeficiente de correlación lineal. El coeficiente de determinación toma valores entre 0 y 1 y representa el porcentaje de variabilidad de la variable dependiente que es explicada por la regresión. En el caso de la regresión entre *Temp* y *Ozone*, del conjunto de datos *airquality*, el coeficiente de cor-



relación lineal es 0,698 y el coeficiente de determinación es 0,488, que nos diría que el 48,8 % de la variabilidad del ozono es explicada por la temperatura según la recta de regresión.

Otra forma de calcular el coeficiente de determinación es mediante la fórmula dada por: $R^2 = 1 - \frac{s_R^2}{s_y^2}$ donde $s_R^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ que es conocida como varianza residual. Esta segunda manera de calcularla es válida para cualquier modelo de regresión que planteemos mientras que calcular el coeficiente de determinación como el cuadrado del coeficiente de correlación sólo es válido para rectas de regresión.

Generalización al caso k-dimensional

Estudiaremos las características de una población de la cual obtenemos una muestra $(x_{11},...,x_{k1}),\ldots,(x_{1n},...,x_{kn})$. Podemos proceder igual que en el apartado de variables bidimensionales definiendo la frecuencia absoluta como $n_{i_1,...,i_k}$ y la frecuencia relativa como $f_{i_1,...,i_k} = \frac{n_{i_1,...,i_k}}{N}$. Las propiedades de estas frecuencias son idénticas al caso bidimensional. La distribución de frecuencias conjunta de la variable $(X_1,...,X_k)$ es el resultado de organizar en una tabla de k dimensiones las modalidades de las variables unidimensionales junto con las correspondientes frecuencias absolutas (relativas). Llamaremos **distribuciones marginales** a las distribuciones de frecuencias unidimensionales que resultan de agregar todas las frecuencias que incluyen una determinada modalidad de alguna variable unidimensional. Ahora hablaremos de **vector de medias** como el vector k-dimensional que en cada componente presenta la media de cada variable unidimensional, es decir, $(\bar{x}_1, ..., \bar{x}_k)$. La covarianza entre dos variables X_i y X_j será: $\mathrm{Cov}(X_i, X_j) = s_{ij} = \frac{1}{n}\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)$. La varianza de X_i será $s_{ii} = s_i^2$ y el coeficiente de correlación lineal se definirá como $r(X_i, X_j) = r_{ij} = \frac{s_{ij}}{s_i s_j}$. Finalmente, llamaremos matriz de varianzas-covarianzas y matriz de correlaciones respectivamente a:

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{12} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_1^2 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{12} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}.$$

Como la matriz de varianzas-covarianzas no es un número y por tanto no se puede interpretar como dispersión, se conoce como **varianza generalizada** al determinante de la matriz de varianzas-covarianzas que ahora, al ser un número, sí se puede interpretar como cantidad de incertidumbre. Este determinante es mayor o igual que cero ya que la matriz de varianzas-covarianzas cumple la propiedad de ser semidefinida positiva (equivalente k-dimensional a decir en el caso unidimensional que un número es mayor o igual que cero). Entonces la varianza generalizada mide el "volumen ocupado" por los datos k-dimensionales generalizando el concepto de varianza para datos unidimensionales.

1.4. Anexo

```
# El conjunto de datos Titanic contiene 4 variables cualitativas
nominales de los 2201 pasajeros y tripulantes que corresponden a:
la clase del pasajero (1ª, 2ª, 3ª y tripulación), edad
(niño/adulto), supervivencia (si/no) y el sexo (hombre/mujer)#

>data(Titanic)
>fabs<-apply(Titanic,1,sum) # Frecuencia absoluta
>fabs/sum(fabs) # Frecuencia relativa
>facum<-double(length(fabs)) # Se crea un vector de tamaño el de fabs
>for (i in 1:length(fabs)) {facum[i]<-sum(fabs[1:i])} # Frec.Abs.Acum.
>facum/sum(fabs) # Frecuencia Relativa Acumulada

# Ejemplo de cálculo de frecuencias en variables continuas
>data(airquality)
>f<-cut(airquality$Temp,5) # Asigna datos numéricos uno de los 5 grupos
>fabs<-tapply(airquality$Temp,f,length) # Calcula la frec. abs.
(Repetir pasos anteriores para obtener otras frecuencias)
>marca<-numeric() # Creo un vector numérico sin dimensión
>for (i in 1:length(fabs)){
 marca[i]<-sum(type.convert(unlist(strsplit(chartr("(]","  ",
              names(fabs[i])),","))))/2
 } # Calculo las marcas de clase

# Ejemplo de diagrama de barras y de polígono de frecuencias
>par(mfrow=c(1,2))
>barplot(fabs,xlab="Temperatura (°F)",main="Diagrama de barras")
>plot(marca,fabs,type="l",lwd=3,xlab="Temperatura (°F)",
 main="Polígono de frecuencias")

# Ejemplo de histograma, diagrama de sectores y diagrama de tallo
y hojas
>hist(airquality$Temp)
>pie(fabs)
>stem(airquality$Temp)

# Ejemplo de diagrama de caja y de las medidas de centralización,
dispersión y forma
>data(airquality);attach(airquality)
>boxplot(Ozone)
>mOzone<-mean(Ozone,na.rm=T)
>text(rep(1,5),boxplot.stats(Ozone)$stats,c("LI","Q1","Med","Q3","LS"))
```

```
>text(1,mOzone,"Media")
>title("Diagrama de caja para variable Ozono")
>segments(0.8,mOzone,1.2,mOzone)
>quantile(Ozone,probs=c(0.05,0.25,0.50,0.75,0.95),na.rm=T)
    5%   25%   50%    75%   95%
  7.75 18.00 31.50 63.25 108.50
>var(Ozone,na.rm=T);sd(Ozone,na.rm=T)
[1] 1088.201 # Varianza
[1] 32.98788 # Desviación estandar
>mean(abs(Ozone[!is.na(Ozone)]-mOzone))
[1] 26.35018 # Desv. Abs.
>mean(abs(Ozone[!is.na(Ozone)]-median(Ozone,na.rm=T)))
[1] 24.88793 # Desv. Absoluta Mediana
>momento.centrado<-function(x,orden){
x.new<-x[!is.na(x)]
mean((x.new-mean(x.new))^orden)}
>momento.centrado(Ozone,3)/sd(Ozone,na.rm=T)^3
[1] 1.209866 # Asimetría de Fisher
>momento.centrado(Ozone,4)/sd(Ozone,na.rm=T)^4
[1] 4.112243 # Kurtosis

# Ejemplo de gráfico de dispersión y de histograma bidimensional
>data(airquality)
>attach(airquality)
>plot(Temp,Ozone,xlab="Temp.°F",ylab="Ozono",main="Gráf. de dispersión")
>library(gregmisc) # Librería que dispone de la función hist2d
>hist2d(Temp,Ozone,nbins=6,xlab="Temperatura °F",ylab="Ozono",
main="Histograma bidimensional")

# Ejemplo de ajuste de recta de regresión
>data(airquality)
>attach(airquality)
>regre<-lm(Ozone~Temp)
>plot(Temp,Ozone)
>abline(regre)
>coef(regre)
>text(60,150,expression(hat(beta[0])==-146.9955))
>text(60,120,expression(hat(beta[1])==2.428703))
>cor(Temp,Ozone,use="pairwise.complete.obs") # Coef. de correlación
>cor(Temp,Ozone,use="pairwise.complete.obs")^2 # Coef. de determinación
```

1.5. Ejercicio resuelto

Ejercicio: Una empresa de informática dedicada al análisis de virus en ordenadores, contabiliza los virus detectados con su producto en 20 ordenadores de domicilios particulares. Los resultados obtenidos son los siguientes:

$$46, 29, 35, 61, 54, 37, 53, 57, 52, 51, 43, 67, 66, 31, 53, 51, 48, 59, 55, 47.$$

a) Construir una tabla con las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas del conjunto de datos.

b) Dibujar un histograma del número de virus.

c) Obtener la media, mediana, moda, cuartiles, desviación típica, MEDA, coeficiente de variación, percentil del 40 %, el rango y el rango intercuartílico.

Solución:

a) Este apartado se resuelve con la siguiente tabla:

Pesos	Frec. absolutas	Frec. relativas	Frec. abs. acum.	Frec. rel. acum.
	n_i	f_i	N_i	F_i
$28 \leq x < 36$	3	0,15	3	0,15
$36 \leq x < 44$	2	0,10	5	0,25
$44 \leq x < 52$	5	0,25	10	0,50
$52 \leq x < 60$	7	0,35	17	0,85
$60 \leq x < 68$	3	0,15	20	1
	20	1		

b) La Figura 2.1 resuelve este apartado.

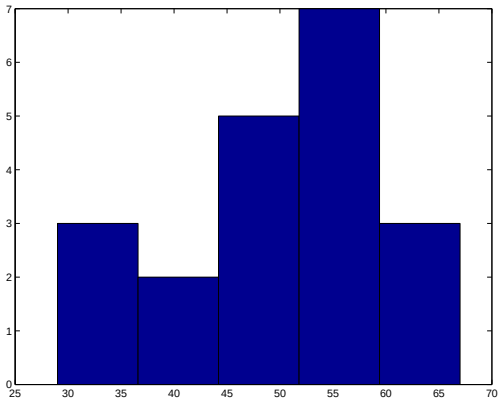


Figura 1.1: Histograma del número de virus

c) Se tiene:

- Media: $\overline{x} = 49{,}75$.
- Mediana: $M_e = 51{,}5$.
- Moda=51 y 53.
- Cuartiles: $Q_1 = 43,\ Q_3 = 55$.
- Desviación típica: $s = 10{,}32$.
- MEDA = 5,5.
- Coeficiente de variación: $CV = 0{,}20$.
- Percentil del 40 %: Per(40)=48.
- Rango: $R = \text{máx}(x_i) - \text{mín}(x_i) = 38$.
- Rango intercuartílico: $RI = Q_3(x) - Q_1(x) = 12$.

Capítulo 2

Modelos de distribución de probabilidad

2.1. Introducción

El concepto de probabilidad indica la posibilidad de ocurrencia de un suceso futuro, por ello está asociado a experimentos donde existe incertidumbre sobre el resultado final. Esta es la razón de que la Teoría de la Probabilidad sea importante por los muchos problemas prácticos que permite resolver. Además, supone un soporte teórico para la Estadística, más concretamente para la Inferencia Estadística, que es la que nos permite conocer (inferir) la distribución de una población a partir del conocimiento de una parte de ella (muestra).

La Teoría de la Probabilidad surgió de los estudios realizados sobre los juegos de azar, y estos se remontan miles de años atrás. Como primeros trabajos con cierto formalismo cabe destacar los realizados por Cardano y Galilei (siglo XVI), aunque las bases de esta teoría fueron desarrolladas por Pascal y Fermat en el siglo XVII. De ahí en adelante grandes científicos han contribuido al desarrollo de la Probabilidad, como Bernouilli, Bayes, Euler, Gauss,... en los siglos XVIII y XIX. Será a finales del siglo XIX y principios del XX cuando la Probabilidad adquiera una mayor formalización matemática, debida en gran medida a la llamada Escuela de San Petesburgo en la que cabe destacar los estudios de Tchebychev, Markov y Liapunov.

2.2. Espacio probabilístico

2.2.1. Experimentos y sucesos

Consideraremos que un **experimento** es "un proceso por medio del cual se obtiene una observación". Bajo este enfoque podemos distinguir entre experimentos **deterministas** y **aleatorios**. Los primeros son aquellos que siempre que se repitan bajo condiciones análogas llevan al mismo resultado, por tanto éste se puede predecir. Por el contrario, un experimento aleatorio es el que puede dar lugar a varios resultados, cono-

cidos previamente, sin que sea posible saber de antemano cuál de ellos se va a producir. Estos últimos son los que interesan a la Teoría de la Probabilidad. Como ejemplo de los mismos tenemos el observar qué número sale al lanzar un dado al aire. Muchos experimentos de la vida real entran en el campo de los experimentos aleatorios, ya que son muchas las situaciones en las que no se puede tener un control total sobre las variables de las que depende que se llegue a una u otra realización.

A continuación, describimos los principales conceptos necesarios para el estudio de un experimento aleatorio:

- **Suceso elemental:** Es cada uno de los posibles resultados del experimento aleatorio. Se denotan con la letra griega ω.
- **Espacio Muestral:** Conjunto formado por todos los sucesos elementales. Se denota por $\Omega = \{\omega \ / \ \omega \text{ es un suceso elemental}\}$.
- **Suceso:** Se llama suceso a cualquier subconjunto del espacio muestral. Se denota por $\emptyset$ al suceso imposible y Ω se corresponde con el suceso seguro.

  Ejemplo. Experimento aleatorio: Lanzamiento de un dado.

  Suceso elemental: el 3.

  Espacio Muestral: $\Omega = \{1,2,3,4,5,6\}$.

  Suceso: "Salir par" $= \{2,4,6\}$.

  Denotaremos por A^C al complementario del suceso A, es decir, $A^C = \Omega - A$.

- **Operaciones con sucesos:**
  - **Unión de sucesos:** Dados dos sucesos A y B, se define el suceso unión, $A \cup B$, como el que está formado por todos los sucesos elementales que están en A **o** en B.
  - **Intersección de sucesos:** Dados dos sucesos A y B, se define el suceso intersección, $A \cap B$, como el que está formado por todos los sucesos elementales que están en A **y** en B.
  - **Diferencia de sucesos:** Dados dos sucesos A y B, se define el suceso diferencia, $A \backslash B$, como el que está formado por todos los sucesos elementales que están en A y **no** en B, $A \backslash B = A \cap B^C$.

    Dos sucesos A y B se dicen incompatibles si $A \cap B = \emptyset$.

Para mayor comodidad en el momento en el que se asignen probabilidades a los sucesos, en vez de trabajar con todos los posibles sucesos asociados a un experimento aleatorio se trabaja con una familia de los mismos que se pretende sea suficiente:

- **Álgebra de sucesos.** Es un subconjunto del conjunto de todos los sucesos asociados a un experimento aleatorio, se denota por $\mathcal{A}$ y ha de cumplir las siguientes condiciones:

1. $\omega, \emptyset \in \mathcal{A}$.
2. $A \in \mathcal{A} \quad \Rightarrow \quad A^C \in \mathcal{A}$.
3. $A, B \in \mathcal{A} \quad \Rightarrow \quad A \cup B \in \mathcal{A}, A \cap B \in \mathcal{A}$.

Llamamos espacio probabilizable es un par (Ω,$\mathcal{A}$); un espacio muestral y un álgebra de sucesos definida a partir del mismo.

2.2.2. Definiciones de probabilidad

El principal objetivo de un experimento aleatorio suele ser determinar con qué probabilidad ocurre cada uno de los sucesos elementales. A continuación citamos las tres definiciones más manejadas para asignar probabilidades a los sucesos:

- **Definición frecuentista:** Dadas n repeticiones de un experimento aleatorio, si denotamos por n_A el número de veces que se ha obtenido el suceso A, se define la frecuencia de dicho suceso como $fr(A) = \frac{n_A}{n}$ donde $0 \leq fr(A) \leq 1$. Cuando n es grande la frecuencia de un suceso se estabiliza en torno a un valor al que se llama probabilidad del suceso A.

- **Definición clásica o de Laplace:** En el caso de que el espacio muestral sea finito y de que todos los sucesos elementales tengan la misma probabilidad, se define la probabilidad de un suceso A como:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{casos favorables}}{\text{casos posibles}},$$

  donde $|A|$ denota el número de sucesos elementales que componen el suceso A.

- **Definición axiomática (Kolmogorov 1933):** Dado el espacio probabilizable (Ω,$\mathcal{A}$), diremos que P es una probabilidad sobre dicho espacio si cumple:

  1. $P(\Omega) = 1$.
  2. Si $A \cap B = \emptyset$, entonces $P(A \cup B) = P(A) + P(B)$.
  3. $0 \leq P(A) \leq 1$.

El espacio probabilizable (Ω,$\mathcal{A}$), junto con la medida de probabilidad P, se denomina espacio de probabilidad y se representa como (Ω,$\mathcal{A}$,P).

Ejercicio: Prueba que en un espacio de probabilidad (Ω,A,P) se satisfacen las siguientes propiedades:

1. $P(\emptyset) = 0$.
2. $P(A) = 1 - P(A^C)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. Si $A \subseteq B$, entonces $P(A) \leq P(B)$.

2.2.3. Probabilidad condicionada

Es posible que, al realizar un experimento aleatorio, se disponga de cierta información que permite reducir el espacio muestral. Para esto se introduce la probabilidad condicionada; $P(A/B)$ denota la probabilidad de que se produzca el suceso A sabiendo que se va a producir el B. Por ejemplo, si sabemos que al lanzar un dado ha salido un número par y queremos saber la probabilidad de que este sea el 4, habría que calcular $P(\{4\}/\{2,4,6\})$.

De este modo, dado un suceso B tal que $P(B) > 0$ se define la probabilidad del suceso A **condicionada** al suceso B como:

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

Es importante destacar que dado un suceso B, la función P_B, que a cada suceso A le asigna la probabilidad de A condicionada a B, es una función de probabilidad que satisface las propiedades de la definición axiomática de Kolmogorov. Es decir, $P_B(A) = P(A/B)$.

2.2.4. Independencia de sucesos

Dos sucesos A y B son independientes si el hecho de que se produzca o no uno de ellos no afecta a la posible ocurrencia del otro. Formalmente, A y B son independientes si $P(A \cap B) = P(A) \cdot P(B)$ o equivalentemente $P(B/A) = P(B)$ si $P(A) > 0$ (y también $P(A/B) = P(A)$ si $P(B) > 0$).

Ejercicio: Comprobar que en el lanzamiento de un dado los sucesos $A = \{4\} = \{$Salir un 4$\}$ y $B = \{1,2,3,4\} = \{$salir menor que 5$\}$ no son independientes. Sin embargo los sucesos C $= \{2,4,6\} = \{$salir par$\}$ y B sí lo son.

2.2.5. Regla del producto

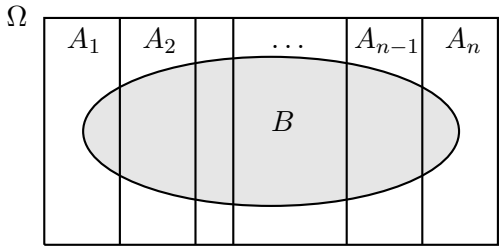
Dados los sucesos $A_1, A_2, \ldots, A_n$, tales que $P(\bigcap_{i=1}^{n-1} A_i) > 0$. Entonces:

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/(A_1 \cap A_2) \cdot \ldots \cdot P(A_n/\bigcap_{i=1}^{n-1} A_i).$$

2.2.6. Teorema de las probabilidades totales

Dados los sucesos $A_1, A_2, \ldots, A_n$, tales que $\Omega = \bigcup_{i=1}^{n} A_i$ y además $A_i \cap A_j = \emptyset$ *si* $i \neq j$. Entonces, dado un suceso B se tiene que

$$P(B) = \sum_{i=1}^{n} P(B/A_i) \cdot P(A_i)$$



Lo que nos dice este teorema es que dado un conjunto de sucesos mutuamente excluyentes tales que su unión sea el suceso seguro Ω, entonces la probabilidad de un suceso cualquiera B se puede descomponer como la suma de las probabilidades de B dentro de cada uno de los sucesos (rectángulos del dibujo) por la probabilidad de caer en dicho suceso. Con otras palabras, la probabilidad del suceso B se reparte entre los sucesos en los que hemos particionado Ω.

2.2.7. Regla de Bayes

Dados los sucesos $A_1, A_2, \ldots, A_n$, tales que $\Omega = \bigcup_{i=1}^{n} A_i$ y además $A_i \cap A_j = \emptyset$ *si* $i \neq j$. Entonces, dado un suceso B se tiene que

$$P(A_j/B) = \frac{P(B/A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B/A_i) \cdot P(A_i)}.$$

Esta fórmula sale de combinar las fórmulas de la probabilidad condicionada con el teorema de las probabilidades totales. La utilidad de la misma radica en que conociendo como son las probabilidades de B condicionadas a los sucesos en los que hemos descompuesto el espacio muestral, podemos calcular también cuánto valen las probabilidades cuando quien condiciona es el propio suceso B.

Ejercicio: En un hospital se realiza una prueba para detectar una enfermedad. Se sabe que la padecen 1 de cada 10.000 personas. Asimismo, también se sabe que cuando un paciente tiene la enfermedad la prueba da positivo el 90 % de las veces y que cuando está sano el test da positivo un 10 % de las veces.

a) ¿Cuál es la probabilidad de que el test dé positivo?

b) ¿Hasta qué punto es fiable el test? Es decir, si una persona da positivo, ¿qué probabilidad hay de que tenga la enfermedad?

Solución:

a) Aquí se usa el teorema de las probabilidades totales. Denotemos por A el suceso "tener la enfermedad" y por B "el test da positivo", de modo que sabemos que $P(A) = 0{,}0001$, $P(B/A) = 0{,}9$, $P(B/A^C) = 0{,}1$. Entonces:

$$P(B) = P(B/A)P(A) + P(B/A^C)P(A^C) = 0{,}9 \times 0{,}0001 + 0{,}1 \times 0{,}9999 = 0{,}10008.$$

b) Ahora utilizamos el teorema de Bayes, se nos pide $P(A/B)$.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B/A)P(A) + P(B/A^C)P(A^C)} = \frac{0{,}00009}{0{,}10008} = 0{,}000899.$$

De modo que aunque alguien dé positivo en el test, la probabilidad de que tenga la enfermedad es todavía muy pequeña. Haría falta una prueba más fiable.

2.3. Variables aleatorias unidimensionales

Es posible que en un experimento aleatorio tengamos interés en cuantificar los sucesos del espacio muestral. Por ejemplo, si tenemos un experimento que consiste en tirar 2 monedas al aire, es posible que lo que nos interese sea simplemente contar el número de caras y no nos importe en qué monedas han salido las mismas. Para esto se define una variable aleatoria que asigna a cada suceso elemental un número real, después de esto, utilizando la función de probabilidad del espacio muestral de partida se puede definir una nueva probabilidad sobre la recta real. Además, el trabajar con números reales nos permite hacer uso de herramientas matemáticas a la hora de estudiar las propiedades de un determinado experimento aleatorio.

Dado un espacio de probabilidad (Ω,A,P), la **variable aleatoria** $\boldsymbol{X}$ se define como una función que asocia un número real a cada suceso elemental de Ω, verificando la propiedad de que el conjunto $\{\omega \in \Omega$ tal que $X(\omega) \leq r\} = X^{-1}(\text{-}\infty,r])$ pertenece a $\mathcal{A}$. Este requerimiento nos permite definir una probabilidad sobre la recta real de la siguiente manera: $P_X(B) = P(X^{-1}(B))$.

Ejemplo: Supongamos que nuestro experimento aleatorio consiste en tirar dos monedas al aire, el espacio muestral es $\{(c,c),(c,+),(+,c),(+,+)\}$, siendo estos 4 sucesos elementales equiprobables. Considérese ahora la variable aleatoria $X =$"Número de caras".

Entonces, $X(c,c) = 2$; $X(c,+) = 1$; $X(+,c) = 1$; $X(+,+) = 0$ y además:

$$\begin{aligned} P_X(0) &= P(+,+) = 0{,}25 \\ P_X(1) &= P((c,+) \cup (+,c)) = 0{,}5 \\ P_X(2) &= P(c,c) = 0{,}25 \end{aligned}$$

$P_X(0) = P(+,+) = 0{,}25$, $P_X(1) = P((c,+) \cup (+,c)) = 0{,}5$, $P_X(2) = P(c,c) = 0{,}25$. Del mismo modo podríamos tener:

$$\begin{aligned} P_X([2,3)) &= P(c,c) = 0{,}25 \\ P_X((-\infty,1]) &= P((c,+) \cup (+,c) \cup (+,+)) = 0{,}75. \end{aligned}$$

A continuación definimos algunos conceptos que siempre acompañan al de variable aleatoria y nos permiten conocer mejor sus propiedades (y por tanto las del experimento aleatorio del que proceden).

2.3.1. Función de distribución de una variable aleatoria

La función de distribución F de una variable aleatoria X, es una función definida en la recta real que toma valores en el intervalo [0,1].

$F(x\ ) = P(X \leq x) = P(\{\omega \in \Omega \text{ tales que } X(\omega) \leq x\}) = P(X^{-1}((\text{-}\infty,x]))$

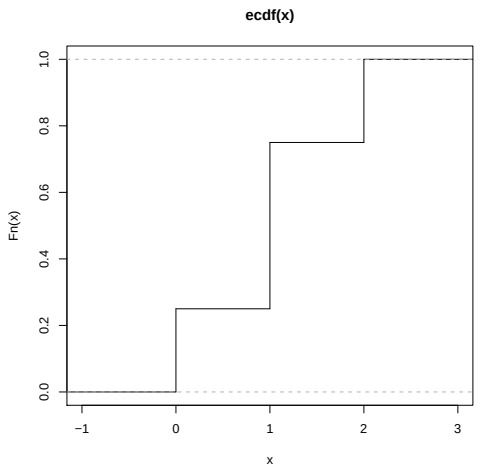
Para cada valor "x" que puede tomar la variable, la función F nos devuelve la probabilidad de que la variable tome un valor menor o igual que x.

Propiedades de una función de distribución:

1. $0 \leq F(x) \leq 1$.
2. F es no decreciente.
3. $\lim_{x \to +\infty} F(x) = 1$.
4. $\lim_{x \to -\infty} F(x) = 0$.
5. F es continua por la derecha.

Ejemplo: Considérese la variable aleatoria del anterior ejemplo (contar el número de caras al lanzar dos monedas al aire).

```
library(stepfun)  # paquete para funciones de distribución
x<-c(0,1,1,2) # valores que toma la variable
plot(ecdf(x),do.points=FALSE,verticals=TRUE) # representamos F
# La función ecdf calcula F y plot la dibuja
```



Vemos en la gráfica que el salto en 1 es el doble de grande que los saltos que se producen en 0 y 3. Además se pueden comprobar fácilmente las 5 propiedades.

Antes de seguir con los conceptos asociados a las variables aleatorias debemos distinguir entre variables aleatorias discretas y continuas.

2.3.2. Variables aleatorias discretas

Son aquellas que sólo toman valores dentro de un conjunto finito o infinito numerable.

Función de masa de probabilidad de una variable discreta: Es la que nos indica la probabilidad de cada uno de los valores de la variable (no es acumulada como la función de distribución). Se denota por p, por tanto $p(x) = P(X = x)$.

En el caso de las variables discretas se cumple que $F(x) = \sum_{y \leq x} p(y)$, la función de distribución se obtiene "acumulando los valores" que va tomando la función de masa de probabilidad.

2.3.3. Variables aleatorias continuas

Una variable aleatoria es continua si toma todos los valores en uno o varios intervalos de la recta real (por tanto toma una cantidad de valores infinita no numerable). Imaginemos que tenemos un experimento aleatorio que nos permite sacar un número al azar entre 0 y 1 de tal manera que todos son equiprobables. En este caso todos ellos tienen probabilidad 0 y sin embargo tenemos que la probabilidad total es 1 o que la probabilidad de obtener un número menor o igual que 0,5 es $F(0{,}5) = 0{,}5$. La función de densidad nos mide como crece la función de distribución en cada punto (que no es lo mismo que la probabilidad en ese punto).

Función de densidad de una variable continua: Se denota por f, y se calcula:

$$f(x) = F'(x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{P(x - h \leq X \leq x + h)}{2h}.$$

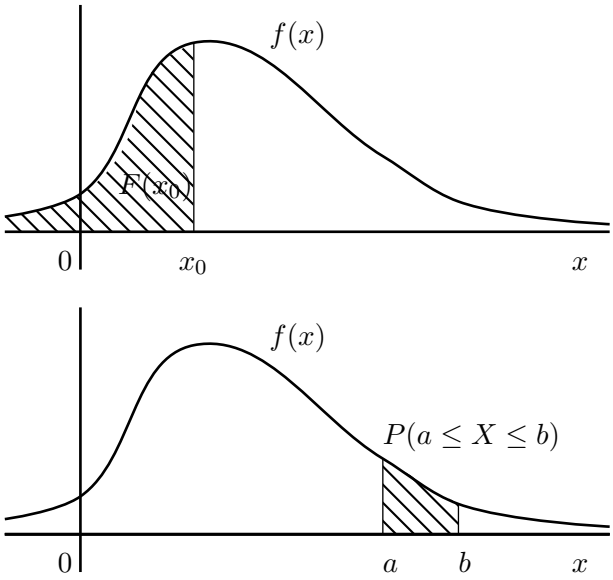
La función de distribución se obtiene "acumulando los valores" que va tomando la función de densidad $F(x_0) = \int_{-\infty}^{x_0} f(x)dx$. La función de densidad no indica probabilidad, es el área bajo la curva quien lo hace, de ahí que haya que integrar.

Propiedades:

1. $f(x) \geq 0, \ -\infty < x < +\infty$.
2. $\int_{-\infty}^{+\infty} f(x)dx = F(+\infty) = 1$.
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$.
4. Todo punto tiene probabilidad 0, $P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$.

Función de masa de probabilidad y función de densidad son conceptos análogos, el uso de uno u otro depende de la naturaleza de la variable en estudio.

Aunque aquí no las vamos a estudiar, también es posible encontrarse con variables aleatorias en las cuales determinados valores puntuales se toman con una probabilidad positiva y el resto de valores se toman dentro de uno o varios intervalos de acuerdo a una función de densidad. En estos casos se hablará de **variables aleatorias mixtas**.



2.3.4. Cambio de variable

Supongamos que tenemos una variable aleatoria X que nos mide la temperatura en un determinada región. Es posible que nosotros estemos interesados en estudiar cómo de lejos están sus valores de los veinte grados, para esto habría que estudiar cosas del tipo $P(20-\alpha \leq X \leq 20+\alpha)$. Sin embargo, si consideramos la variable aleatoria $Y = |X - 20|$, tendríamos probabilidades de la forma $P(Y \leq \alpha)$, porque ahora el punto de interés ha pasado a ser el 0 y además todos los valores son positivos (sólo queríamos estudiar cómo de lejos estaban los valores del 20, no hacia que lado). Los cambios de variable son todavía más útiles a la hora de trabajar con las medidas características de una variable aleatoria (siguiente tema).

A partir de una variable aleatoria X, definimos la variable $Y = g(X)$, donde g ha de ser una función continua y monótona (esto es para poder trabajar cómodamente con inversas, aunque también se pueden estudiar transformaciones más generales). Ahora veremos cómo calcular la función de distribución asociada a la variable Y conociendo la de X.

En general, si denotamos por G a la función de distribución de la variable Y tenemos:

$$G(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y])).$$

Ahora veremos cómo adaptar esta fórmula según la variable sea discreta o continua:

Caso discreto: Sea X una variable aleatoria discreta que toma valores x_i, con función de masa de probabilidad p, es decir $P(X = x) = p(x)$. Entonces para la variable aleatoria Y, que toma los valores y_j tenemos:

$$P(Y = y_j) = P(g(X) = y_j) = P(g^{-1}(y_j)) = \sum_i P(x_i \text{ tales que } g(x_i = y_j)).$$

Caso continuo: Sea X una variable continua con función de densidad $f_X(x)$, sea g una función continua y monótona. Entonces $Y = g(x)$ y, equivalentemente, $X = g^{-1}(Y\ )$. Denotamos por $f_Y(y)$ a la función de densidad de la variable transformada, entonces:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| = f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right|.$$

El valor $J = \left| \frac{dx}{dy} \right|$ se conoce como el jacobiano de la transformación.

Ejemplo: Si tuviésemos $Y = g(X) = aX + b$, entonces $X = g^{-1}(Y) = \frac{Y-b}{a}$. Por tanto $J = \left| \frac{dx}{dy} \right| = \left| \frac{dg^{-1}(y)}{dy} \right| = \left| \frac{1}{a} \right|$, de modo que $f_Y(y) = f_X\left(\frac{y-b}{a} \right) \left| \frac{1}{a} \right|$.

2.4. Medidas características de una variable aleatoria

La interpretación de conceptos como media o varianza o momentos es la misma que se hacía en el primer tema.

2.4.1. Media o esperanza matemática de una variable aleatoria

Caso discreto: Sea X una variable aleatoria discreta que toma valores x_1, x_2,..., x_i,..., x_n,..., con probabilidades $p_1, p_2, \ldots, p_i, \ldots, p_n, \ldots$ La **media o esperanza matemática** de la variable X es el número real:

$$\mu = E(X) = \sum_i x_i p_i \text{ (supuesto que } \sum_i |x_i|\, p_i < \infty).$$

Caso continuo: Sea X una variable aleatoria continua con función de densidad f, la **media o esperanza matemática** de la variable X es el número real:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx \text{ (supuesto que } \int_{-\infty}^{+\infty} |x|\, f(x) dx < \infty).$$

Nótese que las dos definiciones son en realidad la misma, sólo que utilizando en cada caso la herramienta matemática apropiada. En el primer caso se realiza el promedio a través de sumatorios. En el segundo, debido a la imposibilidad de hacer un promedio sobre un continuo con un sumatorio, se ha de echar mano de las integrales.

Propiedades:

1. $E(aX + b) = aE(X) + b$.
2. $E(X + Y) = E(X) + E(Y)$.

3. $E(X \cdot Y) = E(X) \cdot E(Y) \Leftrightarrow X$ e Y son independientes.

4. Si $Y = g(X)$ entonces:

   - $E(Y) = \sum_i g(x_i)p_i$ (caso discreto).
   - $E(Y) = \int_{-\infty}^{+\infty} g(x)f(x)dx$ (caso continuo).

2.4.2. Varianza de una variable aleatoria

Sea X una variable aleatoria con media $\mu = E(X)$, la **varianza** de X es el valor esperado de los cuadrados de las diferencias con respecto de la media:

$$\sigma^2 = \mathrm{Var}(x) = E((X - E(X))^2).$$

Caso discreto: La fórmula de la varianza en el caso discreto se puede escribir como:

$$\sigma^2 = \sum_{x \in X} (x - E(x))^2 p(x).$$

Caso continuo: La fórmula ahora es $\sigma^2 = \int_{-\infty}^{+\infty} (x - E(x))^2 f(x)dx$.

La **desviación típica** es la raíz positiva de la varianza: $\sigma = +\sqrt{\mathrm{Var}(X)}$. La principal ventaja de la desviación típica sobre la varianza es que los resultados vienen en las mismas unidades que los valores de la variable.

Propiedades:

1. $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

2. $\mathrm{Var}(X) = E((X - E(X))^2) = E(X^2 - 2E(X)X + E(X)^2) = E(X^2) - E(X)^2$ (usando que E(X) es una constante). Esta es la fórmula reducida para el cálculo de la varianza.

2.4.3. Coeficiente de variación

El coeficiente de variación se define $CV(X) = \frac{\sigma}{\mu}$ siempre que μ sea distinto de 0.

Se usa para comparar entre sí los grados de dispersión de distintas variables, este coeficiente no varía ante cambios de escala.

2.4.4. Momentos

Los momentos se dividen en dos tipos; momentos respecto del origen y momentos respecto de la media. A continuación damos una pequeña intuición de la utilidad de los momentos. Los primeros, entre los que se incluye la media, tienen como objetivo calcular la esperanza de las variables X, X^2,...,X^n. Los momentos respecto de la media, cuando

tienen orden par miden dispersión y cuando tienen orden impar se usan para medir asimetrías.

El **momento de orden r** de una variable X, denotado por a_r, es la esperanza de la variable X^r

$$a_r = E(X^r).$$

El momento central de orden r o momento de orden r con respecto de la media, denotado por m_r se calcula como

$$m_r = E\left((X - E(X))^r\right).$$

A la vista de esta fórmula se ve fácilmente porqué los momentos centrales de orden par miden dispersión y los de orden impar asimetrías: si el orden es par todas las diferencias con respecto de la media se elevan a una potencia par, haciéndose positivas, de modo que las distancias a izquierda y derecha de la media se suman. Por el contrario, cuando el exponente es impar las diferencias a izquierda y derecha se van cancelando, llevando a asimetrías positivas o negativas según el signo.

2.4.5. Mediana

La **mediana** de una variable aleatoria es una medida de centralización, divide la distribución en dos partes de igual probabilidad. Se denota por M_e y ha de cumplir que $F(M_e) = 0{,}5$.

Nótese que esta definición no implica que la mediana sea única (cosa que sí pasaba con la media).

2.4.6. Cuantiles

Suponen una generalización de la mediana.

Los **cuantiles de orden p**, con $0 < p < 1$, denotados por Q_p son aquellos valores x_p tal que la probabilidad de los valores a su izquierda coincide con p, esto es $F(x_p) = p$ (para el caso discreto se toma ínf $\{x : F(x) \geq p\}$).

Los cuantiles más usados son aquellos con probabilidades 0,25, 0,5 (mediana) y 0,75 denominados primer, segundo y tercer cuartil respectivamente (Q_1, Q_2, Q_3).

2.4.7. Recorrido semi-intercuartílico

Es una medida de dispersión, se denota por **SIQR** y viene dado por

$$\text{SIQR} = (Q_3 - Q_1)/2.$$

2.4.8. Moda

La **moda** de una variable aleatoria, denotada por M_o, es el valor que maximiza la función de probabilidad o la función de densidad, según se trate de una variable discreta o continua.

2.4.9. Coeficientes de asimetría

Nos sirven para saber si la función de probabilidad tiene más peso a un lado u otro de la media:

- Coeficiente de asimetría de **Pearson**: $As_P = (\mu - M_e)/\sigma$.
- Coeficiente de asimetría de **Fisher**: $As_F = m_3/\sigma^3$.

2.4.10. Coeficiciente de apuntamiento o curtosis

Mide el grado de concentración de los datos alrededor de la media, se denota por Sk_F y se calcula como $Sk_F = m_4/\sigma^4$. Un valor superior a 3 indica mayor concentración que en la distribución normal, y una variable con este valor se denomina **leptocúrtica**. Análogamente si el valor es 3 se habla de **mesocúrtica** y el término **platicúrtica** se usa cuando $Sk_F < 3$.

2.4.11. Desigualdad de Markov

Si X es una variable aleatoria que sólo toma valores no negativos, entonces la desigualdad de Markov nos dice que

$$P(X \geq k) \leq \frac{E(x)}{k},\ k > 0.$$

Si pensamos en transformaciones de la variable X tenemos también que

$$P(g(X) \geq k) \leq \frac{E(g(x))}{k},\ k > 0,$$

donde lo único que debemos exigirle a g es que sea no negativa.

Conocida la media de la variable aleatoria, esta desigualdad nos permite conocer una cota para la probabilidad de que la variable tome valores por encima de un valor arbitrario k.

2.4.12. Desigualdad de Tchebychev

Dada una variable aleatoria X con media μ y desviación típica σ, la desigualdad de Tchebychev nos dice que para cualquier constante positiva k tenemos

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Esta desigualdad nos permite dar una cota para la dispersión de la variable en función de la desviación típica. Por ejemplo, para $k = 3$ nos dice que la probabilidad de que una variable aleatoria tome valores en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ es siempre superior a 0,88. La desigualdad de Tchebychev es un caso particular de la desigualdad de Markov cuando tomamos $g(x) = (X - E(X))^2$ y $k = k^2\sigma^2$.

2.4.13. Tipificación de una variable aleatoria

Una variable aleatoria está **estandarizada** o **tipificada** si su media es 0 y su varianza 1. Una variable aleatoria con media μ y desviación típica σ se puede estandarizar mediante la transformación $Y = \frac{X-\mu}{\sigma}$. Tipificar variables es de gran utilidad a la hora de trabajar con variables cuyas probabilidades están tabuladas.

2.5. Principales distribuciones unidimensionales discretas

2.5.1. Distribución de Bernoulli

Los experimentos de Bernoulli son aquellos que sólo presentan dos posibles resultados: éxito/fracaso. La variable X toma entonces los valores $\{0,1\}$. La probabilidad p de 1 (éxito) se conoce de antemano. Esta probabilidad es siempre la misma, no varía a medida que se repite el experimento.

La distribución de Bernoulli es la que estudia un experimento de Bernoulli que se realiza una sola vez.

$$X = \begin{cases} 0 & \text{si fracaso,} \\ 1 & \text{si éxito.} \end{cases}$$

La **función de probabilidad** de una distribución de Bernoulli de parámetro p es $P(X = 1) = p$ y $P(X = 0) = 1 - p$.

Características: $E(X) = p,\ \mathrm{Var}(X) = p(1-p)$.

2.5.2. Distribución binomial

Se denota como $\mathrm{B}(n, p)$ a la repetición n veces de un proceso de Bernoulli de parámetro p (por tanto una distribución $\mathrm{B}(1,p)$ es una distribución de Bernoulli).

Ejemplos de variables que se pueden estudiar con este modelo podrían ser "número de caras al tirar 10 veces una moneda" o "número de piezas defectuosas en un proceso de fabricación".

Función de probabilidad: $P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n.$

Características: $E(X) = np$, $\mathrm{Var}(X) = np(1-p)$.

Esta distribución se utiliza en procesos de control de calidad y en el muestreo con reemplazamiento.

Dadas las variables $X \in \mathrm{B}(n, p)$ e $Y \in \mathrm{B}(m, p)$, entonces la variable aleatoria $X + Y$ se distribuye según una $\mathrm{B}(n + m, p)$.

Ejercicio: En un proceso de fabricación de microchips, la probabilidad de que una pieza salga defectuosa es $p = 0{,}001$, si cada día se producen 10.000 piezas ¿Cuál es la probabilidad de que un día no haya ningún chip defectuoso?¿y de que haya como mucho 10 defectuosos?

Solución:

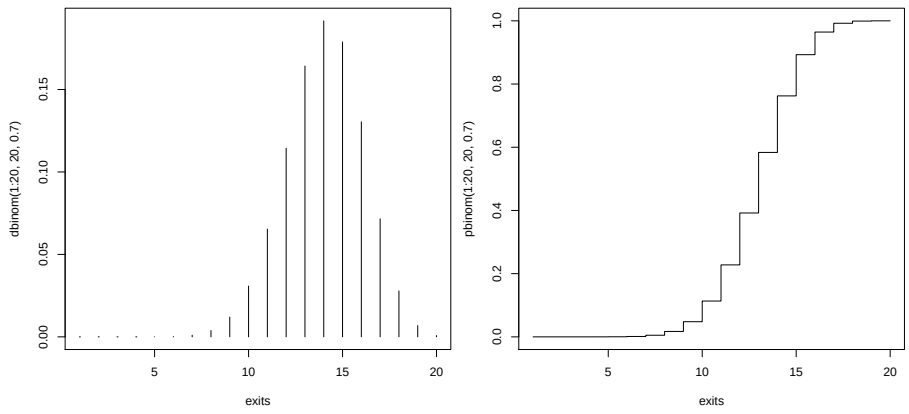
$$P(X=0) = \begin{pmatrix} 10000 \\ 0 \end{pmatrix} \cdot 0{,}001^0 \cdot 0{,}999^{10000} = 0{,}00004517.$$

Para el segundo caso tendríamos:

$$P(X \leq 10) = \sum_{x=0}^{10} \begin{pmatrix} 10000 \\ x \end{pmatrix} \cdot 0{,}001^x \cdot 0{,}999^{10000-x}.$$

Haciendo el cálculo con R obtenemos:

```
pbinom(10,10000,0.001)
[1] 0.5830398
# Función de masa de probabilidad de una B(20,0.7)
plot(dbinom(1:20,20,0.7),xlab="exits",type=''h'')
# Función de distribución de una B(20,0.7)
plot(pbinom(1:20,20,0.7),xlab="exits",type="S")
```



La distribución binomial está tabulada (existen tablas para consultar las probabilidades), de todos modos, cuando el valor de n es suficientemente grande, se puede aproximar una $\mathrm{B}(n,p)$ por una distribución de Poisson de parámetro $\lambda = np$. Esta aproximación se considera buena cuando $n > 30$ y $p < 0{,}1$. Además, si $n > 30$ y $0{,}1 < p < 0{,}9$, consideraremos buena la aproximación por una distribución normal $\mathrm{N}(np, \sqrt{np(1-p)})$ (Esto lo veremos en más detalle al final del tema, al hablar de las relaciones entre distribuciones).

2.5.3. Distribución geométrica

Consideramos nuevamente el experimento que consiste en repetir n veces un experimento de Bernoulli y tomamos la variable aleatoria $X =$ "número de fracasos antes de

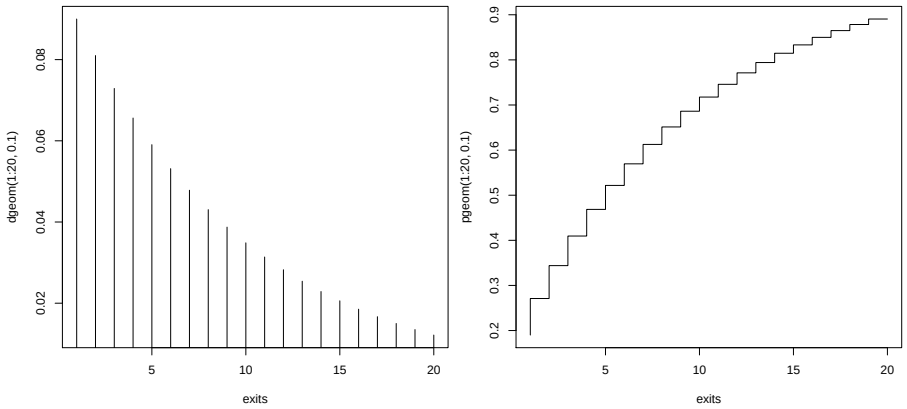
obtener el primer éxito". (Nuevamente se considera que la probabilidad de éxito en cada repetición viene dada por el parámetro p)

La **función de probabilidad** es de la forma:

$$P(X = x) = (1 - p)^x p, \quad x = 0, 1, ..., n.$$

Características: $E(X) = (1 - p)/p$, $\text{Var}(X) = (1 - p)/p^2$.

```
# Función de masa de prob. de una geométrica con prob. de acierto 0.1
plot(dgeom(1:20,0.1),xlab="exits",type="h")
# Función de distribución
plot(pgeom(1:20,0.1),xlab="exits",type="S")
```



2.5.4. Distribución binomial negativa

Es una generalización del caso anterior, consiste en estudiar la variable $X =$ "número de fracasos antes de obtener el éxito n". Se denota por $\text{BN}(n, p)$ (la geométrica es entonces una $\text{BN}(1,p)$)

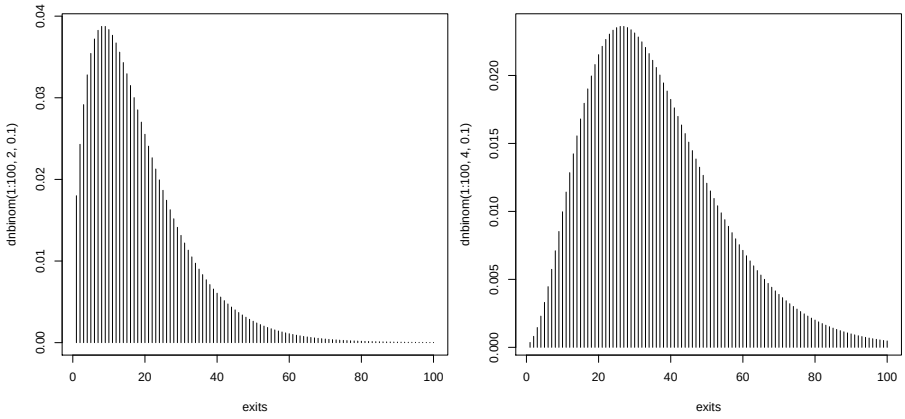
Su **función de probabilidad** es:

$$P(X = x) = \begin{pmatrix} n + x - 1 \\ x \end{pmatrix} p^n (1 - p)^x, \quad x = 0, 1, ..., n, \ldots$$

Características: $E(X) = n(1 - p)/p$, $\text{Var}(X) = n(1 - p)/p^2$.

Se utiliza por ejemplo en estudios de fiabilidad de sistemas.

```
# Función de masa de prob. de una binomial negativa con prob. de
acierto 0.1, en la que pedimos tener 2 aciertos.
plot(dnbinom(1:100,2,0.1),xlab="exits",type="h")
# Lo mismo pero ahora pedimos 4 aciertos
plot(dnbinom(1:100,4,0.1),xlab="exits",type="h")
```



2.5.5. Distribución de Poisson

Un proceso de Poisson generaliza en cierta manera al proceso de Bernoulli. Consiste en observar el número de veces que se presenta un suceso (número de éxitos) en un determinado intervalo (generalmente de tiempo). En estos procesos se asume que hay estabilidad, en el sentido de que el número de sucesos por unidad de tiempo (λ) permanece constante. Como ejemplos tendríamos "número de fallos superficiales en un cable de red por unidad de tiempo (o por unidad de superficie)", "espectadores que llegan a la cola de un cine",... De modo que, considerado un proceso de Poisson, la distribución de Poisson mide el "número de sucesos ocurridos en un intervalo".

La fórmula de la **función de distribución** de la distribución de Poisson es:

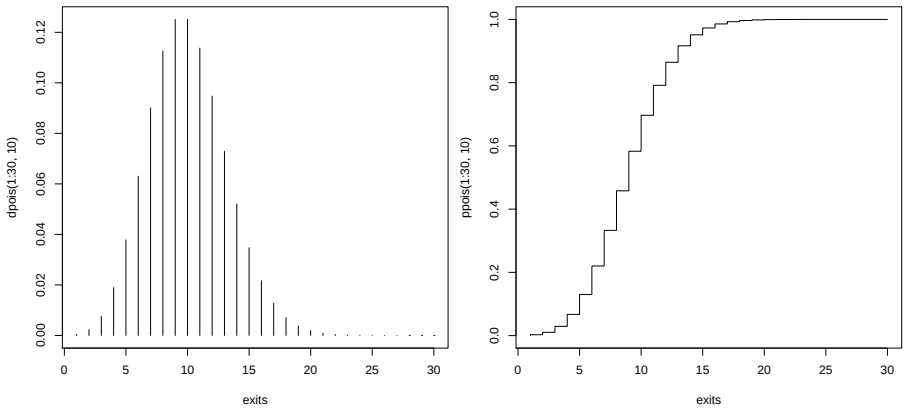
$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, ...$$

Características: $E(X) = \text{Var}(x) = \lambda$.

Dadas dos variables $X \in \text{Pois}(\lambda_1)$ e $Y \in \text{Pois}(\lambda_2)$ la variable $X + Y$ tiene una distribución $\text{Pois}(\lambda_1 + \lambda_2)$.

La distribución de Poisson se obtiene como límite de la binomial cuando $n \to \infty$ y $p \to 0$. Es decir, si repetimos una gran cantidad de veces un proceso con probabilidad muy pequeña de éxito, se podría utilizar la distribución de Poisson para obtener una buena aproximación del resultado (nótese que la distribución de Poisson es, en general, más fácil de calcular que la binomial debido al problema computacional de los números combinatorios).

```
# Función de masa de prob. de una Poisson de parámetro 10
plot(dpois(1:30,10),xlab="exits")
# Función de distribución de una Poisson de parámetro 10
plot(ppois(1:30,10),xlab="exits",type="S")
```



Cuando el valor del parámetro λ es mayor que 5, la Pois(λ) se puede aproximar por una normal N($\lambda, \sqrt{\lambda}$).

2.5.6. Distribución uniforme discreta

Una variable aleatoria X que toma los valores $\{x_1,\ldots,x_n\}$ se dice uniforme si todos ellos son equiprobables.

Función de probabilidad: $P(X = x) = \frac{1}{n}, \quad x = x_1, ..., x_n$.

Características: $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2$.

2.5.7. Distribución hipergeométrica

Si repetimos un experimento aleatorio del tipo "extraer una carta de una baraja", la variable aleatoria "número de oros obtenidos" puede estudiarse como una binomial siempre y cuando la carta extraída sea introducida de nuevo antes de repetir el experimento. Cuando esto no es así, y las extracciones se realizan sucesivamente pero sin reemplazamiento, es necesario recurrir a la distribución hipergeométrica.

Consideremos una población finita de N elementos, k de ellos de la clase D (oros en la baraja) y N- k del resto. Si ahora tomamos una muestra **sin** reemplazamiento y estudiamos la variable aleatoria X = "Número de elementos de la clase D en la muestra de tamaño n", esta sigue una distribución hipergeométrica H(N,n, k). Sea $p = k/N$ la probabilidad de obtener un elemento de la clase D en la primera extracción.

Función de probabilidad:

$$P(X = x) = \frac{\binom{k}{x} \cdot \binom{N-k}{n-x}}{\binom{N}{n}}, \quad \max\{0, n-(N-k)\} \leq x \leq \min\{k, n\}.$$

Características: $E(X) = np,\ \mathrm{Var}(X) = np(1-p)(N-n)/(N-1)$.

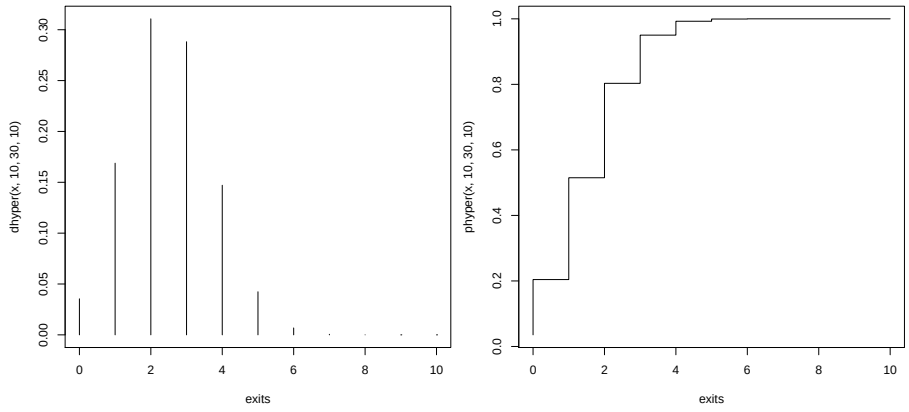
La distribución hipergeométrica se utiliza en el muestreo de una población finita sin reemplazamiento, por contraposición a la binomial que se utiliza cuando hay reemplazamiento. En el caso de que el tamaño de la población sea muy grande, la hipergeométrica se puede aproximar por la normal (la probabilidad de éxito apenas varía entre cada repetición del experimento).

EJEMPLO: Supongamos que tenemos una baraja y extraemos 10 cartas, queremos saber la probabilidad de extraer entre ellas 1,2,... 10 oros:

```
array(c(0:10,dhyper(0:10,10,30,10)),c(11,2))
      [,1]         [,2]
 [1,]    0  3.544463e-02
 [2,]    1  1.687840e-01
 [3,]    2  3.107159e-01
 [4,]    3  2.882003e-01
 [5,]    4  1.471022e-01
 [6,]    5  4.236544e-02
 [7,]    6  6.789333e-03
 [8,]    7  5.747584e-04
 [9,]    8  2.309297e-05
[10,]    9  3.539153e-07
[11,]   10  1.179718e-09
```

Gráficamente esto sería:

```
x<-c(0:10)
# Función de masa de prob. de una hipergeométrica H(40,10,10)
plot(x,dhyper(x,10,30,10),xlab="exits",type="h")
# Función de distribución
plot(x,phyper(x,10,30,10),xlab="exits",type="S")
```



Cuando N se hace muy grande, esta distribución se puede aproximar por la binomial, en general se considera buena esta aproximación cuando $n/N < 0{,}1$.

2.6. Principales distribuciones unidimensionales continuas

2.6.1. Distribución uniforme

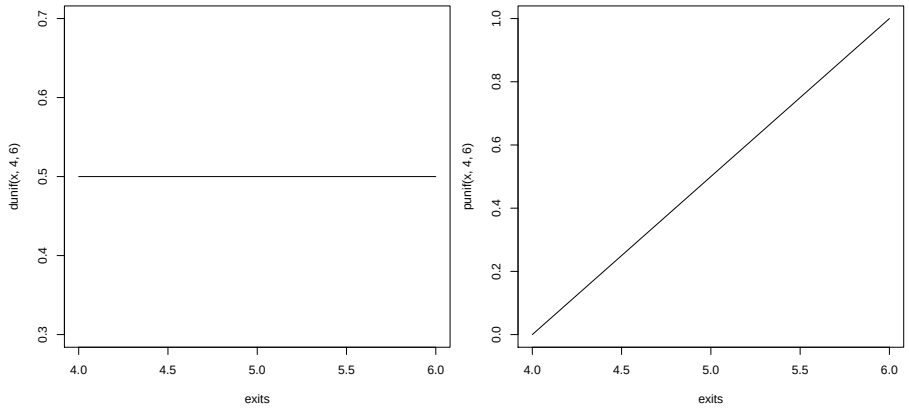
Como el propio nombre indica, una variable aleatoria sigue una distribución uniforme si todos los valores en el intervalo en el que está definida son igual de probables:

Función de densidad, $f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in (a,b), \\ 0 & \text{resto.} \end{cases}$

Función de distribución, $F(x) = \begin{cases} 0 & \text{si } x < a, \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b, \\ 1 & \text{si } x > b. \end{cases}$

Características: $E(X) = \frac{a+b}{2}$, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

```
x<-seq(4,6,0.05)
# Densidad de una uniforme en el intervalo [4,6]
plot(x,dunif(x,4,6),xlab="exits",type="l")
# Función de distribución
plot(x,punif(x,4,6),xlab="exits",type="l")
```



2.6.2. Distribución normal

La distribución normal es la más usada de todas las distribuciones de probabilidad. Gran cantidad de experimentos se ajustan a distribuciones de esta familia, por ejemplo el peso y la talla de una persona, los errores de medición... Además, si una variable es

el resultado de la suma de variables aleatorias independientes, es bastante posible que pueda ser aproximada por una distribución normal.

Una normal de media μ y varianza σ^2 se denota N(μ,σ).

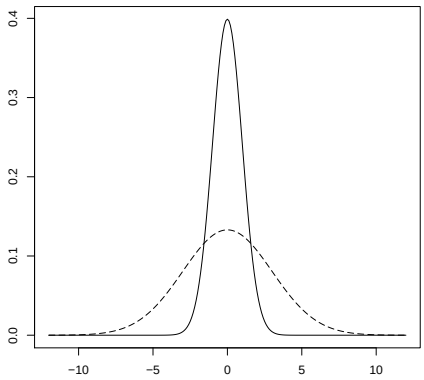
Función de densidad: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$.

A pesar de su complicada expresión, puede ser obtenida como límite de la binomial cuando n tiende a infinito. Cabe destacar que es una función simétrica, con un máximo en $x = \mu$ y puntos de inflexión en $x = \mu - \sigma$ y $x = \mu + \sigma$.

Características: $E(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.

Cuando definimos la curtosis dijimos que normalmente se hablaba de "apuntamiento" con respecto a una variable normal, sin decir nada de la media o varianza de la misma. Esto es porque todas las distribuciones normales tienen la misma curtosis.

```
# Rango en el que representaremos la función
x<-seq(-12,12,by=0.1)
# Representamos una N(0,1)
plot(x,dnorm(x,0,1),ylab="",xlab="",type="l")
# En trazo discontinuo una N(0,3)
lines(x,dnorm(x,0,3),lty=5)
```



El apuntamiento es una medida relativa, se trata de grado de apuntamiento con respecto a la dispersión. Es decir, si las dos distribuciones tuvieran la misma desviación típica, sus gráficas serían igual de apuntadas (por eso la fórmula de la curtosis tiene la desviación típica en el denominador).

Comprobemos a través de una simulación en R que las distribuciones N(0,1) y N(0,3) tienen el mismo apuntamiento. Como se observa en la simulación, el coeficiente de kurtosis calculado para las muestras generadas de ambas distribuciones es prácticamente el mismo.

```
#Función para el cálculo de la curtosis
curtosis<-function(x){
   momento.centrado(x,4)/(sd(x,na.rm=TRUE)\ 4*(length(x[!is.na(x)])-
   1)/length(x[!is.na(x)]))
}
# Simulamos 100000 valores de la N(0,1) y otros tantos de la N(0,3)
simdesv0=rnorm(100000,0,1)
simdesv3=rnorm(100000,0,3)
# Computamos sus curtosis y vemos que en ambos casos nos encontramos
muy cerca de 3 (valor teórico esperado)
curtosis(simdesv0)
[1] 2.990247
curtosis(simdesv3)
[1] 2.993902
```

Aunque este pequeño ejemplo no tiene valor teórico sirve para ilustrar que estas dos distribuciones tienen el mismo coeficiente de apuntamiento.

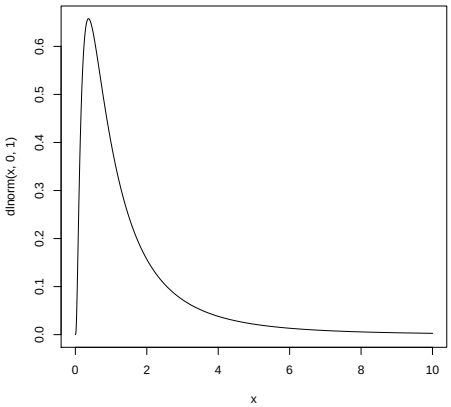
Si $X \in N(\mu_1, \sigma_1)$ e $X \in N(\mu_2, \sigma_2)$ entonces la variable $X + Y$ será del tipo $N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$. De aquí sacamos que la combinación lineal de variables aleatorias normales e independientes también es una variable aleatoria normal.

2.6.3. Distribución lognormal

Una variable X es lognormal si la variable $Y = \ln(X)$ sigue una distribución normal.

Función de densidad:

$$f(x) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} e^{\frac{-(\ln x - \mu)^2}{2\sigma^2}} & \text{si } x > 0, \\ 0 & \text{resto.} \end{cases}$$



Características:

$$E(X) = e^{(2\mu+\sigma^2)/2},$$

$$\mathrm{Var}(X) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}.$$

La distribución lognormal se usa principalmente en estudios de fiabilidad, para modelizar el tiempo de vida de materiales... Otra utilidad es para trabajar con variables relativas a rentas, ventas...

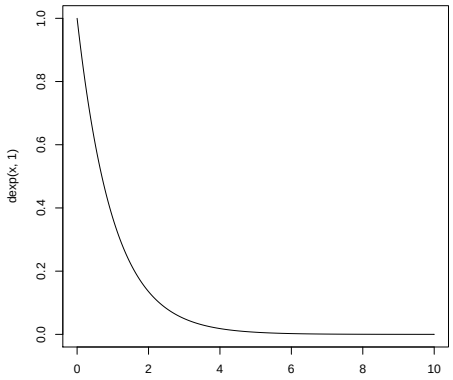
```
x<-seq(0,10,0.01)
plot(x,dlnorm(x,0,1),type="l")
```

2.6.4. Distribución exponencial

Un proceso de Poisson se utilizaba para medir el número de sucesos de un determinado tipo que tenían lugar en un determinado intervalo. Consideramos ahora la variable X que estudia el "tiempo entre dos sucesos consecutivos". Esta seguirá una distribución exponencial de parámetro λ y se denota por $\text{Exp}(\lambda)$. Podemos decir entonces que una distribución de Poisson mide el "número de sucesos por unidad de tiempo" y una exponencial el "tiempo que tarda en producirse un suceso".

Función de densidad:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0, \\ 0 & \text{en otro caso.} \end{cases}$$



Función de distribución: $F(x) = 1 - e^{-\lambda x}$.

Características:

$$E(X) = \frac{1}{\lambda},$$

$$\text{Var}(X) = \frac{1}{\lambda^2}.$$

Nótese que este valor para la media encaja con la motivación que dimos para la exponencial a través de la Poisson, si en una Poisson de parámetro λ se producen, en media λ sucesos por unidad de tiempo, cabe esperar que, en media, la separación entre sucesos sea $1/\lambda$ (así tenemos que $\lambda(1/\lambda) = 1$ unidad de tiempo).

La distribución exponencial es la generalización al caso continuo de la distribución geométrica y, al igual que esta tiene una importante propiedad que es la ausencia de memoria. Veámoslo con un ejemplo, supongamos que estamos midiendo la probabilidad de que un proceso en cadena tenga algún fallo debido al azar (no al desgaste). Si sabemos que en la primera hora no se ha producido ningún error y queremos saber la probabilidad de que en la segunda hora tampoco se produzca ninguno (esto es $P(X > 2/X > 1)$) esto se calcula directamente como la probabilidad de que durante una hora no haya ningún fallo (no importa que ya lleve una hora funcionando sin error, no tiene memoria). Escribiendo con más rigor esta propiedad: $P(X > t_0 + t_1/X > t_0) = P(X > t_1)$. Es un buen modelo para describir la aparición de fallos al azar, pero no para estudiar sistemas que se deterioran con el tiempo.

```
x<-seq(0,10,0.01)
plot(x,dexp(x,1),type="l")
```

2.6.5. Distribución gamma

Es una generalización de la distribución exponencial.

Llamaremos función gamma a: $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$, y tiene las siguientes propiedades:

1. $\Gamma(p) = (p-1)!$ si p es un número natural distinto de 0.

2. $\Gamma(p) = (p-1)\Gamma(p-1)$.

3. $\Gamma(1/2) = \sqrt{\pi}$.

Una variable X que mide el "tiempo de espera hasta el suceso número p en un proceso de Poisson" sigue una distribución gamma. Se denota por $\Gamma(\lambda,p)$, donde el primer parámetro representa el número medio de sucesos por unidad de tiempo y p es el número de sucesos que queremos que ocurran. De modo que una exponencial no será más que una $\Gamma(\lambda,1)$.

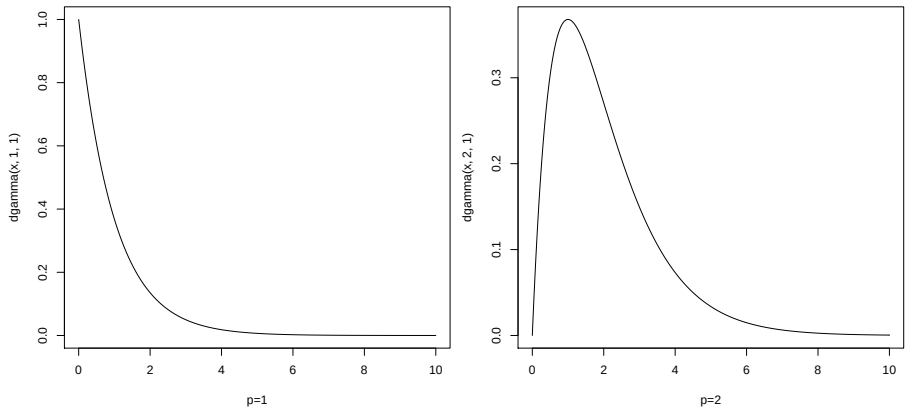
Función de densidad: $f(x) = \begin{cases} \frac{\lambda^p}{\Gamma(p)} e^{-\lambda x} x^{p-1} & \text{si } x > 0, \\ 0 & \text{resto.} \end{cases}$

Características:

$$E(X) = \frac{p}{\lambda},\ \mathrm{Var}(X) = \frac{p}{\lambda^2}.$$

Una importante propiedad de la distribución gamma es que, si $X \in \Gamma(\lambda,p_1)$ e $Y \in \Gamma(\lambda,p_2)$, entonces $X + Y \in \Gamma(\lambda,p_1 + p_2)$.

```
x<-seq(0,10,0.01)
# Parámetro p=1, coincide con la exponencial
plot(x,dgamma(x,1,1),xlab="p=1",type="l")
# Parámetro p=2, se ve que tarda más en decaer (pedimos
tiempo hasta el suceso p=2 en vez de p=1)
plot(x,dgamma(x,2,1),xlab="p=2",type="l")
```



2.6.6. Distribución de Erlang

Supone un importante caso particular de la distribución gamma. En la gamma permitíamos que el valor p fuese un número real cualquiera. En muchos modelos sólo interesa el caso en que p es un número entero positivo, es decir, no podemos partir sucesos (este

modelo surgió al modelizar el uso de las líneas telefónicas y estudiar las llamadas entrantes a un operador). Además de verla como un caso particular de la distribución gamma también se puede ver como una generalización de la exponencial; más concretamente, una Erlang de parámetro p es la suma de p exponenciales.

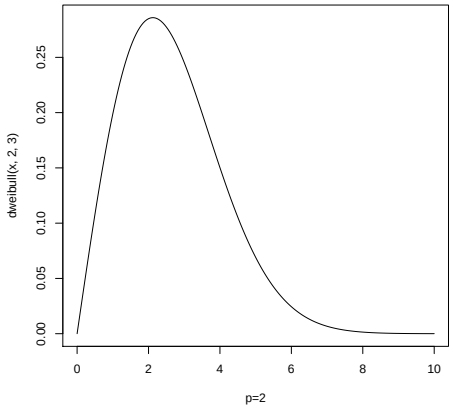
En este caso la **función de distribución** queda:

$$F(X) = 1 - \sum_{i=0}^{p-1} e^{-\lambda x} \frac{(\lambda x)^i}{i!}.$$

2.6.7. Distribución de Weibull

Diremos que una variable X sigue una distribución de Weibull W(b,λ) si tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \lambda^b b x^{b-1} e^{-(\lambda x)^b} & \text{si } x \geq 0, \\ 0 & \text{si } x < 0. \end{cases}$$



Función de distribución:

$$F(x) = \begin{cases} 1 - e^{-(\lambda x)^b} & \text{si } x \geq 0, \\ 0 & \text{si } x < 0. \end{cases}$$

Características:

$$E(X) = \frac{1}{\lambda}\Gamma(1 + \frac{1}{b}),$$

$$\text{Var}(X) = \frac{1}{\lambda^2}\left\{\Gamma(1 + \frac{2}{b}) - \left(\Gamma(1 + \frac{1}{b})\right)^2\right\}.$$

La distribución Weibull se utiliza en estudios de fiabilidad y para modelizar el tiempo de vida de diferentes materiales.

```
x<-seq(0,10,0.05)
plot(x,dweibull(x,2,3),xlab="p=2",type="l")
```

2.6.8. Distribución de tiempo de fatiga

También conocida con el nombre de distribución de **Birnbaum-Saunders**, surgió para modelizar la siguiente situación: tenemos un material al que continuamente se le va sometiendo a distintas presiones, éstas lo van desgastando poco a poco hasta que el material cede y se rompe. Por ejemplo, una pieza de una cadena de montaje. La distribución de tiempo de fatiga tiene como objetivo modelizar el tiempo que esta rotura tardará en producirse. Una variable X sigue una distribución de tiempo de fatiga, TF(μ,γ), si viene dada por una **función de densidad** del tipo $f(x) = h(x) \cdot \varphi(g(x))$, donde la letra griega φ denota la función de densidad de una N(0,1), más concretamente:

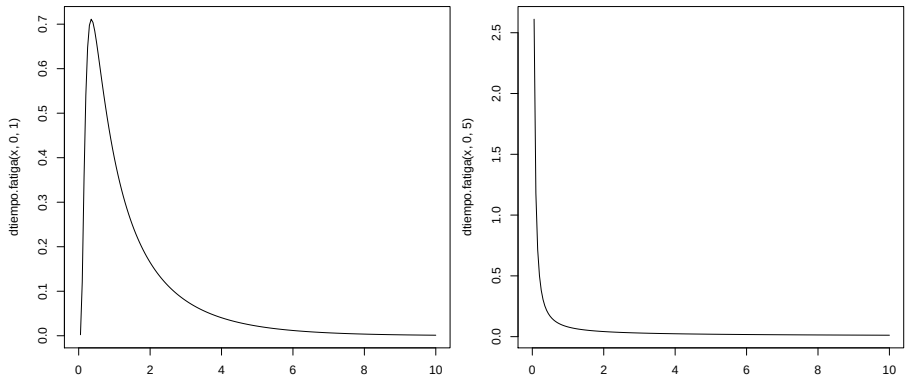
$$f(x) = \begin{cases} \frac{\sqrt{(x-\mu)}+\sqrt{\frac{1}{(x-\mu)}}}{2\gamma(x-\mu)} \cdot \varphi\left(\frac{\sqrt{(x-\mu)}-\sqrt{\frac{1}{(x-\mu)}}}{\gamma}\right) & \text{si } x \geq \mu, \\ 0 & \text{si } x < \mu. \end{cases}$$

El parámetro μ es el parámetro de localización, depende de la resistencia de la pieza en estudio y de la intensidad del desgaste al que es sometida. Por su parte el parámetro de forma, γ, depende de la variabilidad en las presiones que va sufriendo el material.

Características (cuando $\mu = 0$)**:**

$$E(X) = (1 + \frac{\gamma^2}{2}), \quad \text{Var}(X) = \gamma^2 \left(1 + \frac{5}{4}\gamma^2\right).$$

```
# Esta distribución no está definida en R, debemos definirla nosotros:
dtiempo.fatiga<-function(x,loc,shape){
  a<-(sqrt(x-loc)+ sqrt(1/(x-loc)))/(2*shape*(x-loc))
  b<-(sqrt(x-loc)-sqrt(1/(x-loc)))/shape
  a*dnorm(b)
}
x<-seq(0,10,0.05)
plot(x,dtiempo.fatiga(x,0,1),xlab="",type="l")
plot(x,dtiempo.fatiga(x,0,5),xlab="",type="l")
```



2.6.9. Distribución beta

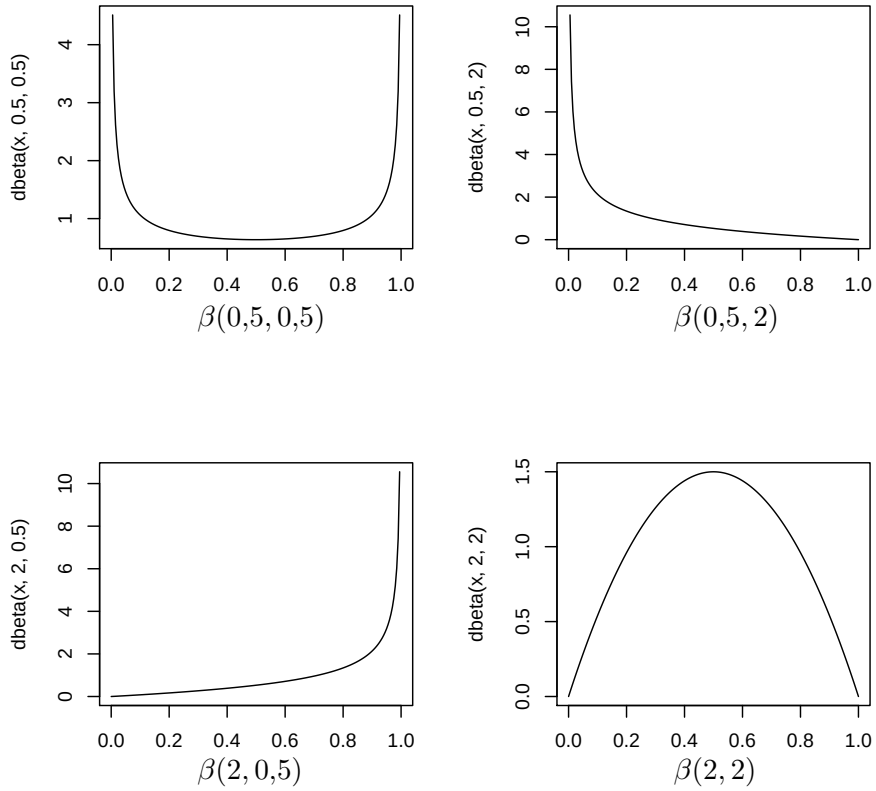
Llamaremos función beta a la aplicación: $\beta(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}$.

Diremos que una variable aleatoria X sigue una distribución $\beta(p,q)$ con p y q positivos si su **función de densidad** viene dada por

$$f(x) = \begin{cases} \frac{x^{p-1}(1-x)^{q-1}}{\beta(p,q)} & \text{si } 0 < x < 1, \\ 0 & \text{resto.} \end{cases}$$

Características: $E(X) = \frac{p}{p+q}$, $\text{Var}(X) = \frac{pq}{(p+q)^2(p+q+1)}$.
Se utiliza principalmente cuando se trabaja con estadística bayesiana.

```
x<-seq(0,1,0.005)
plot(x,dbeta(x,0.5,0.5),xlab="",type="l")
plot(x,dbeta(x,0.5,2),xlab="",type="l")
plot(x,dbeta(x,2,0.5),xlab="",type="l")
plot(x,dbeta(x,2,2),xlab="",type="l")
```



2.6.10. Distribuciones asociadas a la normal

Distribución χ^2 de Pearson

Dadas Z_1, ..., Z_n variables aleatorias independientes y N(0,1) entonces la variable

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2$$

sigue una distribución χ^2 con n grados de libertad (χ^2_n). Su **función de densidad** viene dada por

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} & \text{si } x > 0, \\ 0 & \text{resto.} \end{cases}$$

que no es más que una distribución gamma con parámetros $\lambda = 1/2$ y $p = n/2$. Por tanto, la esperanza de χ^2_n es n y su varianza es $2n$.

Como propiedad destacable cabe destacar que la variable $\sqrt{2\chi^2_n} \approx N\left(\sqrt{2n-1}, 1\right)$ aunque esta aproximación es lenta.

Distribución t de Student

Sea $Z \in$N(0,1) e $Y \in \chi^2_n$ variables aleatorias independientes. La distribución t de Student con n grados de libertad se define como la distribución de la variable

$$t_n = \frac{Z}{\sqrt{Y/n}}.$$

Su **función de densidad** viene dada por

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

Características: $E(t_n) = 0,\ n > 1$, $\text{Var}(t_n) = \frac{n}{n-2}$, $n > 2$.

Esta distribución se aproxima a una N(0,1) cuando n es grande y presenta más dispersión y menor curtosis que esta. En el caso particular de $n = 1$ se llama distribución de Cauchy y no tiene media.

Distribución F de Fisher-Snedecor

Sea $X \in \chi^2_n$ e $Y \in \chi^2_m$ dos variables aleatorias independientes. Se define la distribución F con (n,m) grados de libertad como la distribución de la variable

$$F_{n,m} = \frac{X/n}{Y/m}.$$

Función de densidad:

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right) \left(\frac{n}{m}x\right)^{\frac{n}{2}-1} \left(1 + \frac{n}{m}x\right)^{-\frac{n+m}{2}}, \forall x > 0.$$

Características:

$$E(F_{n,m}) = \frac{m}{m-2},\ m > 2, \quad \text{Var}(F_{n,m}) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}\ ,\ m > 4.$$

Ademas si $X \in F_{n,m}$ entonces $1/X \in F_{m,n}$.

2.7. Variables aleatorias multidimensionales

A veces no es suficiente con saber como se distribuyen una a una las características de una determinada población sino que se necesita estudiar cómo se distribuyen todas ellas conjuntamente. Esto lleva a la aparición de los **vectores aleatorios** como generalización natural de las variables aleatorias.

Ejemplo: Consideremos un proceso de control de calidad en el que por un lado tenemos el número de productos defectuosos y por otro el tiempo de fabricación de cada producto. En este caso un análisis conjunto de ambas variables aportaría mucha más información que un estudio independiente de las mismas.

En este tema nos dedicaremos principalmente a estudiar el caso bidimensional, pudiéndose extender todos los conceptos y resultados al caso n-dimensional. Más concretamente nos centraremos en el caso de que tengamos variables aleatorias tales que ambas sean bien continuas o bien discretas.

Dado un espacio de probabilidad $(\Omega,\mathcal{A},P)$, la **variable (vector) aleatoria bidimensional (X,Y)** se define como una función que asocia a cada suceso elemental un par de números reales, siendo cada componente (X e Y) variables aleatorias unidimensionales sobre $(\Omega,\mathcal{A},P)$.

2.7.1. Función de distribución de una variable aleatoria bidimensional

La **función de distribución conjunta** de una variable aleatoria bidimensional es la función que a cada par de números reales (x,y) les asigna el valor $F(x,y) = P(X \leq x, Y \leq y)$.

1. $0 \leq F(x) \leq 1$.
2. F es continua por la derecha y no decreciente en cada componente.
3. $F(+\infty,+\infty) = \lim_{x,y\to+\infty} F(x,y) = 1$.
4. $F(-\infty,y) = F(x,-\infty) = 0$.

Caso discreto

Dadas dos variables aleatorias discretas X e Y, definidas sobre el mismo espacio de probabilidad, la **función de masa de probabilidad conjunta** de la variable aleatoria bidimensional discreta (X,Y) es la que asigna una probabilidad a cada una de las posibles realizaciones de la variable conjunta:

$$p(x_i,y_j) = P(\omega \in \Omega / X(\omega) = x_i, Y(\omega) = y_j), \text{ que verifica } \sum_{x_i}\sum_{y_j} p(x_i,y_j) = 1.$$

La fórmula de la función de distribución para este caso queda

$$F(x,y) = \sum_{x_i \leq x}\sum_{y_j \leq y} p(x_i,y_j).$$

Caso continuo

Dadas dos variables aleatorias continuas X e Y, definidas sobre el mismo espacio de probabilidad, la **función de densidad conjunta** de la variable aleatoria bidimensional continua (X,Y) es la función f definida como $f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$, esta función puede ser definida también como aquella que verifica:

1. $f(x,y) \geq 0$.
2. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)dxdy = 1$.
3. $F(x_0,y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f(x,y)dxdy$.

Esta función cumple además que

$$P(X = x_0, Y = y_0) = \int_{x_0}^{x_0} \int_{y_0}^{y_0} f(x,y)dxdy = 0,$$

es decir, los valores puntuales tienen probabilidad 0, y

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x,y)dydx.$$

2.7.2. Distribuciones marginales

Esto nos permite hacer el proceso inverso al descrito en el momento en el que se introdujeron los vectores aleatorios. Si tenemos la distribución conjunta del vector (X,Y) y estamos interesados en estudiar la variable X por separado, necesitamos recurrir a las distribuciones marginales.

Dada una variable aleatoria bidimensional (X,Y), con función de distribución $F(x,y)$, se pueden definir dos funciones de distribución univariantes, $F_X(x)$ y $F_Y(y)$. Por comodidad pondremos sólo las fórmulas de relativas a la marginal de X.

Función de distribución marginal de la variable X:

$$F_X(x) = P(X \leq x) = P(X \leq x, -\infty \leq Y \leq +\infty).$$

Caso discreto:

$$F_X(x) = \sum_{x_i \leq x} \sum_{y_j} P(X = x_i, Y = y_j).$$

Función de probabilidad marginal de X:

$$p_X(x_i) = P(X = x_i) = \sum_{y_j} P(X = x_i, Y = y_j).$$

Caso continuo:

$$F_X(x_0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{x_0} f(x,y)dxdy.$$

Función de densidad marginal de X:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy.$$

Observaciones:

1. Distintas funciones de densidad conjunta pueden dar lugar a las mismas marginales.
2. La idea intuitiva de marginal se puede ver claramente en el primer tema, en la descripción estadística de varias variables. En la representación en formato de tabla, la marginal de la X se obtiene haciendo tomar a la Y todos los posibles valores para cada valor fijado de la X (y sus valores se ponen al margen).

2.7.3. Distribuciones condicionadas

Trabajando con la idea de probabilidad condicionada, se pueden obtener nuevas distribuciones a partir de una variable conjunta, condicionando los valores de una variable a posibles realizaciones de la otra.

Caso discreto

Dada una variable bidimensional discreta (X,Y), la **función de probabilidad condicionada** de la variable X, dado el suceso y de la variable Yserá

$$p_{X|Y=y}(x) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P_Y(y)}, \text{ dado que } P_Y(y) > 0.$$

Caso continuo

Dada una variable bidimensional continua (X,Y), la **función de densidad condicionada** para X, dado el suceso y de la variable Y será

$$f_{X|Y=y}(x) = f(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{si } f_Y(y) > 0, \\ 0 & \text{si } f_Y(y) = 0. \end{cases}$$

Ejemplo: Dada la variable bidimensional (X,Y) con densidad conjunta dada por

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{si } x > 0,\ y > 0, \\ 0 & \text{resto,} \end{cases}$$

calcular la distribución marginal de X y la de Y condicionada a X.

SOLUCIÓN: La función de densidad marginal de X se calcula como,

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy.$$

Teniendo en cuenta la expresión de la densidad conjunta $f(x,y)$ se tiene que

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = e^{-x} \int_{0}^{+\infty} e^{-y}dy = e^{-x}.$$

Por lo tanto

$$f_X(x) = \begin{cases} e^{-x} & \text{si } x > 0, \\ 0 & \text{resto.} \end{cases}$$

Análogamente

$$f_Y(y) = \begin{cases} e^{-y} & \text{si } y > 0, \\ 0 & \text{resto.} \end{cases}$$

La distribución marginal de Y condicionada a X se calcula como

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{e^{-(x+y)}}{e^{-x}} = e^{-y},$$

por tanto la distribución condicionada queda

$$f(y|x) = \begin{cases} e^{-y} & \text{si } y > 0, \\ 0 & \text{resto.} \end{cases}$$

En este caso vemos que $f(y|x) = f_Y(y)$, esto es lo que pasa cuando ambas variables son independientes, que es el siguiente punto de este tema.

2.7.4. Independencia de variables aleatorias

La idea de independencia entre variables aleatorias es la misma que en su momento vimos al definir la independencia entre sucesos. Si el conocimiento de la realización de una variable aleatoria X no influye para nada en el posible resultado de otra variable Y, decimos que ambas son independientes. De acuerdo a esta idea, dos variables serán independientes si las distribuciones marginales y las condicionadas coinciden.

Sea (X,Y) una variable aleatoria bidimensional, se dirá que ***X* es independiente de *Y*** si la distribución conjunta es igual al producto de las distribuciones marginales,

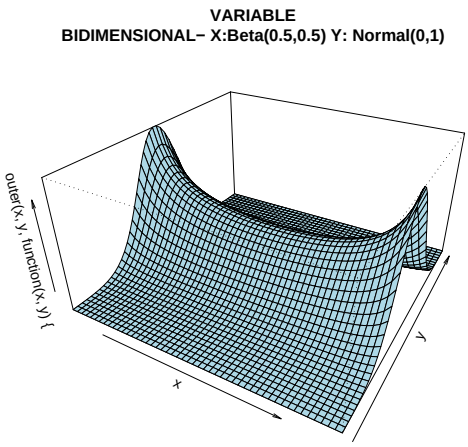
$$F(x,y) = F_X(x) \cdot F_Y(y).$$

Caso discreto, en este caso X e Y son independientes si para cada posible realización (x,y) de la variable conjunta tenemos $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$.

Caso continuo, X e Y son independientes si para toda realización (x,y) de la variable conjunta tenemos $f(x,y) = f_X(x) \cdot f_Y(y)$. Equivalentemente, X e Y son independientes si para todo valor y de la variable Y tenemos que $f(x|y) = f_X(x)$.

EJEMPLO: Ahora somos capaces de construir la distribución conjunta de dos variables X e Y independientes conocidas, basta tomar el producto de las marginales para cada posible realización.

Dibujamos a continuación la variable (X,Y) donde X e Y son independientes, la primera una distribución beta, β(0.5,0.5), y la segunda una distribución normal, N(0,1).



El código R sería:

```
x<-seq(0.05,0.95,length=50)
y<-seq(-4,4,length=50)
persp(x, y, outer(x,y,function(x,y){dbeta(x,0.5,0.5)*dnorm(y)}),
theta = 30,phi = 30, expand = 0.5, col = "lightblue", main="VARIABLE
BIDIMENSIONAL- X:Beta(0.5,0.5) Y: Normal(0,1)")
```

2.7.5. Medidas características de una variable aleatoria bidimensional

A continuación generalizaremos los conceptos más relevantes del caso unidimensional para el contexto actual.

- **Esperanza matemática de un vector aleatorio.** Dada una variable aleatoria bidimensional (X,Y), se denomina vector de medias de la variable conjunta al vector ($E(X)$,$E(Y)$).

- **Matriz de varianzas-covarianzas.** El cálculo de las varianzas de las variables X e Y se hace utilizando la correspondiente definición obtenida para el caso unidimensional.

  Dadas dos variables X e Y podemos estar interesados en estudiar la varianza conjunta de ambas variables, ver qué hace la variable Y cuando X aumenta,... Para esto se utiliza la **covarianza**, que se define como el número real:

  $$\mathrm{Cov}(X,Y) = \sigma_{XY} = \sigma_{YX} = E\left((X - E(X))\,(Y - E(Y))\right).$$

  La matriz de varianzas-covarianzas, Σ, resume toda la información relativa a las mismas:

  $$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.$$

  La covarianza también se puede definir como $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$.

  Propiedades:

  1. Si la covarianza tiene valor positivo quiere decir que cuando una de las variables crece, también la otra tiende a crecer. Si por el contrario el signo es negativo, quiere decir que las variables van en direcciones opuestas.

2. Si dos variables son independientes su covarianza es cero y por lo tanto se tiene que $E(XY) = E(X)E(Y)$. El recíproco sin embargo no es necesariamente cierto, dos variables pueden tener covarianza 0 y no ser independientes.
3. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$.
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$,
   $\text{Var}(X\text{- } Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y)$.

De estas fórmulas se deduce que la varianza de las variables no tiene porqué aumentar al sumarlas, esto es porque las variabilidades se pueden compensar unas con otras.

- **Matriz de correlaciones.** La covarianza nos mide el grado de dependencia entre dos variables en término absolutos, depende de la escala en que estemos trabajando. Por tanto, si queremos comparar dos covarianzas distintas para ver dónde hay una mayor dependencia entre las variables, podemos llevarnos a engaño. Al igual que se definió en su momento el coeficiente de variación como una medida adimensional (no depende de las unidades de medida) de la variabilidad de una variable aleatoria, ahora se define el **coeficiente de correlación** como una medida de tipo relativo para la **variabilidad conjunta** de dos variables aleatorias.

  Dada la variable bidimensional (X,Y), el **coeficiente de correlación lineal**, ρ, se define como:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

  Este coeficiente cumple las siguientes propiedades:

  1. $-1 \leq \rho_{XY} \leq 1$.
  2. Si $\rho_{XY} = 0$ ambas variables son incorreladas (no necesariamente independientes). No hay relación lineal entre ellas, pero puede haber alguna relación no lineal.
  3. Si $\rho_{XY} = 1$ entonces $Y = aX+b$ con $a > 0$ (relación lineal positiva), del mismo modo $\rho_{XY} = -1$ implica que $Y = aX+b$ sólo que ahora $a < 0$. En esta propiedad se puede apreciar que este coeficiente no depende de la medida, $Y = 100X$ tienen correlación $\rho_{XY} = 1$, al igual que las variables $Z = 2X$ ($\rho_{XZ} = 1$).

  Al igual que hacía la matriz de varianzas-covarianzas, la **matriz de correlaciones** resume la información de las mismas. Dadas dos variables X e Y la matriz de correlaciones R, es la matriz

$$R = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{pmatrix}.$$

- **Esperanza condicionada.** Dadas dos variables aleatorias X e Y, la **esperanza de Y condicionada a X** es la función que nos dice como se distribuye la variable Y para cada valor de X. Veremos la definición sólo para el caso continuo.

  Sean X e Y dos variables aleatorias continuas, se define la esperanza condicionada de Y, sabiendo que $X = x$ como $E(Y|X = x) = \int_{-\infty}^{+\infty} yf(y|X = x)dy$.

  Esta función nos dice, dado un valor de la variable X, lo que cabe esperar de la variable Y.

- **Función de regresión.** Hemos visto como el coeficiente de correlación nos indica la posible relación existente entre dos variables aleatorias. Cuando este coeficiente toma valores próximos a 1 ó -1 sabemos que existe una fuerte dependencia lineal entre las mismas, lo cual nos permite predecir los valores de una de las variables a partir de los que haya tomado la otra. Por ejemplo, dada la longitud l de un cuadrado su perímetro tomará el valor $4l$. En este caso la dependencia es total y el valor de la longitud de un lado nos permite determinar con toda precisión el valor del perímetro. En otras muchas situaciones esta relación, aún existiendo puede no ser tan clara, el peso de una persona está relacionada con su altura, aunque la relación no es tan fuerte como en el caso anterior. En general esto se hace a través de la llamada **función de regresión**. Dadas dos variables aleatorias X e Y, intentaremos utilizar la primera de ellas para predecir los valores de la segunda. Nos restringiremos al caso de variables continuas. En este contexto la variable X se llamará variable explicativa o independiente y la Y será la variable respuesta o dependiente. Por tanto, buscaremos un modelo explicativo con la siguiente forma

$$Y = g(X) + e,$$

  donde e denota el error en la predicción (que a su vez será una variable aleatoria). Para cada valor x de la variable X, se predice un valor y de la variable Y. En esta predicción se comete un error. Se trata de buscar una función g tal que el error esperado en la predicción sea mínimo. Buscamos entonces la función minimizando el valor esperado del error cuadrático

$$\begin{aligned} \min_g E[(Y - g(X))^2] &= \min_g \int_{\mathrm{R}^2} (y - g(x))^2 f(x, y)dxdy \\ &= \min_g \int_{\mathrm{R}^2} (y - g(x))^2 f(y|x)f(x)dydx. \end{aligned}$$

  En la integral de la última de las igualdades vemos aparecer la esperanza de Y condicionada a X. Siguiendo con los cálculos, se puede comprobar que el valor de la función g que minimiza este error esperado para cada valor de x es $g(x) = E(Y \mid X = x)$, es decir, dado un valor de la variable X, el mejor predictor para la variable Y no es otro que la esperanza de Y condicionada a ese valor de X.

- **Momentos.** Al igual que en el caso de las variables unidimensionales, también podemos definir los momentos que, entre otras cosas, son una herramienta de gran utilidad a la hora de conseguir resultados teóricos. Dada una variable bidimensional (X,Y) definimos:

  Momento con respecto del origen de orden (r,s):

  $$a_{rs} = E(X^r Y^s).$$

  Momento con respecto de la media de orden (r,s):

  $$m_{rs} = E\left((X - E(X))^r (Y - E(Y))^s\right).$$

  Al igual que hicimos en el caso unidimensional, también aquí se podrían obtener expresiones con integrales o sumatorios para los casos continuo y discreto, respectivamente.

2.7.6. Transformaciones de variables bidimensionales

Del mismo modo que en el caso unidimensional, también aquí es importante estudiar que pasa si dada una variable (X,Y) queremos estudiar una determinada transformación de la misma. Consideraremos sólo el caso en el que (X,Y) es una variable continua $g(X,Y) = (Z,T)$ es una nueva variable bidimensional (esto en general no tiene porqué suceder), por tanto tendremos una función g tal que $g(x,y) = (g_1(x,y), g_2(x,y)) = (z,t)$, verificando que tanto g_1 como g_2 admiten inversas u_1, u_2, ambas diferenciables. Por tanto tenemos $x = u_1(z,t)$ e $y = u_2(z,t)$.

Esto dará lugar al jacobiano $J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial t} \end{vmatrix}$.

Y la densidad w de la nueva variable (Z,T) se calcula como

$$w(z,t) = f(u_1(z,t), u_2(z,t))\,|J|\,.$$

Una vez que tenemos la densidad de la nueva variable ya estamos en condiciones de estudiar sus características y propiedades.

2.7.7. Caso n-dimensional

Todos los conceptos que hemos definido para el caso bidimensional pueden extenderse inmediatamente a la situación en la que nuestras variables están en un espacio cualquiera de dimensión $n > 1$. A continuación damos un pequeño apunte de cómo serían algunas de estas extensiones para el caso de una variable n-dimensional $(X_1, X_2, \ldots, X_n)$:

1. **Vector de medias:** $(E(X_1), E(X_2), \ldots, E(X_n))$.
2. **Covarianzas dos a dos,** covarianza entre X_i y X_j**:**
   $\sigma_{ij} = E\left((X_i - E(X_i))\,(X_j - E(X_j))\right)$.

3. **Matriz de varianzas-covarianzas**:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

Nótese que esta matriz es simétrica.

Del mismo modo se podrían generalizar las **correlaciones** y la **matriz de correlaciones**.

2.8. Modelos multidimensionales de distribución de probabilidad

2.8.1. Distribución multinomial

Esta distribución generaliza a la binomial, estudiada en el caso unidimensional. En este caso también se realizan n repeticiones independientes del mismo experimento. La diferencia radica en que los resultados del mismo no son dicotómicos (verdadero-falso), sino que en cada repetición tenemos como resultados posibles los sucesos $A_1, A_2, ..., A_k$, con probabilidades $p_1, p_2, ..., p_k$ respectivamente, verificando que $p_1 + p_2 + ... + p_k = 1$. La variable $X = (X_1, X_2, ..., X_k)$, donde $X_i =$ "número de veces que ha ocurrido el suceso A_i en las n repeticiones" se denomina **multinomial** y se denota por $M_k(n; p_1, ..., p_k)$. Su **función de probabilidad** viene dada por:

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

Características: $E(X_i) = np_i$, $\mathrm{Var}(X_i) = np_i$ (1- p_i), $\mathrm{Cov}(X_i, X_j) = -np_ip_j$.

2.8.2. Distribución normal multidimensional

Como ya hicimos notar en el momento de introducir la distribución normal unidimensional, se trata de la distribución más usada en estadística. Sus buenas propiedades se hacen todavía más patentes cuando se estudian modelos estadísticos multidimensionales. Ahora introduciremos la generalización de la distribución normal y citaremos sus propiedades más relevantes.

Un vector aleatorio $X = (X_1,\ldots,X_n)$ se distribuye según una **normal n-dimensional** con vector de medias μ y matriz de varianzas-covarianzas Σ si su **función de densidad** es de la forma:

$$f(x) = \frac{1}{|\Sigma|^{1/2}\,(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right\},\ x \in \mathbb{R}^n.$$

A pesar de su complejidad se observa fácilmente que es una generalización natural de la fórmula para el caso unidimensional. El papel de la varianza lo juega ahora la matriz de varianzas-covarianzas y el vector de medias hace lo que antes hacía la media.

Propiedades:

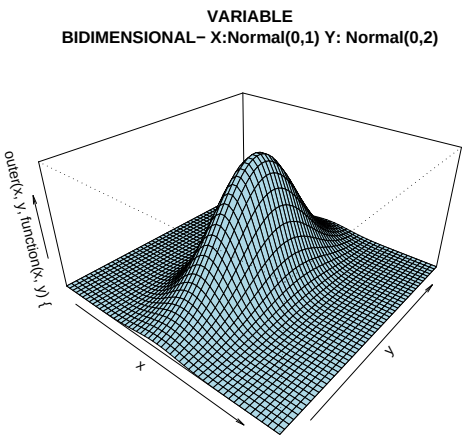
1. Las distribuciones condicionales y marginales de un vector normal n-dimensional también siguen distribuciones normales.

2. Para vectores normales, dos componentes son independientes si y sólo si son incorreladas (recuérdese que esto en general no tiene por qué ser cierto).

3. Cualquier combinación lineal de variables normales independientes da lugar a una nueva distribución normal.

Ejemplo: En el caso de que tengamos una normal bidimensional (X,Y) con matriz de correlaciones $R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, entonces ambas variables son incorreladas y por tanto independientes. Esto nos permite escribir la densidad conjunta como el producto de las densidades marginales.

Representación de una variable normal bidimensional (X,Y), con $X \in$ N$(0,1)$ e $Y \in$N(0,2), ambas independientes, usando la propiedad de que la densidad conjunta se puede poner como producto de densidades.



En el gráfico se puede ver como la variable Y tiene más variabilidad que la X, que está más concentrada en torno a la media (pues en el gráfico ambos ejes están en la misma escala).

El código R sería:

```
x<-seq(-5,5,length=50)
y<-seq(-5,5,length=50)
persp(x, y, outer(x,y,function(x,y){dnorm(x)*dnorm(y,0,2)}),
theta = 40, phi = 30, expand = 0.5, col = "lightblue", main="VARIABLE
BIDIMENSIONAL- X:Normal(0,1) Y: Normal(0,2)")
```

Comentario:

El concepto de tipificación $Z = \frac{X-\mu}{\sigma}$ descrito para variables unidimensionales también tiene su equivalente en el caso general. Tendremos que hacer una traslación utilizando el vector de medias y valernos de la matriz de varianzas-covarianzas para hacer el cambio de escala. Aunque esto no es tan fácil, ya que en el caso unidimensional nos bastaba con dividir por la raíz cuadrada de la varianza y aquí lo que tenemos es una matriz. En el caso particular de la distribución normal esto tiene fácil solución.

Supongamos que queremos obtener una muestra aleatoria de una variable normal $X = (X_1,\ldots,X_p)$ con vector de medias μ y matriz de varianzas-covarianzas Σ, valiéndonos

de distribuciones N(0,1) (fáciles de generar aleatoriamente). En este caso tenemos la siguiente propiedad, si encontramos una matriz L tal que $\Sigma = LL^t$, entonces si $Z = (Z_1,\ldots,Z_p)$ es un vector de normales estándar independientes, la variable $X = \mu + LZ$ tiene distribución N(μ,Σ), como estábamos buscando. De modo que para simular la variable X nos bastaría con simular p normales N(0,1) y aplicarles la correspondiente transformación. La matriz L juega ahora el papel que en las tipificaciones unidimensionales jugaba la desviación típica (es, en cierto modo, la raíz de la matriz de varianzas-covarianzas). El único problema técnico es cómo hallar esta matriz L, aunque hay multitud de algoritmos eficientes que nos permiten calcularla.

2.9. Sucesiones de variables aleatorias

Hasta ahora hemos estudiado los conceptos de variable aleatoria unidimensional X y vector aleatorio $(X_1, ..., X_n)$, una n-tupla de variables aleatorias. Ahora necesitaremos trabajar con cantidades infinitas de variables, por eso introducimos el concepto de **sucesión de variables aleatorias** $\{X_n\}_{n=1}^{\infty}$, teniendo como objetivo principal estudiar la convergencia de las mismas. En este contexto hemos de definir criterios de convergencia que deben ser distintos de aquellos del análisis matemático, debido al carácter aleatorio de estas sucesiones. Dada la sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$, con funciones de distribución $\{F_n\}_{n=1}^{\infty}$, y la variable aleatoria X con función de distribución F, tenemos los siguientes tipos de convergencia de la sucesión a la variable X:

Convergencia en distribución:

$$X_n \xrightarrow{D} X \Leftrightarrow \lim_{n\to\infty} F_n(x) = F(x)\ \forall x, \text{ punto de continuidad de } F.$$

Las funciones de distribución han de converger puntualmente (donde hay continuidad).

Convergencia en probabilidad:

$$X_n \xrightarrow{P} X \Leftrightarrow \lim_{n\to\infty} P(\omega \in \Omega / \, |X_n(\omega) - X(\omega)| < \varepsilon) = 1,\ \forall \varepsilon > 0.$$

La probabilidad de que los valores de las variables disten más de un determinado valor ha de tender a 0.

Convergencia casi segura:

$$X_n \xrightarrow{c.s.} X \Leftrightarrow P(\omega \in \Omega / \lim_{n\to\infty} X_n(w) = X(w)) = 1.$$

Ahora se pide que haya convergencia puntual entre las variables con probabilidad 1.

Veamos una pequeña explicación de la diferencia entre estos dos últimos tipos de convergencia. Definamos el suceso $A_n = \{\omega \in \Omega / \, |X_n(\omega) - X(\omega)| > \varepsilon\}$, lo que nos dice la convergencia en probabilidad es que $P(A_n) \to 0$, a medida que n aumenta, pero ojo ¡Esto no implica ninguna convergencia puntual!, es decir, podemos tener $P(A_{n+1}) < P(A_n)$, sin que estos dos sucesos tengan sucesos elementales en común. Sin embargo, la convergencia

casi segura añadiría a mayores la condición[1] $A_{n+1} \subseteq A_n$, para asegurar que tenemos convergencia puntual con probabilidad 1.

La relación entre estas tres convergencias es la siguiente:

$$X_n \xrightarrow{c.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X.$$

2.9.1. Leyes de los Grandes Números

Ley Débil de los Grandes Números

Diremos que una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ satisface la Ley Débil de los Grandes Números si existe una sucesión de números reales $\{A_n\}_{n=1}^{\infty}$, y otra de números positivos $\{B_n\}_{n=1}^{\infty}$ tal que $\lim_{n\to\infty} B_n = +\infty$ verificando que

$$\frac{\sum_{k=1}^{n} X_k - A_n}{B_n} \xrightarrow{P} 0.$$

El papel de la sucesión A_n es el de asegurar que la media de la sucesión límite sea 0, es decir, sirve para controlar la "centralización". Por su parte la sucesión B_n tiene una doble función; por un lado controla la escala y por otro evita la divergencia de la serie. En gran medida esto es parecido a lo que se hace al tipificar una variable aleatoria, en este caso teníamos que μ era el parámetro de localización y σ el de escala.

Como ejemplo tenemos el resultado de Markov que nos dice que si la sucesión $\{X_n\}_{n=1}^{\infty}$, cumple que todas sus componentes son incorreladas dos a dos, todas las medias y varianzas son finitas y además $\lim_{n\to\infty} \frac{1}{n^2} \sum_{k=1}^{n} \sigma_k^2 = 0$ entonces se satisface la Ley Débil para $\{X_n\}_{n=1}^{\infty}$ con $A_n = \sum_{k=1}^{n} E(X_k)$, y $B_n = n$. Es decir, toda sucesión de variables aleatorias en estas condiciones verifica que

$$\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} E(X_k)}{n} \xrightarrow{P} 0.$$

Aquí estamos diciendo que la variable aleatoria que se obtiene en cada paso tiende a la variable aleatoria degenerada en 0.

Ley Fuerte de los Grandes Números

Diremos que una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ satisface la Ley Fuerte de los Grandes Números si existe una sucesión de números reales $\{A_n\}_{n=1}^{\infty}$, y otra de números positivos $\{B_n\}_{n=1}^{\infty}$ tal que $\lim_{n\to\infty} B_n = +\infty$ verificando que

$$\frac{\sum_{k=1}^{n} X_k - A_n}{B_n} \xrightarrow{c.s.} 0.$$

[1] Esta condición no está escrita con todo rigor. Se ha puesto de esta manera para proporcionar una mejor intuición de lo que está pasando.

2.9.2. Teorema Central del Límite

Diremos que una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ satisface el Teorema Central del Límite si existe una variable aleatoria X, una sucesión de números reales $\{A_n\}_{n=1}^{\infty}$, y otra de números positivos $\{B_n\}_{n=1}^{\infty}$ tal que $\lim_{n\to\infty} B_n = +\infty$ verificando que

$$\frac{\sum_{k=1}^{n} X_k - A_n}{B_n} \xrightarrow{D} X.$$

Esto da lugar al siguiente resultado, de gran utilidad para la estadística, tanto a nivel teórico como práctico: Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con media μ y desviación típica σ, entonces verifica las hipótesis del teorema central del límite con $X \in N(0,1)$, $A_n = n\mu$ y $B_n = \sqrt{n}\sigma$. Es decir:

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}} \xrightarrow{n\to\infty} N(0,1).$$

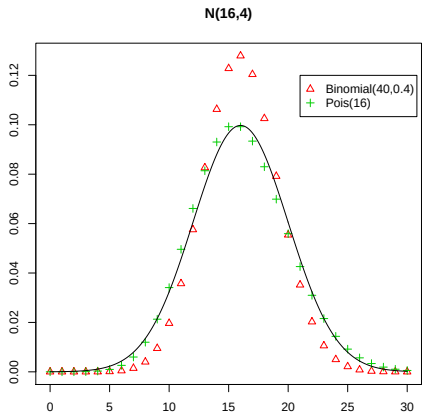
O lo que es lo mismo, la distribución de $\sum_{k=1}^{n} X_k$ se puede aproximar por una $N(n\mu, \sigma\sqrt{n})$ cuando n es suficientemente grande. De hecho, si dividimos numerador y denominador por n el anterior límite se puede escribir

$$\frac{\left(\bar{X} - \mu\right)}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} N(0,1),$$

siendo $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ la media de las variables. En esta nueva expresión se ve de una forma más transparente el papel de μ y σ y la analogía con la tipificación de una variable aleatoria. Este resultado nos dice que la suma de variables aleatorias iguales (misma distribución) e independientes acaba ajustándose a una distribución normal, sin importar la naturaleza de las variables que estemos sumando, de hecho. Este resultado es el que justifica las aproximaciones de la binomial y la Poisson por la normal enunciadas cuando se introdujeron dichas distribuciones.

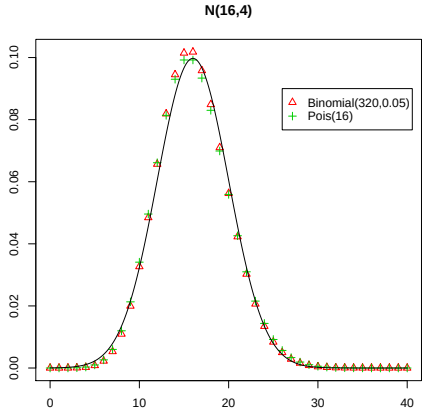
En este primer **ejemplo** vemos que el ajuste de la binomial no es bueno, esto es porque $p = 0{,}4$, y es necesario tener $p < 0{,}1$ y $n > 30$ para tener buenos ajustes.

```
x<-seq(0,30,by=1)
plot(x,dbinom(x,40,0.4),pch=2,col=2,ylab="",xlab="",main="N(16,4)")
points(x,dpois(x,16),pch=3,col=3)
x<-seq(0,30,by=0.01)
lines(x,dnorm(x,16,4))
leg.txt=c("Binomial(40,0.4)","Pois(16)")
legend(22,0.12,leg=leg.txt,pch=c(2,3),col=c(2,3))
```



En el segundo gráfico tenemos un ajuste mucho mejor ya que la probabilidad es más pequeña y n más grande.

```
x<-seq(0,40,by=1)
plot(x,dbinom(x,320,0.05),ylab="",xlab="",pch=2,col=2,main="N(16,4)")
points(x,dpois(x,16),pch=3,col=3)
x<-seq(0,40,by=0.01)
lines(x,dnorm(x,16,4))
leg.txt=c("Binomial(320,0.05)","Pois(16)")
legend(28,0.09,leg=leg.txt,pch=c(2,3),col=c(2,3))
```



2.10. Anexo: repaso de combinatoria

Para aplicar la regla de Laplace, el cálculo de los sucesos favorables y de los sucesos posibles a veces no plantea ningún problema, ya que son un número reducido y se pueden calcular con facilidad. Sin embargo a veces es necesario echar mano de alguna herramienta matemática para realizar dichos cálculos, aquí es donde entra la combinatoria.
EJEMPLO: 5 matrimonios se sientan aleatoriamente a cenar y queremos calcular la probabilidad de que al menos los miembros de un matrimonio se sienten juntos. En este caso, determinar el número de casos favorables y de casos posibles es complejo.

Las reglas matemáticas que nos pueden ayudar son las combinaciones, las variaciones y las permutaciones.

2.10.1. Combinaciones

Dado un conjunto de "n" elementos, las combinaciones nos permiten calcular el número de subgrupos de "m" elementos ($m < n$) que se pueden formar con ellos. Cada subgrupo se diferencia del resto en los elementos que lo componen, el orden no influye.

Para calcular el número de combinaciones se aplica la siguiente fórmula:

$$C_n^m = \left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{m!(n-m)!}.$$

El termino "$n!$" se denomina "factorial de n" y es la multiplicación de todos los números que van desde 1 hasta n. La expresión C_n^mse lee: combinaciones de "n" elementos tomados de "m" en "m".

EJEMPLO: *¿Cuántos grupos de 5 alumnos pueden formarse con los treinta alumnos de una clase? (Un grupo es distinto de otro si se diferencia de otro por lo menos en un alumno)*
No importa el orden (son grupos de alumnos). No puede haber dos alumnos iguales en un grupo evidentemente, luego sin repetición. Por tanto, se pueden formar 142506 grupos distintos:

$$C_{30}^5 = \left(\begin{array}{c} 30 \\ 5 \end{array} \right) = \frac{30!}{5!(30-5)!} = 142506.$$

También podemos analizar que ocurriría con el cálculo de las combinaciones en el supuesto de que al formar los subgrupos los elementos pudieran repetirse.

2.10.2. Combinaciones con repetición

Para calcular el número de combinaciones con repetición se aplica la siguiente fórmula:

$$CR_n^m = C_{n+m-1}^m = \frac{(n+m-1)!}{m!(n-1)!}.$$

EJEMPLO: *En una confitería hay cinco tipos diferentes de pasteles. ¿De cuántas formas se pueden elegir cuatro pasteles?*
No importa el orden (son pasteles). Puede haber dos o más pasteles de un mismo tipo en un grupo, luego con repetición. Por tanto, se pueden formar 70 grupos distintos:

$$CR_5^4 = \begin{pmatrix} 5+4-1 \\ 4 \end{pmatrix} = \frac{8!}{4!(5-1)!} = 70.$$

2.10.3. Variaciones

Dado un conjunto de "n" elementos, las variaciones nos permiten calcular el número de subgrupos de "m" elementos ($m < n$) que se pueden formar con ellos. Cada subgrupo se diferencia del resto en los elementos que lo componen o en el orden de los mismos, el orden influye.

Para calcular el número de variaciones se aplica la siguiente fórmula:

$$V_n^m = \frac{n!}{(n-m)!}.$$

EJEMPLO: *¿Cuántos números de tres cifras distintas se pueden formar con las nueve cifras significativas del sistema decimal?*
Al tratarse de números el orden importa y además nos dice "cifras distintas" luego no pueden repetirse. Por tanto, se pueden formar 504 números : $V_9^3 = \frac{9!}{(9-6)!} = 9 \cdot 8 \cdot 7 = 504$.

2.10.4. Variaciones con repetición

Nos indican el número de subgrupos distintos de "m" elementos que se pueden formar con los "n" dados. Se llaman variaciones con repetición porque un mismo elemento puede entrar más de una vez en cada subgrupo. La fórmula es

$$VR_n^m = n^m.$$

EJEMPLO: *¿Cuántos números de tres cifras se pueden formar con las nueve cifras significativas del sistema decimal?*
Al tratarse de números el orden importa y además no dice nada sobre "cifras distintas" luego sí pueden repetirse. Se pueden formar 729 números : $VR_9^3 = 9^3 = 729$.

2.10.5. Permutaciones

Dado un conjunto de "n" elementos, las permutaciones nos dan el número de posibles ordenaciones de los mismos.

Para calcular el número de permutaciones se aplica la siguiente fórmula:

$$P_n = V_n^n = n!.$$

EJEMPLO: *Con las letras de la palabra DISCO ¿cuantas palabras distintas (con o sin sentido) se pueden formar?*

Evidentemente, al tratarse de palabras el orden importa. Y además $n = m$, es decir tenemos que formar palabras de cinco letras con cinco elementos D, I, S, C, O que no están repetidos. Por tanto, se pueden formar 120 palabras : $P_5 = 5! = 120$.

2.10.6. Permutaciones con repetición

Dado un conjunto de "n" elementos, si queremos calcular las posibles ordenaciones de los mismos, permitiendo que se repitan r_1, r_2,...,r_n veces respectivamente. Sea $k = r_1 + r_2 + \ldots + r_n$, para calcular el número de permutaciones con repetición se aplica:

$$PR_k^{r_1,r_2,\ldots,r_n} = \frac{k!}{r_1!r_2!\ldots r_n!}.$$

EJEMPLO: *¿De cuántas maneras distintas pueden colocarse en línea nueve bolas de las que 4 son blancas, 3 amarillas y 2 azules?*
El orden importa por ser de distinto color, pero hay bolas del mismo color (están repetidas) y además $n = m$, *i.e.* colocamos 9 bolas en línea y tenemos 9 bolas para colocar. Por tanto, tenemos 1260 modos de colocarlas : $PR_9^{4,3,2} = \frac{9!}{4!3!2!} = 1260$.

TABLA RESUMEN:

Agrupaciones	Tipo	¿Importa orden?	¿Pueden repetirse?	Elem. grupo	Elem. total	FÓRMULA
Variaciones	Sin rep.	SI	NO			$V_n^m = \frac{n!}{(m-n)!}$
	Con rep.		SI			$VR_n^m = n^m$
Permutaciones	Sin rep.	SI	NO			$P_n = n!$
	Con rep.		SI	m	n	$P_n^{a,b,c,\ldots} = \frac{n!}{a!b!c!\ldots}$
Combinaciones	Sin rep.	NO	NO			$C_n^m = \binom{n}{m} = \frac{n!}{m!(n-m)!}$ $C_n^m = \frac{V_n^m}{P_n}$
	Con rep.		SI			$CR_n^m = C_{n+m-1}^m = \frac{(n+m-1)!}{m!(n-1)!}$

Una página web en la que estos conceptos de combinatoria figuran explicados más detalladamente es http://thales.cica.es/rd/Recursos/rd99/ed99-0516-02/practica/

2.11. Ejercicios resueltos

EJERCICIO: En un examen de tipo test, un estudiante contesta a una pregunta que ofrece tres soluciones posibles. La probabilidad de que el estudiante conozca la respuesta es 0,8.

a) Si el estudiante contesta correctamente la pregunta, ¿cuál es la probabilidad de que realmente sepa la respuesta correcta?

b) Si el examen consta de 100 preguntas, los aciertos valen un punto y los fallos restan 1/3 y el alumno responde a todas las preguntas, ¿cuál es la probabilidad de que obtenga al menos un notable, si para ello debe obtener al menos 70 puntos?

Solución:

a) Consideramos los sucesos

$$S = \text{“el alumno conoce la respuesta”},$$
$$C = \text{“el alumno contesta correctamente a la pregunta”}.$$

Sabiendo que el alumno ha contestado correctamente a la pregunta, la probabilidad de que el alumno realmente supiese la respuesta es $P(S/C)$, que por el teorema de Bayes se calcula como:

$$P(S/C) = \frac{P(C/S)P(S)}{P(C/S)P(S) + P(C/S^c)P(S^c)}.$$

Evidentemente $P(C/S) = 1$, pues es la probabilidad de que el alumno conteste correctamente a la pregunta teniendo en cuenta que la sabe. Por otra parte, el enunciado del ejercicio establece que $P(S) = 0{,}8$ y en consecuencia $P(S^c) = 1 - P(S) = 0{,}2$ ($P(S^c)$ representa la probabilidad de que el alumno no sepa la respuesta). Por último $P(C/S^c)$ representa la probabilidad de que el alumno conteste correctamente a la pregunta teniendo en cuenta que no sabe la respuesta. Como cada pregunta ofrece 3 posibles respuestas, el alumno elegirá al azar una de esas respuestas. La probabilidad de acertar la correcta es entonces 1/3 y, por lo tanto, $P(C/S^c) = 1/3 = 0{,}33$. En definitiva

$$P(S/C) = \frac{P(C/S)P(S)}{P(C/S)P(S) + P(C/S^c)P(S^c)} = \frac{1 \cdot 0{,}8}{1 \cdot 0{,}8 + 0{,}33 \cdot 0{,}2} = 0{,}923.$$

b) Sea

$$N = \text{“Nota obtenida por el alumno en el examen”}.$$

Así definida N es una variable aleatoria. Nos interesa determinar $P(N \geq 70)$, es decir, la probabilidad de que el alumno obtenga una nota mayor de 70 puntos y que es la condición necesaria para sacar por lo menos un notable. Pero, ¿cómo determinamos la nota de un alumno? Necesitamos saber cuántas preguntas ha contestado correctamente. Sea

$$X = \text{“Número de respuestas correctas contestadas”}.$$

Entonces, como se establece en el enunciado, la nota del alumno se calcula como

$$N = 1 \cdot X - \frac{1}{3}(100 - X).$$

Por tanto,

$$P\left(N \geq 70\right) = P\left(X - \frac{1}{3}(100 - X) \geq 70\right) = P\left(\frac{4}{3}X - \frac{100}{3} \geq 70\right) = P(X \geq 77{,}5).$$

Es fácil ver que X sigue una distribución binomial de parámetros $n = 100$ y $p = P(C) = P(\text{“el alumno contesta correctamente a la pregunta”})$. Fíjate que el examen es la repetición de $n = 100$ experimentos (preguntas) independientes de manera que para todos ellos la probabilidad de éxito p (probabilidad de acertar la pregunta) es la misma. ¿Y cuánto vale dicha probabilidad? El alumno puede acertar una pregunta bien porque la sabe o bien porque aunque la desconoce, la echa a suertes y acierta. Para calcular $P(C)$ usaremos el teorema de las probabilidades totales. Así

$$P(C) = P(C/S)P(S) + P(C/S^c)P(S^c) = 1 \cdot 0{,}8 + 0{,}33 \cdot 0{,}2 = 0{,}8666.$$

Por lo tanto $X \in B(100, 0{,}8666)$ y

$$P(X \geq 77{,}5) = 1 - P(X < 77{,}5) = 1 - 0{,}0058 = 0{,}9941.$$

El valor de $P(X < 77{,}5)$ lo hemos calculado en R con el comando

```
> pbinom(77.5,100,0.8666)
[1] 0.005816317
```

También se puede calcular el valor de $P(X < 77{,}5)$ aproximando por una distribución normal $N(np, \sqrt{np(1-p)})$, por ser $n > 30$ y $0{,}1 < p < 0{,}9$. Y tipificando la variable obtenemos

$$\begin{aligned} P(X < 77{,}5) &= P\left(\frac{X - 86{,}66}{\sqrt{100 \cdot 0{,}8666 \cdot 0{,}1334}} < \frac{77{,}5 - 86{,}66}{\sqrt{100 \cdot 0{,}8666 \cdot 0{,}1334}}\right) \\ &= P(Z < -2, 69), \end{aligned}$$

donde Z se aproxima a una N(0,1). Mirando en las tablas de la distribución normal estándar obtenemos que $P(Z < -2, 69) = 0{,}0035$ y, por lo tanto, $P(X \geq 77{,}5) = 1 - 0{,}0035 = 0{,}996$. Fíjate que el resultado obtenido al aproximar por la normal no difiere mucho del que se obtenía utilizando la distribución binomial.

En definitiva obtenemos que la probabilidad de que el alumno saque al menos un notable es de 0.996.

Ejercicio: Consideremos una variable aleatoria bidimensional (X, Y) con función de densidad conjunta

$$f(x,y) = \begin{cases} kx(1-y^2) & \text{, si } 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0 & \text{, en otro caso.} \end{cases}$$

a) Determinar el valor de k para que f sea función de densidad.

b) Calcular las funciones de densidad marginal de X e Y. ¿Son ambas variables independientes?

c) Calcular la función de densidad de la variable $Z = X/Y$.

SOLUCIÓN:

a) Dadas dos variables aleatorias continuas X e Y, definidas sobre el mismo espacio de probabilidad, la función de densidad conjunta de la variable aleatoria bidimensional (X, Y) es una función f verificando,

i) $f(x, y) \geq 0$,
ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$.

Por lo tanto, para que f tal y como está definida en el enunciado del ejercicio sea función de densidad, debemos garantizar que integra 1 en el dominio de definición. Entonces, teniendo en cuenta que el conjunto de valores donde f es no nula coincide con el cuadrado unidad

$$\{(x, y) \in \mathbb{R}^2 : \ 0 \leq x \leq 1, \ 0 \leq y \leq 1\},$$

se debe verificar

$$1 = \int_0^1 \int_0^1 kx(1 - y^2)dxdy.$$

Resolviendo la integral,

$$1 = \int_0^1 \int_0^1 kx(1 - y^2)dxdy = \int_0^1 k(1 - y^2) \left[\frac{x^2}{2}\right]_{x=0}^{x=1} dy =$$
$$= \int_0^1 \frac{k(1 - y^2)}{2} dy = \frac{k}{2} \left[y - \frac{y^3}{3}\right]_{y=0}^{y=1} = \frac{k}{2} \left(1 - \frac{1}{3}\right) = \frac{k}{3}.$$

Por lo tanto, despejando el valor de k obtenemos que f es función de densidad para $\boldsymbol{k=3}$. Se muestra en la Figura 2.1 la representación gráfica de f para dicho valor de k.

b) Dada una variable aleatoria bidimensional (X, Y) con densidad conjunta f, se defina la función de densidad marginal de X como

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy.$$

En nuestro caso, fijado $x \in [0, 1]$, la variable y se mueve entre 0 y 1. Por lo tanto, para $0 \leq x \leq 1$,

$$f_X(x) = \int_0^1 3x(1 - y^2)dy = 3x \left[y - \frac{y^3}{3}\right]_{y=0}^{y=1} = 3x \left(1 - \frac{1}{3}\right) = 2x.$$

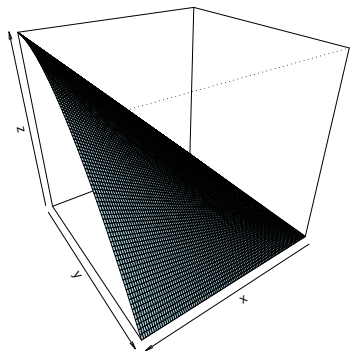


Figura 2.1: Representación de la función f para $k = 3$.

Para $x < 0$ o $x > 1$, la función f es nula y también lo será entonces $f_X(x)$. En definitiva

$$f_X(x) = \begin{cases} 2x & \text{, si } 0 \leq x \leq 1, \\ 0 & \text{, en otro caso.} \end{cases}$$

De forma análoga, se define la función de densidad marginal de Y como

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

Ahora, fijado $y \in [0, 1]$, la variable x se mueve entre 0 y 1. Por lo tanto, para $0 \leq y \leq 1$,

$$f_Y(y) = \int_0^1 3x(1 - y^2)dx = 3(1 - y^2)\left[\frac{x^2}{2}\right]_{x=0}^{x=1} = \frac{3}{2}\left(1 - y^2\right).$$

Para $y < 0$ o $y > 1$, la función f es nula y también lo será entonces $f_Y(y)$. En definitiva

$$f_Y(y) = \begin{cases} \frac{3}{2}\left(1 - y^2\right) & \text{, si } 0 \leq y \leq 1, \\ 0 & \text{, en otro caso.} \end{cases}$$

Por último, X e Y son independientes si la densidad conjunta es igual al producto de las densidades marginales, es decir, si

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

En nuestro caso

$$f_X(x)f_Y(y) = \begin{cases} 2x \cdot \frac{3}{2}\left(1 - y^2\right) = 3x\left(1 - y^2\right) & \text{, si } 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0 & \text{, en otro caso,} \end{cases}$$

que coincide con $f(x, y)$. Por lo tanto podemos concluir que las variables X e Y son independientes.

c) La variable $Z = X/Y$ se obtiene como transformación de la variable (X, Y). Así, podemos escribir

$$(Z, T) = g(X, Y) = (g_1(X, Y), g_2(X, Y)),$$

siendo

$$z = g_1(x, y) = x/y,$$
$$t = g_2(x, y) = y.$$

Tanto g_1 como g_2 admiten inversas u_1 y u_2, siendo

$$x = u_1(z, t) = zt,$$
$$y = u_2(z, t) = t.$$

Podemos aplicar el teorema de cambio de variable, siendo el determinante jacobiano

$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} t & z \\ 0 & 1 \end{vmatrix} = t.$$

Teniendo en cuenta que $0 \leq y \leq 1$ y $0 \leq x \leq 1$ determinamos el campo de variación de (Z, T). Así,

$$t = y \Rightarrow 0 \leq t \leq 1.$$
$$x = zt \Rightarrow z = x/t \Rightarrow 0 \leq z \leq 1/t.$$

En definitiva, la densidad conjunta w de la variable bidimensional (Z, T) será

$$w(z, t) = f(u_1(z, t), u_2(z, t))\,|J| = 3zt(1 - t^2)t = 3zt^2(1 - t^2),$$

para (z, u) verificando $0 \leq t \leq 1$ y $0 \leq z \leq 1/t$. (Ver Figura 2.2).

Finalmente, integrando con respecto a t se obtiene la función de densidad marginal de la variable $Z = X/T$. Debemos tener cuidado al definir los extremos de integración. A la vista de la Figura 2.2, la variable Z toma valores en $[0, \infty)$. Sin embargo debemos tener en cuenta que la variable T tiene distinto campo de variación dependiendo del valor de Z. Así, para $0 \leq z \leq 1$ se tiene que $0 \leq t \leq 1$. Por otra parte, para $z > 1$ se tiene que $0 \leq t \leq 1/z$. Por lo tanto

$$w_Z(z) = \int_{-\infty}^{\infty} w(z, t)dt = \begin{cases} \int_0^1 3zt^2(1 - t^2)dt = \frac{2z}{5} & \text{, si } 0 \leq z \leq 1, \\ \int_0^{1/t} 3zt^2(1 - t^2)dt = \frac{5z^2-3}{5z^4} & \text{, si } z > 1. \end{cases}$$

EJERCICIO: Supongamos que la función de distribución conjunta de dos variables x e y es la siguiente:

$$f\,(x, y) = c\left(x^2 + y\right), \qquad 0 \leq y \leq 1 - x^2.$$

Determínese:

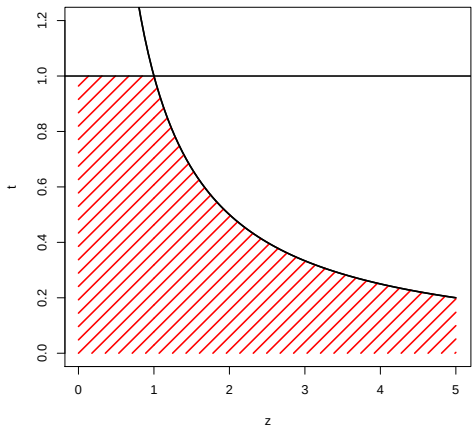


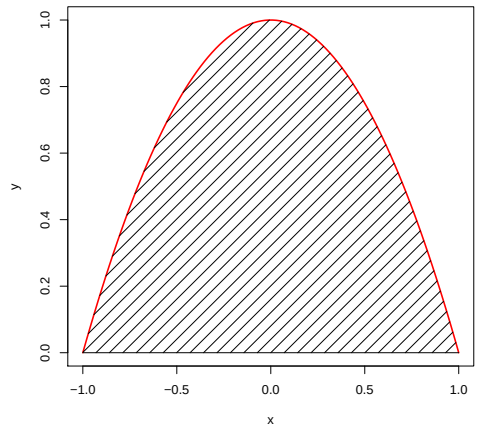
Figura 2.2: Campo de variación de (Z, T).

a) El valor de la constante c y $P\,(0 \leq x \leq 1/2)$.

b) $P\,(y \leq 1 - x)$ y $P\left(y \geq x^2\right)$.

Solución:

a) Para este apartado, lo primero que hay que tener claro es el área donde la función de densidad toma valores no nulos. Esta área se puede ver en la figura adjunta y se corresponde con el área bajo la curva $y = 1 - x^2$ (roja) y los dos ejes.



Para resolver el primer apartado debemos usar que la integral de la densidad debe

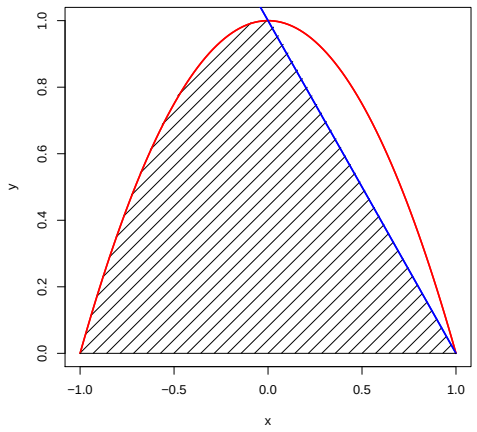
ser 1 y por tanto,

$$c\int_{-1}^{1}\int_{0}^{1-x^2}\left(x^2+y\right)dydx = c\int_{-1}^{1}\left[x^2y+\frac{y^2}{2}\right]_0^{1-x^2}dx =$$
$$= c\int_{-1}^{1} x^2\left(1-x\right)^2+\frac{\left(1-x^2\right)^2}{2}dx =$$
$$= c\int_{-1}^{1}\frac{1-x^4}{2}dx = c\left[\frac{x-\frac{x^5}{5}}{2}\right]_{-1}^{1} = c\frac{4}{5} = 1$$

y despejando tenemos que $c = 5/4$. La segunda parte del apartado a) basta con sustituir y resolver,

$$P\left(0 \leq x \leq 1/2\right) = \frac{5}{4}\int_{0}^{1/2}\frac{1-x^4}{2}dx = \frac{5}{4}\left[\frac{x-\frac{x^5}{5}}{2}\right]_0^{1/2} = \frac{5}{4}\frac{\frac{1}{2}-\frac{\left(\frac{1}{2}\right)^5}{5}}{2} = 0{,}3085.$$

b) Para este apartado simplemente tenemos que ser cuidadosos a la hora de establecer los límites de integración. Así la primera parte es integrar en el área bajo la línea azul y los dos ejes, como se muestra en la siguiente figura:



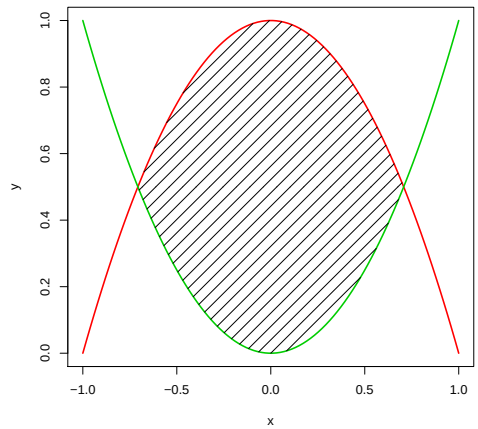
$$P\left(y \leq 1-x\right) = \frac{1}{2} + \frac{5}{4}\int_0^1 \int_0^{1-x} \left(x^2+y\right) dxdy = \frac{1}{2} + \frac{5}{4}\int_0^1 \left[x^2 y + \frac{y^2}{2}\right]_0^{1-x} dx =$$

$$= \frac{1}{2} + \frac{5}{4}\int_0^1 \left(x^2\left(1-x\right) + \frac{\left(1-x\right)^2}{2}\right) dx = \frac{1}{2} + \frac{5}{4}\int_0^1 \left(x^2 - x^3 + \frac{1-2x+x^2}{2}\right) dx =$$

$$= \frac{1}{2} + \frac{5}{4}\int_0^1 \left(\frac{1}{2} - x + \frac{3}{2}x^2 - x^3\right) dx = \frac{1}{2} + \frac{5}{4}\left[\frac{x}{2} - \frac{x^2}{2} + \frac{x^3}{2} - \frac{x^4}{4}\right]_0^1 =$$

$$= \frac{1}{2} + \frac{5}{4}\left[\frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{4}\right] = \frac{1}{2} + \frac{5}{4}\cdot\frac{1}{4} = \frac{13}{16}.$$

El área de integración para la segunda parte de este apartado se muestra en la siguiente figura. Habrá que integrar entre la curva $y = x^2$ (verde) y la curva $y = 1 - x^2$ (roja) teniendo en cuenta que la variable x sólo toma valores entre $-\frac{1}{\sqrt{2}}$ y $\frac{1}{\sqrt{2}}$ (donde se cruzan la curva verde y la roja).



Por tanto, la probabilidad pedida es:

$$P\left(y \geq x^2\right) = \frac{5}{4} \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \int_{x^2}^{1-x^2} \left(x^2 + y\right) dydx = \frac{5}{4} \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \left[x^2 y + \frac{y^2}{2}\right]_{x^2}^{1-x^2} dx =$$

$$= \frac{5}{4} \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \left[x^2\left(1-x^2\right) + \frac{\left(1-x^2\right)^2}{2} - x^4 - \frac{x^4}{2}\right] dx =$$

$$= \frac{5}{4} \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \left[x^2 - x^4 + \frac{1-2x^2+x^4}{2} - x^4 - \frac{x^4}{2}\right] dx =$$

$$= \frac{5}{4} \int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} \left(\frac{1}{2} - 2x^4\right) dx = \frac{5}{4}\left[\frac{1}{2}x - \frac{2}{5}x^5\right]_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} =$$

$$= \frac{5}{4}\left[\frac{1}{\sqrt{2}} - \frac{1}{5\sqrt{2}}\right] = \frac{5}{4}\left[\frac{5-1}{5\sqrt{2}}\right] = \frac{1}{\sqrt{2}}.$$

Capítulo 3

Inferencia paramétrica

3.1. Introducción a la Inferencia Estadística

Una vez asentadas las bases de teoría de probabilidad podemos intentar inferir de la población, es decir, extraer información sobre las distintas características de interés de una cierta población de la que se ha observado un conjunto de datos. Así, puede ser de interés estimar los parámetros de la distribución de probabilidad asociada a la población, construir intervalos de confianza, predecir valores futuros o verificar si ciertas hipótesis son coherentes con los datos observados. Por tanto, la inferencia comprende alguna de las fases del método estadístico. Estas fases son recogida y depuración de datos, estimación, contrastes de simplificación, diagnosis y validación del modelo. Respecto al objetivo de estudio dividiremos la inferencia en **paramétrica**, cuando el objeto de interés sean los parámetros de la distribución de probabilidad que se supone conocida, y en **no paramétrica**, cuando nuestro interés se centra en características más generales de la distribución de probabilidad no referenciados en parámetros.

Otra clasificación de la inferencia se tiene atendiendo al tipo de información considerada que nos clasificaría la inferencia en clásica y bayesiana. El enfoque clásico o frecuentista trata los parámetros poblacionales desconocidos como valores fijos que deben ser estimados. El enfoque bayesiano considera que los parámetros desconocidos son variables aleatorias para las cuales debe fijarse una distribución inicial (a priori). Mezclando la distribución a priori con la información muestral los métodos bayesianos hacen uso de la regla de Bayes para ofrecer una distribución a posteriori de los parámetros.

3.2. Conceptos

Denominaremos **población** a un conjunto homogéneo de individuos sobre los que se estudian una o varias características. Es frecuente que no se pueda observar toda la población por un sinfín de motivos (empezando por el económico) de manera que normalmente trabajaremos con un subconjunto de la población que se denominará la **muestra**. Llamaremos **tamaño muestral** al número de elementos de la muestra. Un **método de muestreo** será el procedimiento empleado para la obtención de la muestra.

El objetivo de los métodos de muestreo es que la muestra represente a la población de la mejor manera posible.

El **muestreo aleatorio simple** es aquel en el que cada vez que seleccionamos un individuo de la muestra, este tiene la misma probabilidad de ser elegido para formar parte de la muestra que cualquier otro independientemente de que otros individuos hayan sido ya seleccionados. Por tanto, bajo muestreo aleatorio simple un individuo podría aparecer en la muestra más de una vez (con reemplazamiento). En este caso las variables aleatorias que conforman la muestra pueden suponerse independientes e idénticamente distribuidas (según la distribución poblacional).

El **muestreo sistemático** consiste en utilizar una lista u orden que presenten los datos. Dada una población de N individuos de la que se quiere extraer n elementos, el muestreo sistemático consiste en elegir aleatoriamente un valor l en el conjunto $\{1,...,k\}$, donde k denota la parte entera de N/n. A partir de ese valor de l consideramos todos aquellos situados en la lista en las posiciones $(i-1)k + l$. Desde el punto de vista computacional este muestreo es más sencillo e incluso puede ser obtener una muestra más representativa que el muestreo aleatorio simple sin embargo un serio inconveniente puede darse si en la muestra existen periodicidades que coincidan con la longitud del salto en la lista.

Otro tipo de muestreo es el **muestreo estratificado** que ocurre cuando es posible dividir la muestra en grupos o estratos entre los que existen diferencias importantes. En este caso se trata de obtener un cierto número de individuos (**afijación**) de cada estrato según muestreo aleatorio simple. Si el número que selecciono en cada estrato es el mismo se dirá que utilizamos una **afijación simple**. Si el número se elige proporcional al tamaño del estrato se dirá que usamos una **afijación proporcional**. La **afijación óptima** consiste en elegir la muestra según la formula siguiente:

$$n_i = \frac{N_i\sigma_i \big/ \sqrt{c_i}}{\sum_{j=1}^{k} N_j\sigma_j \big/ \sqrt{c_j}},$$

donde N_i es la tamaño del estrato, c_i es el coste de muestrear una unidad en el estrato i-ésimo y σ_i es la desviación típica de la característica en el estrato i-ésimo. Esta manera de seleccionar la muestra minimiza la varianza de la característica analizada en la muestra.

En el **muestreo por conglomerados** la población se divide en conglomerados que se suponen homogéneos entre ellos de manera que se puede seleccionar de algunos conglomerados para obtener una muestra representativa (en vez de seleccionar de todos los conglomerados).

Otros tipos de muestreo serían el **muestreo polietápico** (que mezcla varios de los anteriores), el **muestreo opinático** (cada elemento se elige subjetivamente), el **muestreo por cuotas** (opinático por estratos), el **muestreo semialeatorio** y el **muestreo por rutas**. Estos tipos de muestreo sacrifican propiedades deseables de la muestra en aras de facilidad de obtención o menor coste y en general no podrán medir con el mismo rigor que tipos de muestreo anteriores pudiendo producir sesgos en la muestra. Otro problema es que a menudo el causante de los problemas en la representa-

tividad de la muestra viene causado por el diseño del cuestionario o la propia esencia de obtención del dato. Así, en un cuestionario demoscópico debemos tener en cuenta que las preguntas sean entendidas por todos los entrevistados sin lugar a la interpretación personal o que factores culturales induzcan diferencias, que las preguntas no introduzcan en su formulación o en su orden elementos de sesgo que influyan sobre la respuesta o que no provoquen efectos anímicos o de conciencia que afecten a la calidad de las respuestas, su sinceridad o su pertinencia.

3.3. Distribución muestral y función de verosimilitud

A partir de este punto trataremos más profundamente la inferencia paramétrica donde podemos pensar que estamos interesados en el estudio de una variable aleatoria X, cuya distribución, F, es en mayor o menor grado desconocida. Conociendo la distribución podemos extraer conclusiones acerca de la población en estudio. En la inferencia paramétrica suponemos que tenemos una familia de distribuciones cuya distribución de probabilidad se supone conocida salvo los valores que toman ciertos coeficientes (parámetros), es decir, $\mathcal{F} = \left\{F_\theta | \theta \in \Theta \subset \mathbb{R}^k\right\}$ (a Θ se le llama espacio paramétrico). Para tratar de conocer F tomamos una muestra $\{x_1,...,x_n\}$. El método de muestreo empleado influirá decisivamente en los pasos posteriores ya que la distribución conjunta de la muestra $F(x_1,...,x_n)$ será necesaria para el proceso de inferencia. Llamaremos **muestra aleatoria simple** de tamaño n de una variable aleatoria X con distribución teórica F a n variables aleatorias $x_1,...,x_n$ independientes e igualmente distribuidas con distribución común F. Consecuentemente, la función de distribución conjunta de la muestra es $F(x_1,...,x_n) = F(x_1) \times F(x_2) \times ... \times F(x_n)$. Llamaremos **espacio muestral** al conjunto de muestras posibles que pueden obtenerse al seleccionar una muestra de un tamaño determinado de una cierta población. Llamaremos **estadístico** a cualquier función T de la muestra. El estadístico $T(x_1,...,x_n)$ como función de variables aleatorias, es también una variable aleatoria y tendrá por tanto una distribución de probabilidad que llamaremos **distribución en el muestreo** de T.

Ejemplos de estadísticos serían:

- La media muestral: $T(x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- La varianza muestral: $T(x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$.
- La cuasivarianza muestral: $T(x_1, \cdots, x_n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

En general, usaremos la notación $\hat{\theta}_n = T(x_1, \cdots, x_n)$ para referirnos a un estadístico. A continuación presentamos algunas propiedades deseables en un estadístico:

- **Centrado o insesgado**: Si se cumple que $E(\hat{\theta}_n) = \theta$. Llamaremos $\text{Sesgo}_{\hat{\theta}}(\theta) = E(\hat{\theta}_n) - \theta$.

- **Asintóticamente insesgado**: Si $\lim_{n\to\infty} E\left(\hat{\theta}_n\right) = \theta$.
- **Consistente en media cuadrática**: $\lim_{n\to\infty} \text{ECM}\left(\hat{\theta}_n\right) = 0$ siendo $\text{ECM}(\hat{\theta}_n) = \text{Sesgo}_{\hat{\theta}}(\theta)^2 + \text{Var}(\hat{\theta}_n)$.
- **Eficiencia**: $\text{Efic}\left(\hat{\theta}_n\right) = 1 \Big/ \text{ECM}\left(\hat{\theta}_n\right)$.
- **Suficiencia**: Cuando el estadístico utiliza toda la información de la muestra.

Llamaremos **función de distribución empírica** a la función de distribución que resulta de suponer a la muestra como toda la población. Es por tanto, una función de distribución discreta y vendrá definida como

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)}, \\ \frac{r}{n} & \text{si } x_{(r)} \leq x < x_{(r+1)}, \\ 1 & \text{si } x \geq x_{(n)}. \end{cases}$$

donde la notación $x_{(i)}$ significa el dato de la muestra que ordenado ocupa la posición i-ésima.

Las medidas muestrales habituales se pueden escribir ahora en función de la distribución empírica. Así, la media muestral se escribirá como

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \int x dF_n(x) = E_{F_n}(x),$$

que resulta ser la esperanza bajo la distribución empírica.

Una propiedad relevante de la función de distribución empírica viene dada por el teorema de Glivenko-Cantelli que nos asegura que si n tiende a infinito entonces

$$\sup_x |F_n(x) - F(x)| \overset{c.s.}{\to} 0.$$

Esta propiedad nos dice que la función de distribución de la muestra converge uniformemente a la función de distribución de la población y por tanto cualquier funcional de la función de distribución empírica (cumpliendo algunas propiedades) convergerá al funcional de la distribución de la población. Esto asegura que, obteniendo muestras más grandes, podemos aproximarnos al parámetro de interés de la distribución tanto como queramos.

Para construir un estimador podemos emplear varios métodos. El más clásico es el **método de los momentos** que consiste en que si suponemos que la distribución de una variable X viene caracterizada por k parámetros (θ_1,...,θ_k), estos también influirán en los momentos poblacionales ($E(X^j) = g_j(\theta_1, ..., \theta_k)$). De manera que si estamos interesados en un determinado estimador una solución puede venir de considerar el sistema de ecuaciones de igualar los momentos poblacionales con los momentos muestrales. Si el sistema tiene solución ese estimador será el estimador por el método de los momentos.

Supongamos ahora una variable aleatoria continua X con función de densidad $f(\cdot|\Theta)$ donde se especifica Θ para indicar que depende de ciertos parámetros desconocidos. Si tenemos una muestra aleatoria simple $\{x_1,...,x_n\}$ llamaremos **función de verosimilitud** a la función de densidad conjunta de la muestra, es decir, $\ell(\Theta, \{x_1, ..., x_n\}) = \prod f(x_i|\Theta)$. Esta función, dada una muestra particular, nos proporcionaría la probabilidad para cada valor Θ de obtener la muestra dada. El objetivo en la inferencia es obtener información sobre la población a partir de una muestra. En este planteamiento aquel valor Θ que maximice esta función, será el mejor estimador de los parámetros desconocidos de la población. El estimador así construido se llamará **estimador máximo-verosímil**. Entre las propiedades que tienen los estimadores máximo-verosímiles en distribuciones cuyos rangos de valores no depende de parámetros están:

- Asintóticamente centrados.
- Asintóticamente normales.
- Asintóticamente de varianza mínima (eficientes).
- Si existe un estadístico suficiente, el estimador máximo-verosímil es suficiente.

3.4. Distribuciones en el muestreo de poblaciones normales

Supongamos en lo que sigue, y salvo indicación en contra, que estamos muestreando de una población normal de la que obtenemos una muestra $\{x_1,...,x_n\}$. Calculemos los estadísticos más habituales y sus propiedades.

3.4.1. Estimación de la media de una población

Supongamos que la muestra proviene de una variable aleatoria normal con media μ y varianza σ^2. Un estimador razonable del parámetro μ es la media muestral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ que verifica las siguientes propiedades:

- Insesgado: $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$.
- Consistente: $\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$.

Por tanto, la media muestral es una suma de variables normales con la misma media y varianza y por tanto será normal con media μ y varianza σ^2/n.

En el caso de que la población de partida no sea normal, por el teorema central de límite si el tamaño de la muestra es grande (≥ 30) la distribución de la media muestral se aproximará a la normal.

3.4.2. Estimación de la varianza de una población

En el mismo supuesto del apartado anterior un estimador razonable de la varianza de la población puede ser la varianza muestral: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ que presenta las siguientes propiedades:

- Es asintóticamente insesgado:

$$E(nS^2) = E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - n(\mu - \bar{x})^2\right) = (n-1)\sigma^2.$$

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

- Es consistente: Bajo hipótesis de normalidad se puede demostrar que $\frac{nS^2}{\sigma^2} \in \chi^2_{n-1}$ y por tanto $\text{Var}\left(\frac{nS^2}{\sigma^2}\right) = 2(n-1) \Rightarrow \text{Var}(S^2) = \frac{2(n-1)}{n^2}\sigma^4$.

Al efecto de corregir el sesgo de la varianza muestral se utiliza la **cuasivarianza** o **varianza muestral corregida** como $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ que presenta mejores propiedades:

- Es un estimador insesgado de σ^2: $E(\hat{S}^2) = \sigma^2$.
- Es un estimador consistente de σ^2: $\text{Var}(\hat{S}^2) = \frac{2}{n-1}\sigma^4$.
- $\frac{(n-1)\hat{S}^2}{\sigma^2} \in \chi^2_{n-1}$.
- Además la media muestral y la cuasivarianza muestral son variables aleatorias independientes.

Si la población de partida de la muestra no es normal entonces se mantienen las propiedades para la media de los estimadores pero no para la varianza que presentaría una fórmula que depende de los coeficientes de asimetría y curtosis. Tampoco serían válidas las propiedades que ligan la distribución de la varianza muestral (o cuasivarianza) con la distribución χ^2.

3.4.3. Estimación de una proporción

Supongamos una población en la que una proporción de individuos p tiene una determinada característica y supongamos que deseamos estimar esa proporción. Para ello obtenemos una muestra $\{x_1,...,x_n\}$ donde cada x_i es un valor cero si no posee la característica y uno si la posee. Por tanto tenemos una muestra aleatoria simple de una Bernouilli de parámetro p. Un estimador razonable de p es la proporción de elementos de la muestra que presentan la característica, es decir, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Este caso es un caso particular de la estimación de la media de una población con $E(x_i) = p$ y $\text{Var}(x_i) = p(1-p)$ y por tanto las propiedades de la media se derivan del apartado anterior.

3.5. Intervalos de confianza

En muchos casos una estimación puntual no es suficiente ya que no nos proporciona el error que se comete en la estimación. Para tener más información veremos a continuación como calcular un intervalo numérico que con cierta "seguridad" contenga al valor del parámetro en la población.

Definición: Se llama **intervalo de confianza** (IC) para el parámetro θ con **nivel** o **coeficiente de confianza** $1-\alpha$, con $0<\alpha<1$, a un intervalo aleatorio (a,b) tal que $P(a<\theta<b)=1-\alpha$. La aleatoriedad del intervalo (a,b) proviene de la muestra, es decir, a y b son funciones de la muestra.

En general, para construir un intervalo de confianza para un parámetro θ utilizaremos el método de la cantidad pivotal que consiste en seleccionar una muestra y considerar un estadístico $T(x_1,\ldots,x_n;\theta)$, cuya distribución en el muestreo no dependa de θ. Dicho estadístico se denomina **estadístico pivote**. Fijado un nivel de confianza $1-\alpha$ se determinan las constantes a y b tal que cumplen $P(a<T(x_1,\ldots,x_n;\theta)<b)=1-\alpha$. Si es posible despejar θ de la expresión anterior obtendremos dos valores que determinan el intervalo, es decir,

$$P\left(T^{-1}(x_1,\ldots,x_n;a)<\theta<T^{-1}(x_1,\ldots,x_n;b)\right)=1-\alpha.$$

3.5.1. Intervalo de confianza para la media de una población normal

Apliquemos los anteriores conceptos a la estimación por intervalo de la media de una población normal. Sea una muestra aleatoria simple $\{x_1,...,x_n\}$ de una distribución teórica N(μ,σ). Hemos visto que $\bar{x}\in \mathrm{N}(\mu,\sigma/\sqrt{n})$. Como estadístico pivote podemos considerar,

$$T(x_1,\ldots,x_n;\mu)=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}},$$

cuya distribución en el muestreo es $\mathrm{N}(0,1)$ y por tanto no depende de μ. Así, fijado el nivel de confianza y llamando $z_{\alpha/2}$ al cuantil $\alpha/2$ de la normal, se tendrá que

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

donde hemos despejado μ en función de la varianza teórica σ y del tamaño muestral. Por supuesto, este intervalo sólo se puede calcular cuando se conoce la varianza teórica.

Si no se conoce la varianza teórica de la población entonces el estadístico pivote que debemos utilizar sería

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\frac{(n-1)\hat{S}^2}{\sigma^2(n-1)}}=\frac{\bar{x}-\mu}{\hat{S}/\sqrt{n}}\in t_{n-1},$$

que no depende ahora de σ. Por tanto con los mismos pasos anteriores el intervalo de confianza será:

$$\left(\bar{x}-t_{n-1,\alpha/2}\frac{\hat{S}}{\sqrt{n}}\leq\mu\leq\bar{x}+t_{n-1,\alpha/2}\frac{\hat{S}}{\sqrt{n}}\right).$$

Una cuestión interesante es determinar el tamaño muestral para obtener un intervalo de una determinada longitud. Esta cuestión sólo se puede resolver propiamente suponiendo la varianza conocida. En este caso la longitud del intervalo es $L = 2z_{\alpha/2}\sigma/\sqrt{n}$ y por tanto despejando n se tiene

$$n = \frac{4z_{\alpha/2}^2\sigma^2}{L^2}.$$

Si la varianza no es conocida en la fórmula anterior hay que sustituir la varianza teórica por la cuasivarianza y el valor crítico de la normal por el de una t de Student. El valor crítico de la t depende de n aunque para n grande se puede aproximar por el de la normal. La cuasivarianza tiene el problema de que se conoce una vez obtenida la muestra cuando la cuestión planteada es previa a la obtención de la muestra. Se puede solventar esta dificultad bien obteniendo una muestra preliminar o bien obteniendo información previa de estudios anteriores.

3.5.2. Intervalo de confianza para la varianza de una población normal

Para calcular el intervalo de confianza para la varianza de poblaciones normales podemos considerar el estadístico pivote

$$\frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{nS^2}{\sigma^2} \in \chi^2_{n-1},$$

ya visto en apartados anteriores y que sólo depende de σ. Procediendo como en el caso de la media podemos obtener el intervalo

$$1-\alpha = P\left(\chi^2_{n-1,\alpha/2} \leq \frac{nS^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\right) = P\left(\frac{nS^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

donde $\left(\chi^2_{n-1,\alpha/2}, \chi^2_{n-1,1-\alpha/2}\right)$ son los cuantiles de la χ^2.

3.5.3. Intervalo de confianza para la diferencia de medias de poblaciones normales

El objetivo de este apartado es construir intervalos de confianza no sólo para la media de la población sino también para la comparación de dos poblaciones mediante su media. Como vimos, la media es una medida de resumen de la población de tal manera que si obtenemos diferencias significativas entre las medias de dos poblaciones podremos inferir que las poblaciones son diferentes. En este apartado tendremos dos muestras $\{x_1,...,x_n\}$ e $\{y_1,...,y_m\}$ procedentes de dos poblaciones que pueden ser independientes o apareadas. En el caso de muestras independientes los individuos en los que se ha medido la variable X son diferentes de aquellos en los que se mide la variable Y entendiendo que la obtención de la primera muestra no afecta a la segunda. En el caso de muestras apareadas se ha medido para un mismo grupo de individuos la variable X y la variable Y que se pretenden comparar. En este caso $m = n$ y se supone una cierta dependencia entre cada valor de la primera muestra y su homólogo de la segunda.

3.5.4. Muestras independientes, varianzas poblacionales conocidas

Tenemos una muestra $\{x_1,...,x_n\}$ de una variable $X \in \mathrm{N}(\mu_X, \sigma_X)$ y otra muestra $\{y_1,...,y_m\}$ de la variable $Y \in \mathrm{N}(\mu_Y, \sigma_Y)$. En este caso las medias muestrales $(\bar{x},\ \bar{y})$ son independientes y normales y por tanto su diferencia también tendrá una distribución normal.

$$\left.\begin{array}{l} \bar{x} \in \mathrm{N}(\mu_X, \sigma_X/\sqrt{n}) \\ \bar{y} \in \mathrm{N}(\mu_Y, \sigma_Y/\sqrt{m}) \end{array}\right\} \Rightarrow \bar{x} - \bar{y} \in \mathrm{N}\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right).$$

Con esta propiedad podemos calcular ya el intervalo que vendría dado por:

$$1 - \alpha = P\left((\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq \mu_X - \mu_Y \leq (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right).$$

3.5.5. Muestras independientes, varianzas desconocidas e iguales

Como en el caso del intervalo de confianza para la media cuando la varianza es desconocida se pasa de trabajar con valores críticos de la normal a trabajar con valores críticos de una t de Student ya que se estima esta varianza. De igual forma se procede ahora. Si suponemos que las varianzas de las dos poblaciones son iguales el mejor estimador de la varianza será:

$$\hat{S}_T^2 = \frac{(n-1)\hat{S}_X^2 + (m-1)\hat{S}_Y^2}{n+m-2},$$

que no es más que una adecuada ponderación de los mejores estimadores de cada población. Como en casos anteriores se puede demostrar bajo normalidad que

$$\frac{(n+m-2)\hat{S}_T^2}{\sigma^2} \in \chi^2_{n+m-2}.$$

Así, el intervalo de confianza para la diferencia de medias será

$$\left((\bar{x} - \bar{y}) - t_{n+m-2,\alpha/2}\hat{S}_T\sqrt{\frac{1}{n} + \frac{1}{m}} \leq \mu_X - \mu_Y \leq (\bar{x} - \bar{y}) + t_{n+m-2,\alpha/2}\hat{S}_T\sqrt{\frac{1}{n} + \frac{1}{m}}\right).$$

3.5.6. Muestras independientes, varianzas desconocidas y desiguales

Ahora no podemos tener una estimación global de la varianza así que el estadístico pivote que debemos utilizar es:

$$W = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\hat{S}_X^2}{n} + \frac{\hat{S}_Y^2}{m}}},$$

que asintóticamente tiene una distribución N(0,1) aunque la convergencia es lenta y por tanto poco precisa para valores muestrales pequeños. Para este caso (n ó $m < 30$) se

suele utilizar la aproximación de Welch según la cual el estadístico pivote sigue una distribución t con $g = n + m - 2 - \delta$, siendo δ el entero más próximo a:

$$\psi = \frac{\left[(m-1)\frac{\hat{S}_X^2}{n} - (n-1)\frac{\hat{S}_Y^2}{m}\right]^2}{(m-1)\left(\frac{\hat{S}_X^2}{n}\right)^2 + (n-1)\left(\frac{\hat{S}_Y^2}{m}\right)^2}.$$

3.5.7. Muestras apareadas, varianzas poblacionales conocidas

Para el caso de muestras apareadas no podemos utilizar los intervalos vistos anteriormente ya que estos suponen independencia de las poblaciones. En el caso de muestras apareadas precisamente se supone que hay dependencia entre las poblaciones lo que produce que cuando se calcule la varianza de la diferencia de medias: $\mathrm{Var}(\bar{x} - \bar{y}) = \mathrm{Var}(\bar{x}) + \mathrm{Var}(\bar{y}) - 2\mathrm{Cov}(\bar{x}, \bar{y})$ no debamos olvidarnos del término de la covarianza. Lo mejor en este caso en trabajar con la variable $D = X - Y$ y realizar el intervalo de confianza para la media de esta variable que usando lo ya conocido vendrá dado por:

$$\left(\bar{d} - t_{n-1,\alpha/2}\frac{\hat{S}_D}{\sqrt{n}} \leq \mu_D \leq \bar{d} + t_{n-1,\alpha/2}\frac{\hat{S}_D}{\sqrt{n}}\right).$$

3.5.8. Intervalo de confianza para la razón de varianzas de poblaciones normales

Hemos visto hace un momento dos situaciones diferentes para el caso de diferencia de medias de poblaciones normales con muestras independientes. La elección de la situación en la que nos encontramos depende de plantear si las varianzas desconocidas de las dos poblaciones pueden o no ser iguales. Para esto se plantea este intervalo de confianza de razón de varianzas, es decir, σ_Y^2/σ_X^2. Si este intervalo de confianza incluye al 1 entonces podemos suponer que ambas varianzas son iguales y por tanto decidirnos a utilizar el intervalo de confianza para la diferencia de medias supuestas las varianzas desconocidas e iguales. Teniendo en cuenta que las poblaciones son normales tenemos que

$$\frac{(n-1)\hat{S}_X^2}{\sigma_X^2} \in \chi_{n-1}^2 \quad \text{y} \quad \frac{(m-1)\hat{S}_Y^2}{\sigma_Y^2} \in \chi_{m-1}^2$$

y, dado que las poblaciones son independientes, podemos considerar el estadístico pivote:

$$W = \frac{\frac{(n-1)\hat{S}_X^2}{\sigma_X^2(n-1)}}{\frac{(m-1)\hat{S}_Y^2}{\sigma_Y^2(m-1)}} = \frac{\hat{S}_X^2\sigma_Y^2}{\hat{S}_Y^2\sigma_X^2},$$

que por construcción sigue una distribución $F_{n-1,m-1}$. Por tanto el intervalo de confianza vendrá dado por

$$\left(F_{n-1,m-1,1-\alpha/2}\frac{\hat{S}_Y^2}{\hat{S}_X^2}, F_{n-1,m-1,\alpha/2}\frac{\hat{S}_Y^2}{\hat{S}_X^2}\right)$$

donde $F_{n-1,m-1,1-\alpha/2}$ y $F_{n-1,m-1,\alpha/2}$ son los cuantiles de la distribución de orden $1-\alpha/2$ y $\alpha/2$.

3.6. Contrastes de hipótesis

Se trata en este apartado de comprobar empíricamente alguna hipótesis inicial sobre la población en estudio a partir de la información extraída de la muestra. Este es habitualmente el objetivo de la experimentación: avalar o rechazar afirmaciones que involucren alguna característica desconocida de la población. Esta toma de decisiones la englobaremos bajo el nombre de pruebas o contrastes de hipótesis. Se trata de formular una hipótesis sobre la población, se experimenta (se obtiene una muestra adecuada de la población) y por último se juzga si los resultados apoyan la hipótesis de partida.

3.6.1. Hipótesis estadística

Se denomina **hipótesis estadística** a cualquier conjetura sobre una o varias características de interés de un modelo de probabilidad. Ejemplos de este tipo de hipótesis será concretar un valor o rango de valores, establecer comparaciones entre parámetros de distintas poblaciones, especificar alguna característica sobre la forma de la distribución, asumir que una muestra ha sido tomada al azar, *etc.* Una **hipótesis paramétrica** es una afirmación sobre los valores de parámetros poblacionales desconocidos. Una hipótesis paramétrica se dice **simple** si especifica un único valor para cada parámetro poblacional desconocido. Se dice **compuesta** si asigna un conjunto de valores posibles a parámetros poblacionales desconocidos. Se denomina **hipótesis nula** que habitualmente se denota por H_0 a la hipótesis que se contrasta. La hipótesis nula debe ser la hipótesis que el experimentador asume como correcta y que no necesita ser probada (de ahí el nombre de nula). Cuando se acepta la hipótesis nula es porque no hay evidencia suficiente para refutarla. En este sentido si el experimentador quiere respaldar con contundencia un determinado argumento sólo podrá hacerlo a través del rechazo del argumento contrario (el establecido en H_0). Rechazar la hipótesis nula implica asumir como correcta una conjetura o hipótesis complementaria que se conoce como **hipótesis alternativa** y suele denotarse como H_1. La elección de la hipótesis alternativa también puede condicionar las propiedades analíticas del criterio probabilístico seleccionado para discernir entre H_0 y H_1. Una función de los datos muestrales y del valor del parámetro especificado por la hipótesis nula, con distribución conocida cuando H_0 es cierta, se denomina **estadístico de contraste** o **medida de discrepancia**. Habitualmente el estadístico de contraste tomará valores "pequeños" cuando la hipótesis nula es cierta y valores "grandes" cuando es cierta la alternativa. Por tanto, la manera de razonar será: "Preferimos la hipótesis nula salvo que la medida de discrepancia sea suficientemente grande en cuyo caso preferiremos la hipótesis alternativa". En este proceso podemos cometer dos tipos de errores. El **error tipo I** se produce cuando se toma la decisión de rechazar la hipótesis nula siendo esta cierta. El **error tipo II** es el que se comete al no rechazar la hipótesis nula cuando esta es falsa. Asumido que toda decisión está sujeta a error y

establecidos los dos tipos de error posibles, parece claro que cualquier criterio para optar por una u otra hipótesis atenderá a controlar el riesgo de equivocarse. El planteamiento clásico del contraste de hipótesis consiste en controlar el riesgo de cometer un error tipo I. Este enfoque significa que el que decide otorga su crédito inicial a la hipótesis nula y sólo está dispuesto a rechazarla si la evidencia en su contra es muy importante (como en un juicio penal). Con esta forma de proceder y dado que el estadístico de contraste tiene distribución conocida cuando H_0 es cierta bastará con prefijar de antemano la probabilidad de cometer un error tipo I. Llamaremos nivel de significación de un contraste (α) a la probabilidad de cometer un error tipo I, es decir, $\alpha = P(\text{rechazar } H_0/H_0 \text{ es cierta})$. Los valores más habituales para este nivel son 0,01, 0,05 y 0,1. Una elección cualquiera dividirá en dos regiones el conjunto de posibles valores del estadístico de contraste: una de probabilidad α (bajo H_0) que se llamará **región de rechazo o crítica** y otra de probabilidad $1-\alpha$ que llamaremos **región de aceptación**. Si el estadístico toma valores en la región de aceptación diremos que el contraste es estadísticamente no significativo. Si por el contrario toma valores en la región de rechazo diremos que el contraste es estadísticamente significativo. La ubicación y forma de las regiones dependerá del tipo de hipótesis alternativa. Si la región de rechazo está formada por las dos colas de la distribución del estadístico de contraste bajo H_0 se dice un contraste bilateral. Si por el contrario la región de rechazo está formada por una cola de la distribución del estadístico bajo H_0 se dirá un contraste unilateral.

Cuando sea necesario controlar el error tipo II hay que tener en cuenta que normalmente la hipótesis alternativa es una hipótesis compuesta y por tanto este error tipo II es una función de los elementos que pudieran estar bajo la hipótesis H_1. El error tipo II se denota normalmente por β.

Los pasos a seguir para realizar un contraste son:

1. Especificar las hipótesis nula (H_0) y la alternativa (H_1).
2. Elegir un estadístico de contraste apropiado.
3. Fijar el nivel de significación α en base a como de importante se considere el error tipo I.
4. Prefijado α y el estadístico, construir las regiones de rechazo y aceptación.
5. Determinar cuál es el primer valor de la hipótesis alternativa que, de ser correcto, deseamos detectar con el contraste. Especificar el tamaño del error tipo II que estamos dispuestos a asumir.
6. En base a las probabilidades α y β calcular el tamaño muestral adecuado para garantizar ambas probabilidades de error.
7. Tomar la muestra y evaluar el estadístico de contraste.
8. Concluir si el test es estadísticamente significativo o no al nivel de significación α según se ubique en la región de rechazo.

Los pasos 5 y 6 sólo se ejecutarán si es necesario controlar el error tipo II.

Se llama **nivel crítico** o ***p*-valor** a la probabilidad de obtener una discrepancia mayor o igual que el estadístico de contraste cuando H_0 es cierta. Claramente, cuando el nivel crítico sea mayor que el valor de α, se aceptará la hipótesis nula en tanto que valores menores que α no decantan por la hipótesis alternativa.

Se llama **función de potencia** a la probabilidad de rechazar H_0 en función del parámetro de interés. Se suele denotar por $\pi(\theta)$. Por tanto, para los valores de la hipótesis alternativa $\pi(\theta) = 1-\beta(\theta)$ y para los valores de la hipótesis nula $\pi(\theta) \leq \alpha$ (Si la hipótesis nula es simple entonces $\pi(\theta) = \alpha$).

3.6.2. Contraste de hipótesis para la media de una población normal

Sea una muestra aleatoria simple $\{x_1,...,x_n\}$ de una distribución teórica N(μ,σ). Nos ocuparemos ahora de hipótesis del estilo: $H_0 : \mu = \mu_0$ contra alternativas compuestas ($H_1 : \mu \neq \mu_0$, $H_1 : \mu < \mu_0$, $H_1 : \mu > \mu_0$). Para ello utilizaremos los estadísticos pivote estudiados en anteriores apartados:

- $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \in N(0,1)$, si la varianza poblacional es conocida.
- $\frac{\bar{x}-\mu_0}{\hat{S}/\sqrt{n}} \in t_{n-1}$, si la varianza poblacional es desconocida.

Ambos estadísticos miden la discrepancia entre la información muestral ($\bar{x}$) y el valor conjeturado por la hipótesis nula (μ_0) y además sabemos las distribuciones de ambos si H_0 es cierta. Por tanto lo único que nos quedará por hacer es calcular las regiones de aceptación y rechazo para cada una de las posibles alternativas. En el caso de la primera hipótesis alternativa la región de aceptación coincide con el intervalo de confianza para la media de una población normal y por tanto la región de rechazo será $\left(-\infty, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}, +\infty\right)$. En el caso de la hipótesis alternativa de las otras hipótesis alternativas la región de rechazo será, respectivamente, $(-\infty, -z_\alpha)$ y $(z_\alpha, +\infty)$.

Otra cuestión interesante es averiguar el tamaño que debe tener la muestra para que el contraste realizado a un nivel α resulte significativo cuando el valor real de μ sea $\mu_0 + \varepsilon$ tenga una potencia fijada de antemano $1 - \beta(\mu_0 + \varepsilon)$. Consideremos la alternativa $H_1 : \mu > \mu_0$ y supondremos conocida la varianza (el caso más simple). El estadístico pivote

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

sigue una N(0,1) si H_0 es cierta pero ahora se quiere calcular

$$P(\text{no rechazar } H_0 / H_1 \text{ es cierta}).$$

Por tanto, bajo la hipótesis de que H_1 es cierta tenemos que la verdadera media es $\mu = \mu_0 + \varepsilon$ y el estadístico pivote

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \in N\left(\frac{\varepsilon}{\sigma/\sqrt{n}}, 1\right).$$

La probabilidad que queremos calcular es por tanto

$$P\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \middle| H_1\right) = P\left(Z \leq -\frac{\varepsilon}{\sigma/\sqrt{n}} + z_\alpha \middle| N(0,1)\right)$$
$$\Rightarrow -\frac{\varepsilon}{\sigma/\sqrt{n}} + z_\alpha = z_{1-\beta(\mu_0+\varepsilon)} = -z_{\beta(\mu_0+\varepsilon)}$$

de donde se deduce que

$$n = \left(\frac{z_\alpha + z_{\beta(\mu_0+\varepsilon)}}{\varepsilon}\right)^2 \sigma^2,$$

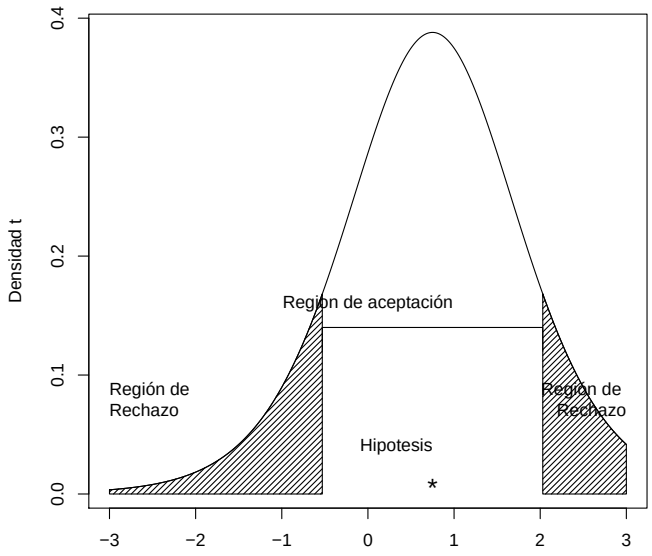
que especifica el tamaño muestral como función del nivel de significación y del error tipo II cuando $\mu = \mu_0 + \varepsilon$. Este tamaño es el tamaño muestral mínimo para que con cierta seguridad $(100(1-\beta)\,\%)$ que realicemos el contraste con significación α detectemos que la verdadera media dista ε de μ_0.

Si la varianza no fuese conocida entonces tenemos la dificultad de calcular la distribución bajo H_1 ya que esta sería una t no central cuyo centro sería ε/σ. Como σ es desconocida sólo podríamos resolver el problema cuando $\varepsilon = k\ \sigma$.

Si la población de partida no fuese normal, el teorema central de límite garantiza que para un tamaño muestral grande la distribución de la media se aproxima a la normal. Por supuesto para muestras pequeñas deberíamos utilizar otros contrastes.

```
data(sleep);attach(sleep)
x<-extra[group==1];alpha<-0.05
test.stat<-t.test(x,y=NULL,mu=mean(x),paired=F,conf.level=1-alpha)
bound.left<--3.0;bound.right<-3.0
df<-length(x)-1
xaxis<-seq(bound.left,bound.right,length=1000)
yaxis<-dt(xaxis-test.stat$estimate,df)
plot(xaxis,yaxis,type="l",xlab="",ylab="Densidad t")
crit.left<-test.stat$conf.int[1]
crit.right<-test.stat$conf.int[2]
xaxis<-seq(bound.left,crit.left,length=100)
yaxis<-c(dt(xaxis-test.stat$estimate,df),0,0)
xaxis<-c(xaxis,crit.left,bound.left)
polygon(xaxis,yaxis,density=25)
xaxis<-seq(crit.right,bound.right,length=100)
yaxis<-c(dt(xaxis-test.stat$estimate,df),0,0)
xaxis<-c(xaxis,bound.right,crit.right)
polygon(xaxis,yaxis,density=25)
points(test.stat$null.value,0.01,pch="*",cex=1.5)
text(test.stat$null.value,0.04,"Hipotesis",adj=1)
text(bound.left,0.08,"Región de \nRechazo",adj=0)
text(bound.right,0.08,"Región de \nRechazo",adj=1)
text((bound.left+bound.right)/2,0.16,"Region de aceptación")
```

```
xaxis<-c(rep(crit.left,2),rep(crit.right,2))
yaxis<-c(0.12,0.14,0.14,0.12)
lines(xaxis,yaxis)
```



3.6.3. Contraste de hipótesis para la varianza de una población normal

Sea una muestra aleatoria simple $\{x_1,...,x_n\}$ de una distribución teórica N(μ,σ). Nos ocuparemos ahora de hipótesis del estilo: $H_0 : \sigma^2 = \sigma_0^2$ contra las usuales alternativas compuestas ($H_1 : \sigma^2 \neq \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$). Para ello utilizaremos el estadístico pivote

$$\frac{(n-1)\hat{S}^2}{\sigma^2} \in \chi^2_{n-1}$$

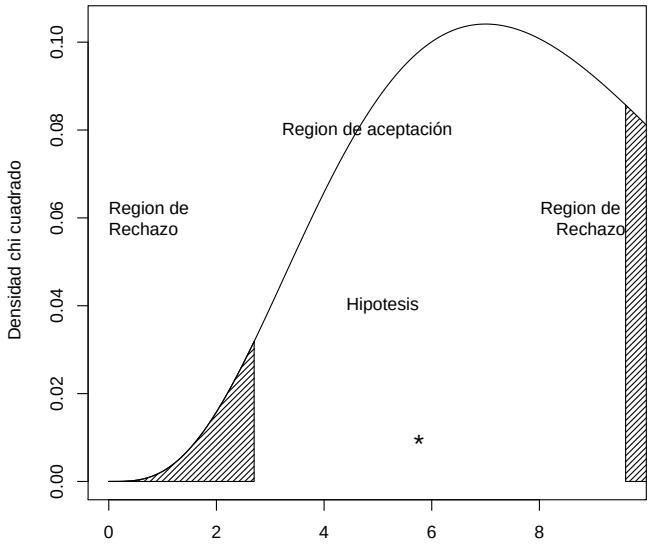
estudiado anteriormente. Las regiones de rechazo serán, respectivamente,

$$\left(0, \chi^2_{n-1,1-\alpha/2}\right) \cup \left(\chi^2_{n-1,\alpha/2}, +\infty\right), \quad \left(0, \chi^2_{n-1,1-\alpha}\right) \quad \text{y} \quad \left(\chi^2_{n-1,\alpha}, +\infty\right)$$

a imitación del caso anterior.

```
data(sleep);attach(sleep)
x<-extra[group==1];alpha<-0.05
bound.left<-0.0;bound.right<-3*var(x)
df<-length(x)-1;H0<-5.0
xaxis<-seq(bound.left,bound.right,length=1000)
yaxis<-dchisq(xaxis,df)
plot(xaxis,yaxis,type="l",xlab="",ylab=" Densidad chi cuadrado")
crit.left<-qchisq(0.025,df=df)
```

```
crit.right<-qchisq(0.975,df=df)
xaxis<-seq(bound.left,crit.left,length=100)
yaxis<-c(dchisq(xaxis,df),0,0)
xaxis<-c(xaxis,crit.left,bound.left)
polygon(xaxis,yaxis,density=25)
xaxis<-seq(crit.right,bound.right,length=100)
yaxis<-c(dchisq(xaxis,df),0,0)
xaxis<-c(xaxis,bound.right,crit.right)
polygon(xaxis,yaxis,density=25)
estad<-var(x)*(length(x)-1)/H0
points(estad,0.01,pch="*",cex=1.5)
text(estad,0.04,"Hipotesis",adj=1)
text(bound.left,0.06,"Region de \nRechazo",adj=0)
text(bound.right,0.06,"Region de \nRechazo",adj=1)
text((bound.left+bound.right)/2,0.08,"Region de aceptación")
xaxis<-c(rep(crit.left,2),rep(crit.right,2))
yaxis<-c(0.12,0.14,0.14,0.12)
lines(xaxis,yaxis)
```



3.6.4. Contraste de hipótesis para la diferencia de medias de poblaciones normales

Sean dos muestras aleatorias simples $\{x_1,...,x_n\}$ y $\{y_1,...,y_m\}$ de distribuciones teóricas $N(\mu_X,\sigma_X)$ y $N(\mu_Y,\sigma_Y)$. Nuestro objetivo ahora será contrastar la hipótesis nula $H_0 : \mu_X = \mu_Y$. El contraste se puede plantear equivalentemente como $H_0 : \mu_X - \mu_Y = 0$.

Debemos distinguir las diferentes situaciones que ya consideramos en el caso de intervalos de confianza.

Muestras independientes, varianzas poblacionales conocidas

Ya sabemos que el estadístico que la diferencia de medias de muestras independientes sigue la distribución normal que además bajo H_0 significa que

$$\bar{x} - \bar{y} \in N\left(0, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right),$$

lo que nos sugiere como estadístico de contraste a:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \in N(0,1).$$

Las regiones de rechazo son como en el caso del contraste para la media.

Muestras independientes, varianzas desconocidas e iguales

Imitando el apartado homólogo en intervalos de confianza ahora el estadístico de contraste será

$$\frac{\bar{x} - \bar{y}}{\hat{S}_T \sqrt{\frac{1}{n} + \frac{1}{m}}} \in t_{n+m-2},$$

donde como antes

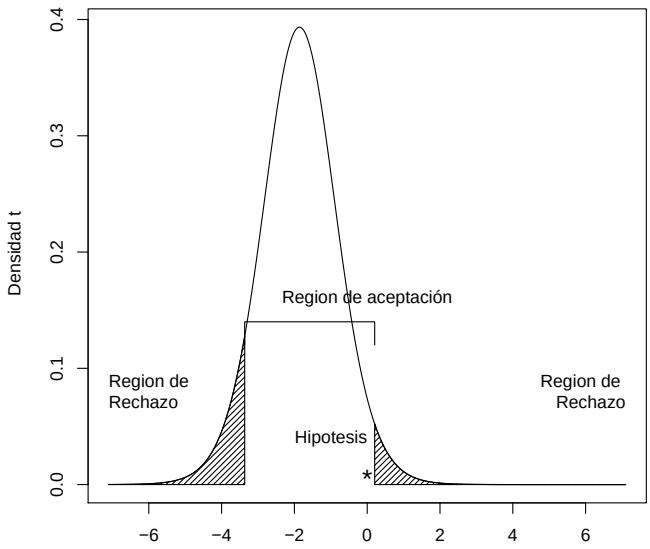
$$\hat{S}_T = \frac{(n-1)\hat{S}_X^2 + (m-1)\hat{S}_Y^2}{n+m-2}.$$

La región de aceptación será entonces

$$\left((\bar{x} - \bar{y}) - t_{n+m-2,\alpha/2}\hat{S}_T\sqrt{\frac{1}{n} + \frac{1}{m}}, (\bar{x} - \bar{y}) + t_{n+m-2,\alpha/2}\hat{S}_T\sqrt{\frac{1}{n} + \frac{1}{m}}\right).$$

```
data(sleep);attach(sleep)
x<-extra[group==1];y<-extra[group==2]
df<-length(x)+length(y)-2;alpha<-0.05
bound.left<-min((min(x)-max(y)),(min(y)-max(x)))
bound.right<-max((max(x)-min(y)),(max(y)-min(x)))
test.stat<-t.test(x,y,paired=F,conf.level=1-alpha)
xaxis<-seq(bound.left,bound.right,length=1000)
yaxis<-dt(xaxis-test.stat$statistic,df)
plot(xaxis,yaxis,type="l",xlab="",ylab=" Densidad t")
crit.left<-test.stat$conf.int[1]
crit.right<-test.stat$conf.int[2]
xaxis<-seq(bound.left,crit.left,length=100)
```

```
yaxis<-c(dt(xaxis-test.stat$statistic,df),0,0)
xaxis<-c(xaxis,crit.left,bound.left)
polygon(xaxis,yaxis,density=25)
xaxis<-seq(crit.right,bound.right,length=100)
yaxis<-c(dt(xaxis-test.stat$statistic,df),0,0)
xaxis<-c(xaxis,bound.right,crit.right)
polygon(xaxis,yaxis,density=25)
points(test.stat$null.value,0.01,pch="*",cex=1.5)
text(test.stat$null.value,0.04,"Hipotesis",adj=1)
text(bound.left,0.08,"Region de \nRechazo",adj=0)
text(bound.right,0.08,"Region de \nRechazo",adj=1)
text((bound.left+bound.right)/2,0.16,"Region de aceptación")
xaxis<-c(rep(crit.left,2),rep(crit.right,2))
yaxis<-c(0.12,0.14,0.14,0.12)
lines(xaxis,yaxis)
```



Muestras independientes, varianzas desconocidas y desiguales

Como en intervalos de confianza, el estadístico de contraste será:

$$W = \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{\hat{S}_X^2}{n} + \frac{\hat{S}_Y^2}{m}}},$$

cuya distribución bajo H_0 tiende lentamente a la N(0,1). Para muestras pequeñas (n ó $m < 30$) se suele utilizar la aproximación de Welch según la cual el estadístico pivote sigue una distribución t con $g = n + m - 2 - \delta$ siendo δ el entero más próximo a:

$$\psi = \frac{\left[(m-1)\frac{\hat{S}_X^2}{n} - (n-1)\frac{\hat{S}_Y^2}{m}\right]^2}{(m-1)\left(\frac{\hat{S}_X^2}{n}\right)^2 + (n-1)\left(\frac{\hat{S}_Y^2}{m}\right)^2}.$$

Muestras apareadas, varianzas poblacionales conocidas

Al igual que se hizo anteriormente la solución natural para este problema es considerar la variable $D = X - Y$ y tratar ahora el contraste $H_0 : \bar{d} = 0$, (no debemos olvidarnos de lo conocido para $\mathrm{Var}(\bar{x} - \bar{y})$ en este caso) suponiendo varianza desconocida. Por tanto el estadístico de contraste será

$$\frac{\bar{d}}{\hat{S}_D \Big/ \sqrt{n}} \in t_{n-1}.$$

3.6.5. Contraste de hipótesis para la razón de varianzas de poblaciones normales

Sean dos muestras aleatorias simples $\{x_1,...,x_n\}$ y $\{y_1,...,y_m\}$ de distribuciones teóricas $\mathrm{N}(\mu_X,\sigma_X)$ y $\mathrm{N}(\mu_Y,\sigma_Y)$. Se trata ahora, siguiendo el paralelismo empleado en intervalos de confianza, de verificar la hipótesis de igualdad de varianzas mediante el estadístico razón de varianzas, es decir, σ_Y^2/σ_X^2. Si se puede aceptar el valor uno como perteneciente a la hipótesis nula podemos suponer que las varianzas de ambas poblaciones son iguales. Como ya se ha razonado anteriormente el estadístico pivote que vamos a utilizar es:

$$W = \frac{\frac{(n-1)\hat{S}_X^2}{\sigma_X^2(n-1)}}{\frac{(m-1)\hat{S}_Y^2}{\sigma_Y^2(m-1)}} = \frac{\hat{S}_X^2\sigma_Y^2}{\hat{S}_Y^2\sigma_X^2},$$

que sigue una distribución $F_{n-1,m-1}$. Por tanto la región de aceptación en el caso de un test bilateral vendrá dada por

$$\left(F_{n-1,m-1,1-\alpha/2}\frac{\hat{S}_Y^2}{\hat{S}_X^2}, F_{n-1,m-1,\alpha/2}\frac{\hat{S}_Y^2}{\hat{S}_X^2}\right),$$

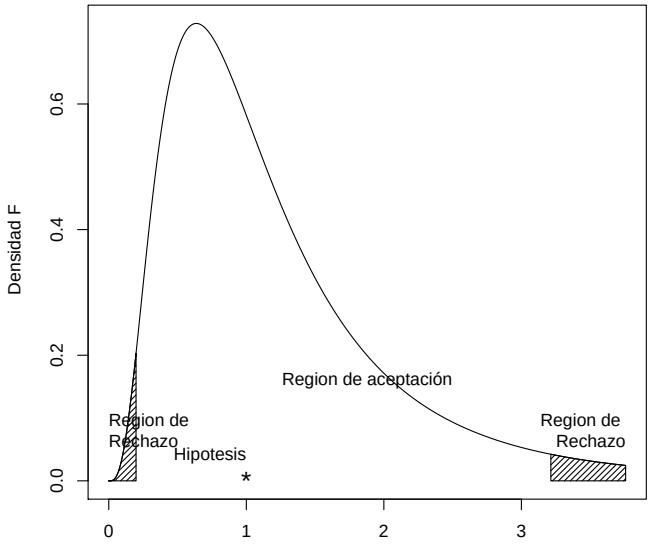
donde $F_{n-1,m-1,1-\alpha/2}$ y $F_{n-1,m-1,\alpha/2}$ son los cuantiles de la distribución de orden $1-\alpha/2$ y $\alpha/2$. En el caso de hipótesis unilaterales, las regiones de aceptación serán

$$\left(0, F_{n-1,m-1,\alpha}\frac{\hat{S}_Y^2}{\hat{S}_X^2}\right) \text{ o } \left(F_{n-1,m-1,1-\alpha}\frac{\hat{S}_Y^2}{\hat{S}_X^2}, \infty\right)$$

dependiendo del sentido de la desigualdad en la hipótesis nula.

```
data(sleep);attach(sleep)
x<-extra[group==1];y<-extra[group==2]
```

```
df<-length(x)+length(y)-2;alpha<-0.05
test.stat<-var.test(x,y,ratio=1,conf.level=1-alpha)
bound.left<-0
bound.right<-3*var(y)/var(x)
xaxis<-seq(bound.left,bound.right,length=1000)
yaxis<-df(xaxis,df1=length(y)-1,df2=length(x)-1)
plot(xaxis,yaxis,type="l",xlab="",ylab=" Densidad F")
crit.left<-test.stat$conf.int[1]
crit.right<-test.stat$conf.int[2]
xaxis<-seq(bound.left,crit.left,length=100)
yaxis<-c(df(xaxis,df1=length(y)-1,df2=length(x)-1),0,0)
xaxis<-c(xaxis,crit.left,bound.left)
polygon(xaxis,yaxis,density=25)
xaxis<-seq(crit.right,bound.right,length=100)
yaxis<-c(df(xaxis,df1=length(y)-1,df2=length(x)-1),0,0)
xaxis<-c(xaxis,bound.right,crit.right)
polygon(xaxis,yaxis,density=25)
points(test.stat$null.value,0.01,pch="*",cex=1.5)
text(test.stat$null.value,0.04,"Hipotesis",adj=1)
text(bound.left,0.08,"Region de \nRechazo",adj=0)
text(bound.right,0.08,"Region de \nRechazo",adj=1)
text((bound.left+bound.right)/2,0.16,"Region de aceptación")
xaxis<-c(rep(crit.left,2),rep(crit.right,2))
yaxis<-c(0.12,0.14,0.14,0.12)
lines(xaxis,yaxis)
```



3.6.6. Relación entre intervalos de confianza y contrastes de hipótesis.

A la vista de lo explicado hasta este momento, parece evidente que existe relación entre los intervalos de confianza y los contrastes de hipótesis. De hecho, las regiones de aceptación de los diversos contrastes a nivel de significación α son intervalos de confianza a nivel de confianza $1-\alpha$. Estamos hablando de las dos caras de la misma moneda. Si un determinado valor θ_0 pertenece al intervalo de confianza $1-\alpha$, entonces la hipótesis nula $H_0 : \theta = \theta_0$ será aceptada frente a la alternativa $H_1 : \theta \neq \theta_0$ cuando el contraste se realiza a nivel α. Y viceversa. Esta analogía responde, además de a la similitud de ambos planteamientos, al hecho de que los estadísticos pivotes utilizados para la construcción de intervalos de confianza coinciden con los estadísticos de contraste empleados en contraste de hipótesis.

3.7. Ejercicio resuelto

Ejercicio: En 20 días lectivos y a la misma hora se ha observado el número de terminales de una universidad conectados a Internet. Los resultados son los siguientes

$$1027, 1023, 1369, 950, 1436, 957, 634, 821, 882, 942,$$
$$904, 984, 1067, 570, 1063, 1307, 1212, 1045, 1047, 1178.$$

Se pide:

a) Calcular el intervalo de confianza al 95 % para el número medio de terminales conectados a Internet.

b) Calcular el intervalo de confianza al 95 % para la varianza del número de terminales conectados a internet.

c) ¿Qué tamaño muestral es necesario para obtener el intervalo de confianza al 95 % para el número medio de terminales conectados a Internet con una longitud inferior a 30 unidades?

Solución:

a) Para este apartado necesitamos los estadísticos básicos de la muestra que son: $n = 20$, $\overline{x} = 1020{,}9$, y $\widehat{s} = 215{,}74$. El intervalo de confianza para la media viene determinado por el estadístico pivote:

$$-t_{n-1}^{\alpha/2} \leq \frac{\overline{x}-\mu}{\frac{\widehat{s}}{\sqrt{n}}} \leq t_{n-1}^{\alpha/2} \Longrightarrow \mu \in \left[\overline{x} \pm t_{n-1}^{\alpha/2}\frac{\widehat{s}}{\sqrt{n}}\right].$$

Haciendo los cálculos

$$\mu \in \left[\overline{x} \pm t_{n-1}^{\alpha/2}\frac{\widehat{s}}{\sqrt{n}}\right] = \left[1020{,}9 \pm 2{,}09 \times \frac{215{,}74}{\sqrt{20}}\right] = [920{,}08, 1121{,}72].$$

b) Para el segundo apartado usaremos el estadístico,

$$\frac{(n-1)\,\widehat{s}^2}{\sigma^2} \in \chi^2_{n-1} \Longrightarrow \sigma^2 \in \left[\frac{(n-1)\,\widehat{s}^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,\widehat{s}^2}{\chi^2_{n-1,1-\alpha/2}}\right].$$

Entonces

$$\sigma^2 \in \left[\frac{(n-1)\,\widehat{s}^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,\widehat{s}^2}{\chi^2_{n-1,1-\alpha/2}}\right] = \left[\frac{884331{,}2}{32{,}9}, \frac{884331{,}2}{8{,}91}\right] = [26879{,}37, 99251{,}54]\,.$$

c) Para este apartado, si conociesemos la varianza, la longitud del intervalo viene dada por $L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. En este caso no conocemos σ, pero como el tamaño muestral que vamos a necesitar va a ser mayor que el que tenemos del apartado 1 (ese tiene longitud 201.68), la t que vamos a necesitar tendrá muchos más grados de libertad y podremos aproximarla por la distribución normal estándar. Por tanto buscamos n tal que:

$$L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 30 \Longrightarrow \sqrt{n} > 2z_{\alpha/2}\frac{\sigma}{30}$$
$$\Longrightarrow \sqrt{n} > 2 \times 1{,}96 \times \frac{215{,}74}{30} = 28{,}19 \Longrightarrow n > 795.$$

Capítulo 4

Inferencia no paramétrica

4.1. Introducción

En general, al estimar los parámetros de una población se supone que los datos siguen una distribución conocida. Esto permite extraer conclusiones que lógicamente dependen de esas suposiciones iniciales. A estas hipótesis se las conoce como *hipótesis estructurales*. Por ejemplo, si queremos construir un intervalo de confianza para la varianza de una población, debemos recordar que la fórmula vista en apartados anteriores es solamente aplicable cuando la población es normal. Si esta suposición falla, esta fórmula sólo nos dará buenos resultados cuando tengamos muchos datos. Por tanto, conviene tener herramientas que nos permitan estudiar si estas hipótesis básicas están o no en contradicción con los datos.

En concreto, nos fijaremos en las siguientes hipótesis

- Hipótesis sobre la distribución.
- Hipótesis sobre la posición de dos o más grupos.
- Hipótesis de independencia.
- Hipótesis de homogeneidad.

4.2. Hipótesis sobre la distribución

En muchos de los contrastes paramétricos se hacen suposiciones de normalidad que pueden no cumplirse en la práctica. De hecho, las inferencias sobre la varianza son muy sensibles a la hipótesis de normalidad. Nuestro interés estará centrado ahora en comprobar si los datos provienen de una distribución determinada y esto se conseguirá a través de dos contrastes básicos: el χ^2 de Pearson y el test de Kolmogorov-Smirnov. También se comentarán los tests específicos para normalidad.

4.2.1. El contraste χ^2 de Pearson

Es el contraste de ajuste a una distribución más antiguo debido a Pearson, cuya idea es comparar las frecuencias observadas en un histograma o diagrama de barras con las especificadas por el modelo teórico que se contrasta. Sirve tanto para distribuciones continuas como discretas, a diferencia del test de Kolmogorov-Smirnov que veremos luego que sólo sirve para distribuciones continuas.

El test se plantea en los siguientes términos, contrastamos

$H_0 : X \in F_0$

$H_1 : X \notin F_0$

Como en cualquier problema de contraste de hipótesis a H_0 la llamaremos hipótesis nula, y a H_1 la llamaremos hipótesis alternativa. A partir de una muestra $X = (x_1, \ldots, x_n)$ aleatoria simple de n datos ($n \geq 25$) debemos determinar cual de los dos hipótesis anteriores es preferible.

El contraste se realiza como sigue:

1. Agrupar los n datos en k clases, donde $k \geq 5$. Las clases se eligen de manera que cubran todo el rango posible de valores de la variable y que cualquier posible dato quede clasificado sin ambigüedad. Es conveniente además, cuando esto sea posible, intentar elegirlas con el mismo número de datos en cada clase. Cada clase debe contener al menos tres datos. Llamaremos O_i al número de datos de la muestra que caen en la clase i (*observados* en dicha clase).

2. Calcular la probabilidad p_i que el modelo supuesto (distribución H_0) asigna a cada clase y calcular para cada clase $E_i = np_i$ (*esperados* en dicha clase).

3. Calcular la discrepancia entre lo observado y lo esperado mediante

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

que se distribuye como una χ^2 cuando el modelo especificado en H_0 es correcto. Los grados de libertad de esta distribución serán $k - r - 1$ donde r son los parámetros estimados de la distribución especificada en H_0.

4. Rechazaremos el modelo cuando la probabilidad de obtener una discrepancia mayor o igual que la observada sea suficientemente baja. Es decir, cuando $X^2 \geq \chi^2_{k-r-1,\alpha}$ para un cierto α pequeño (en general $\alpha = 0{,}05$). ($\chi^2_{k-r-1,\alpha}$ representa el valor crítico de una χ^2 con $k-r-1$ grados de libertad que deja a la derecha un área α).

Notas:

Por supuesto, cuantas más clases tengamos (con las limitaciones mencionadas antes) mejor funcionará el test puesto que más grados de libertad tendrá la distribución del estadístico. Un inconveniente de este contraste es que dos modelos distintos, que asignen la misma probabilidad a las mismas clases obtendrán las mismas discrepancias, y por tanto el mismo resultado.

También resulta interesante comprobar el signo de $O_i - E_i$ para detectar posibles tendencias de la distribución.

Ejemplo: Se han recogido 6115 grupos de 12 individuos de una especie de mosquito contándose para cada grupo el número de machos y hembras. Los resultados fueron los siguientes:

M-H	12-0	11-1	10-2	9-3	8-4	7-5	6-6	5-7	4-8	3-9	2-10	1-11	0-12
Obs.	7	45	181	478	829	1112	1343	1033	670	286	104	24	3

Se trata de comprobar si puede suponerse que existe la misma proporción de machos que de hembras en dicha especie.

Solución: Si estamos contando el número de machos y hembras que hay en cada grupo de 12, podemos pensar que estamos realizando un experimento aleatorio de 12 tiradas donde el éxito del experimento es el número de machos que tenemos (o de hembras, es indiferente). Bajo este punto de vista, estos experimentos sabemos que siguen una distribución binomial de tamaño 12 y probabilidad de éxito p (ahora desconocida) (B(n,p)). También como es conocido la probabilidad de obtener i éxitos en una distribución B(n,p) es:

$$P(X = i) = P(i) = \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1-p)^{n-i}.$$

Queremos contrastar que existe la misma proporción de machos que de hembras, por tanto para determinar nuestra distribución supondremos p=0.5. De manera que la probabilidad de cada grupo con i machos será:

$$P(i) = \left(\begin{array}{c} 12 \\ i \end{array} \right) 0{,}5^i (1-0{,}5)^{12-i}.$$

Como hemos hecho el experimento 6.115 veces, el número de grupos esperado de cada clase con i machos será $E(i) = 6.115 \times P(i)$. Por tanto, tendremos llamando $X_i^2 = (O_i - E_i)^2 / E_i$ y calculando el signo del numerador antes de elevar al cuadrado:

N^o	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
O$_i$	7	45	181	478	829	1112	1343	1033	670	286	104	24	3
E$_i$	1,49	17,92	98,5	328,4	738.9	1182,4	1379,5	1182,4	738,9	328,4	98,53	17,92	1,49
X_i^2	20,4	40,92	69,03	68,10	10,98	4,19	0,96	18,87	6,43	5,48	0,30	2,06	1,53
Signo	+	+	+	+	+	-	-	-	-	-	+	+	+

Podemos calcular ahora el valor del estadístico X^2 sin más que sumar la 4^a fila y obtenemos $X^2 = 249{,}23$. Este valor debemos compararlo con el teórico $\chi^2_{k-r-1,\alpha}$, donde $k = 13$ (número de clases), $r = 0$ (no estimamos ningún parámetro) y $\alpha = 0{,}05$ (habitual cuando no se dice nada). $\chi^2_{12,0,05} = 21{,}0$ (se mira en las tablas de la χ^2 donde cruzan 12 y 0,95). Como $X^2 = 249{,}23 > \chi^2_{0,05}(12) = 21{,}0$ entonces rechazamos la hipótesis de que exista la misma proporción de machos que de hembras. Además, si nos fijamos en la fila

de los signos, primero tenemos una racha de varios signos +, luego otra de varios signos −, y finalmente una última racha de tres signos +. Esto indicaría que hay una cierta tendencia a que existan más machos que hembras en esta especie, aunque los últimos signos + podrían indicar una subpoblación dentro de la especie de este mosquito, donde la hembra predomine.

```
> golesdep<-c(2,3,5,2,2,3,2,1,0,1,2,1,2,2,1,2,1,0,4)
> lambda<-mean(golesdep)
> oi<-table(factor(golesdep,0:3))
> oi<-c(oi,sum(table(factor(golesdep)))-sum(oi))
> probs<-dpois(0:3,lambda=lambda)
> probs<-c(probs,1-sum(probs))
> probs*sum(oi)
[1] 2.856800 5.412884 5.127996 3.238734 2.363586
> chisq.test(oi,p=probs,simulate.p.value=T)

        Chi-squared test for given probabilities

data: oi
X-squared = 2.4267, df = 4, p-value = 0.6578
```

4.2.2. El test de Kolmogorov-Smirnov

Este constraste compara la función de distribución teórica con la empírica y sólo es válido para variables continuas. El test se plantea en los mismos términos que el de Pearson visto anteriormente:

$H_0 : X \in F_0$.

$H_1 : X \notin F_0$.

Suponemos que tenemos una muestra $X = (x_1, \ldots, x_n)$ aleatoria simple que proviene de un modelo continuo $F(x)$. El contraste se realiza como sigue:

1. Ordenar los valores muestrales en orden creciente: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

2. Calcular la función de distribución empírica de la muestra $F_n(x)$, con :

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)}, \\ \frac{r}{n} & \text{si } x_{(r)} \leq x < x_{(r+1)}, \\ 1 & \text{si } x \geq x_{(n)}. \end{cases}$$

3. Calcular la discrepancia máxima entre las funciones de distribución observada (o empírica) y teórica con el estadístico:

$$D_n = \text{máx}\,|F_n(x) - F(x)|,$$

cuya distribución, bajo H_0 se ha tabulado. Si la distancia calculada D_n es mayor que la encontrada en las tablas fijado α, (D(α,n)), rechazaremos el modelo $F(x)$.

Notas

- Este contraste tiene la ventaja de que no es necesario agrupar los datos.
- Si estimamos los parámetros de la población mediante la muestra, la distribución de D_n es sólo aproximada convirtiéndose en un contraste conservador (es decir, tendente a aceptar siempre). Existen versiones de este test para esos casos (ver test de Kolmogorov-Smirnov-Lilliefors para normalidad).
- Para calcular el estadístico D_n dada una muestra hay que pensar que la distribución empírica es constante por tramos, de manera que la máxima distancia entre $F_n(x)$ y $F(x)$ se da en los puntos de salto (los puntos de la muestra). En un punto de la muestra x_h la distancia entre las funciones $F_n(x)$ y $F(x)$, $D_n(x_h)$, se calcula como:

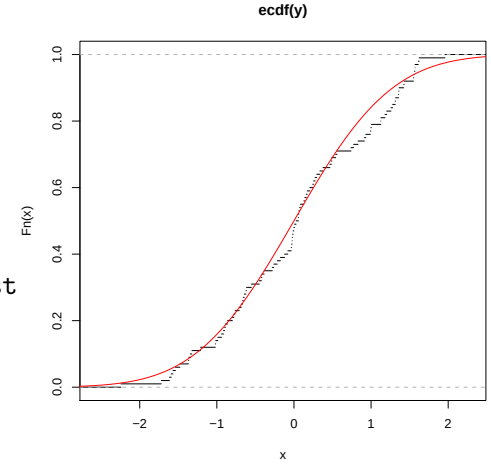
$$D_n(x_h) = \text{máx}\left\{\left|F_n(x_{h-1}) - F(x_h)\right|, \left|F_n(x_h) - F(x_h)\right|\right\}.$$

  y para calcular D_n sólo hay que calcular el máximo de los $D_n(x_h)$.

```
> library(stats)
> y<-rnorm(100)
> plot(ecdf(y),do.points=F)
> x<-seq(-3.5,3.5,length=100)
> lines(x,pnorm(x),col=2)
> ks.test(y,"pnorm",mean=0,sd=1)

   One-sample Kolmogorov-Smirnov test

data:  y
D = 0.0801, p-value = 0.5423
alternative hypothesis: two-sided
```



4.2.3. El contraste de Shapiro-Wilks

Este contraste mide el ajuste de la muestra al dibujarla en papel probabilístico normal a una recta. Se rechaza normalidad cuando el ajuste es malo que se corresponde a valores pequeños del estadístico. El estadístico que se utiliza es:

$$w = \frac{1}{ns^2}\left[\sum_{j=1}^{h} a_{j,n}\left(x_{(n-j+1)} - x_{(j)}\right)\right]^2 = \frac{A^2}{ns^2}.$$

donde $ns^2 = \sum (x_i - \bar{x})^2$; h es la parte entera de $n/2$; los coeficientes $a_{j,n}$ están tabulados y $x_{(j)}$ es el valor ordenado en la muestra que ocupa el lugar j. La distribución de w está tabulada y se rechazará normalidad cuando el valor calculado es menor que el valor crítico dado en las tablas.

4.2.4. Contrastes de asimetría y curtosis

El coeficiente de asimetría muestral viene dado por la expresión: $\alpha_1 = \frac{\sum (x_i - \bar{x})^3}{ns^3}$ donde s es la desviación típica de los datos. Si la hipótesis de normalidad es cierta, entonces la población es simétrica y el coeficiente de asimetría debe acercarse a cero. Para muestras grandes (de al menos 50 datos) la distribución de α_1 es aproximadamente normal con media 0 y varianza aproximadamente igual a $6/n$. Esta propiedad nos permite contrastar la hipótesis de que los datos proceden de una distribución simétrica.

El grado de apuntamiento o curtosis -concentración de probabilidad cerca de la moda- se mide por el coeficiente: $\alpha_2 = \frac{\sum (x_i - \bar{x})^4}{ns^4}$. Este coeficiente toma el valor 3 para una distribución normal. Para muestras grandes -más de 200 datos- este coeficiente se distribuye asintóticamente como una distribución normal con media 3 (valor teórico del coeficiente de curtosis bajo distribución normal) y varianza aproximadamente igual a $24/n$. De nuevo esta propiedad nos permite contrastar la hipótesis de que los datos presentan un apuntamiento similar al de la distribución normal.

También se puede elaborar un test conjunto para las dos características sin más que estandarizar cada una de las distribuciones y sumarlas. En este caso, se obtiene la siguiente expresión: $X_2^2 = \frac{n\alpha_1^2}{6} + \frac{n(\alpha_2-3)^2}{24} \sim \chi_2^2$ que puesto que resulta la suma de dos N(0,1) independientes sigue una distribución χ^2 con dos grados de libertad. Este test se conoce habitualmente con el nombre de Jarque-Bera.

4.2.5. Transformaciones para conseguir normalidad

Una posible solución cuando no hemos conseguido aceptar la hipótesis de que los datos son normales es transformar estos para que su distribución se parezca lo más posible a la normal. La siguiente familia de transformaciones para conseguir normalidad ha sido propuesta por Box y Cox:

$$x^{(\lambda)} = \begin{cases} \frac{(x+m)^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln(x+m) & \text{si } \lambda = 0, \end{cases} \qquad \text{siendo } x + m > 0.$$

donde λ es el parámetro de la transformación que sería necesario estimar y la constante m se elige de tal forma que $x + m$ sea siempre positivo. La estimación de λ se suele hacer por máxima-verosimilitud que en este caso sería obtener el máximo de la siguiente función:

$$\ell(\lambda) = -\frac{n}{2} \ln \hat{\sigma}(\lambda)^2 + (\lambda - 1) \sum_{i=1}^{n} \ln(x_i),$$

donde $\hat{\sigma}(\lambda)^2$ representa la varianza de los datos transformados. Usualmente se suele obtener el valor de esta función para valores entre -2 y 2.

4.3. Contrastes de posición

Se trata en este apartado de verificar si dos poblaciones podrían ser homogéneas en cuanto a su posición. Básicamente estamos en la misma situación que el contraste de igualdad de medias vista en el apartado de inferencia paramétrica salvo que ahora la posición no se medirá con la media de la población sino con otras medidas. Como en aquel caso diferenciaremos los casos de muestras independientes y muestras apareadas.

4.3.1. Test de los signos y rangos para muestras apareadas

Supongamos que tenemos una muestra $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ tomada sobre n individuos donde suponemos una cierta dependencia entre ellas (debido al efecto individuo). El contraste que pretendemos es H_0 : "las dos muestras provienen de la misma población" contra su alternativa. Si es cierta H_0 la distribución de la variable $X-Y$ será simétrica respecto al origen y por tanto la $P(X>Y)=P(Y>X)=0{,}5$. Por tanto el número de veces que se espera que la variable X-Y tome valores positivos debe seguir una distribución binomial de parámetros n y $p=0{,}5$. Para $n>10$ se puede aproximar la binomial por una normal de parámetros $\mu=n/2$ y $\sigma^2=n/4$. Por tanto si llamamos ns al número de veces que ocurre $X-Y>0$ entonces

$$Z=\frac{ns-n/2}{\sqrt{n/4}}\sim N(0,1)$$

y por tanto las regiones de aceptación serán $|Z|\leq z_{\alpha/2}$, $Z\leq z_\alpha$ y $Z\leq -z_\alpha$.

También se utiliza para comparar datos apareados el test del signo-rango de Wilcoxon que además del signo se utiliza la magnitud de las diferencias entre los valores de cada par. Tomaremos las diferencias $d_i=x_i-y_i$ y ordenarlas de menor a mayor prescindiendo del signo y asignarles rango desde 1, 2, ...,n. Consideramos el estadístico $T=\text{mín}(T+,T-)$ siendo $T+$ la suma de los rangos correspondientes a las diferencias positivas y $T-$ la suma de los rangos correspondientes a las diferencias negativas. Se rechaza la hipótesis nula de que las dos muestras provienen de la misma población si T es igual o menor que el valor crítico de la tabla de Wilcoxon. Para $n>25$ se puede aproximar T (bajo la hipótesis nula) a una normal de media $\mu=n(n+1)/4$ y varianza $\sigma^2=n(n+1)(2n+1)/24$. Los valores iguales a cero se ignoran y si varias d_i son iguales se les asigna el rango promedio de los valores empatados.

4.3.2. Test de Mann-Whitney-Wilcoxon para muestras independientes

Supongamos que tenemos dos muestras $\{x_1,\ldots,x_n\}$ y $\{y_1,\ldots,y_m\}$ que creemos provenientes de la misma población (hipótesis nula). Si esto es cierto entonces como se razonó para el anterior test $P(X>Y)=P(Y>X)=0{,}5$. Dado que tenemos $n\times m$ pares el número esperado de pares donde se cumpla la condición $x<y$ será $n\times m/2$. Se conoce como el estadístico U de Mann-Whitney al número observado de pares con la anterior propiedad. Para calcularlo rápidamente se ordenan las dos muestras conjuntamente y se asignan rangos 1, 2, ..., $n+m$ de menor a mayor. Entonces $U=W-m(m+1)/2$ donde

W es la suma de los rangos correspondientes de la segunda muestra. (Simétricamente se puede razonar con la otra muestra). Si $n > 10$ y $m > 10$, entonces U se puede aproximar a una distribución normal de media $n \times m/2$ y varianza $n \times m(n+m+1)/12$. Los posibles empates en los rangos se resuelven asignando a las observaciones empatadas el promedio de los rangos empatados. En este caso se modifica levemente la varianza del estadístico aunque la diferencia es muy pequeña si el número de empates es pequeño.

4.3.3. Test de Kruskal-Wallis para múltiples muestras independientes

Supongamos que tenemos ahora k muestras de tamaño $n_1,\ldots,n_k$. Se ordenan conjuntamente y se asignan rangos a la muestra ordenada conjunta. Llamemos R_i a la suma de los rangos de los n_i elementos de la muestra $i-$ésima. Si existe homogeneidad entre las distribuciones de los k grupos y no hay observaciones repetidas, entonces el estadístico

$$H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(n+1) \sim \chi^2_{k-1}$$

si $n_i > 5$. Para valores pequeños de n es necesario consultar tablas especiales. Si existen observaciones repetidas se promedian los rangos repetidos como en el test de Mann-Whitney y se le asigna ese promedio como rango a todas las observaciones repetidas. En este último caso el estadístico H se corrige dividiendo por

$$1 - \sum_{i=1}^{empates} \left(t_i^3 - t_i\right) \Big/ \left(n^3 - n\right)$$

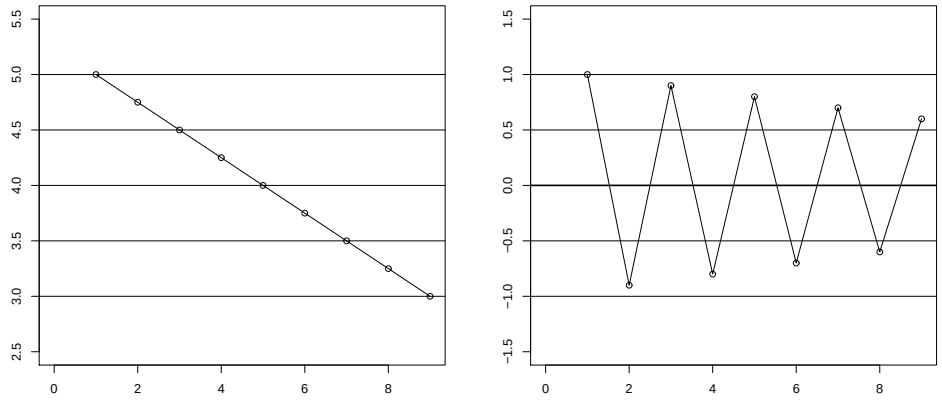
donde t_i es el número de observaciones que empatan en un rango dado.

4.4. Hipótesis de independencia

Cuando las observaciones son dependientes, las expresiones obtenidas para los distintos estimadores cambian radicalmente. Así, por ejemplo, la propiedad que asegura que la varianza de la media muestral de n observaciones N(μ,σ) independientes es σ^2/n es completamente falsa. De hecho, en función de cómo sea esa dependencia esta varianza puede ser mucho más grande o mucho más pequeña. Es por esto que la hipótesis de independencia se convierte entonces en una de las hipótesis más importantes a contrastar ya que de ella suele depender la fiabilidad de las estimaciones posteriores. En general, esta hipótesis debe contrastarse cuando los datos mediante su procedimiento de muestreo procedan de una secuencia temporal o espacial.

Lo primero que se debe hacer con un conjunto de valores es dibujarlo respecto a su orden.

En las figuras siguientes se muestran dos ejemplos de dos sucesiones de valores no independientes; la primera presenta una clara tendencia decreciente, la segunda muestra oscilaciones respecto a un valor fijo y cada dato tiene diferente signo que el anterior. En una secuencia independiente no debemos ver patrones reconocibles respecto al orden, mostrando el dibujo de la secuencia un comportamiento errático.



4.4.1. Contraste de rachas

Definamos ***racha*** como una secuencia de valores que cumplen una condición lógica. Por ejemplo, que estén por encima o por debajo de un valor o que formen una secuencia monótona (creciente o decreciente). Vamos a examinar distintos tipos de rachas para contrastar la independencia.

Sea una sucesión de observaciones X=(x_1, ..., x_n) para la cual construimos la siguiente sucesión z_i:

$$z_i = \begin{cases} 1 & \text{si } x_i < x_{i+1}, \\ 0 & \text{si } x_i > x_{i+1}. \end{cases}$$

Esta nueva sucesión mantiene la información de si estamos en una racha creciente (1) o decreciente (0). Además para contar el número de rachas en la sucesión sólo debemos contar cuantas veces se cambia de valor y sumarle 1. En el ejemplo de las sucesiones dibujadas anteriormente, la primera imagen corresponde al caso extremo donde sólo existe una racha y la segunda a cuando en cada dato se cambia la racha. Por tanto, un número corto o excesivo de rachas significa dependencia.

Asintóticamente, el número de rachas de una sucesión bajo independencia sigue la distribución:

$$\text{n}^{\text{o}} \text{ rachas} \sim N\left(\frac{2n-1}{3}, \sqrt{\frac{16n-29}{90}}\right).$$

También se puede establecer el número esperado de rachas de longitud k crecientes o decrecientes por la siguiente expresión:

$$E\left(RU_k + RD_k\right) = \frac{2\left\{\left(k^2+3k+1\right)n - \left(k^3+3k^2-k-4\right)\right\}}{(k-3)!}, \quad k \ \leq n-2,$$

$$E\left(RU_{n-1} + RD_{n-1}\right) = \frac{2}{n!}.$$

Aunque estas expresiones sólo sirven para realizar un ajuste visual (especialmente útil cuando el test rechaza la hipótesis de independencia).

Otro tipo de rachas que podemos controlar es por encima o debajo de la mediana. Es decir, definimos la siguiente secuencia:

$$s_i = \begin{cases} 1 & \text{si } x_i < \text{mediana}, \\ 0 & \text{si } x_i > \text{mediana}. \end{cases}$$

De nuevo, el número de rachas se cuenta como los cambios de valor de la secuencia s_i +1. Bajo independencia el número de rachas por encima o por debajo de la mediana sigue la siguiente distribución:

$$\text{n}^{\text{o}} \text{ rachas mediana} \sim N\left(s+1, \sqrt{\frac{s(s-1)}{2s-1}}\right),$$

donde s es el número de ceros que tiene la secuencia (igual al número de 1's). También se puede calcular el valor esperado de tener una racha de longitud k que tendría la siguiente expresión: $E\left(R_k\right) = (n-k+3) \times 2^{-(k+1)}$. De nuevo, esta última propiedad es útil cuando se rechaza la hipótesis de que la secuencia es independiente.

4.4.2. Contraste de autocorrelación

Partiendo de una muestra (x_1, ..., x_n) se define el coeficiente de autocorrelación lineal de orden uno $r(1)$ por:

$$r(1) = \frac{\sum_{i=2}^{n}\left(x_i - \bar{x}\right)\left(x_{i-1} - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Este coeficiente viene a ser aproximadamente el coeficiente de correlación lineal de las variables (x_2, ..., x_n) y (x_1, ..., x_{n-1}) y mide la correlación lineal entre las observaciones y sus observaciones precedentes. Análogamente podemos definir el coeficiente de autocorrelación lineal de orden k por:

$$r(k) = \frac{\sum_{i=k+1}^{n}\left(x_i - \bar{x}\right)\left(x_{i-k} - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Su interpretación es también análoga a la del coeficiente de orden uno.

Se denomina correlograma a la representación de estos coeficientes de autocorrelación lineal en función de k (normalmente llamado *retardo*). Cuando las observaciones son independientes y proceden de una distribución normal, los coeficientes de autocorrelación muestrales siguen aproximadamente una distribución normal con media cero y varianza $1/n$. Por lo tanto, podemos considerar significativamente distintos de cero aquellos coeficientes que no estén en el intervalo $(-1,96/\sqrt{n}, 1,96/\sqrt{n})$. Si n es grande ($n > 50$), podemos realizar un contraste conjunto de los primeros coeficientes de autocorrelación.

Si H_0 es la hipótesis de independencia, entonces cada $r(k)$ se distribuye aproximadamente según $N(0,\ 1/\sqrt{n})$ y, por lo tanto,

$$Q = \sum_{k=1}^{m} \left(\frac{r(k)}{1/\sqrt{n}} \right)^2 = n \sum_{k=1}^{m} (r(k))^2,$$

sigue, aproximadamente, una distribución χ^2 con $m-1$ grados de libertad (nótese que hay que estimar un parámetro: la media). Este contraste ha sido mejorado por Ljung y Box que han demostrado que una aproximación más exacta es considerar el estadístico

$$Q = n(n+2) \sum_{k=1}^{m} \frac{(r(k))^2}{n-k}$$

que, como el anterior, si H_0 es cierta, se distribuye aproximadamente como una χ^2 con $m-1$ grados de libertad.

4.4.3. Test de Durbin-Watson

El test se basa en calcular el estadístico

$$R = \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 \Big/ 2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

que cuando la correlación es positiva toma valores cercanos a cero y cuando la correlación es negativa toma valores cercanos a 2. Este estadístico está indicado para detectar autocorrelación de primer orden en el conjunto de datos. Si la muestra es normal y $n > 20$ entonces

$$\frac{R-1}{\sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}} \sim N(0,1),$$

lo que nos sirve para realizar el contraste.

4.5. Hipótesis sobre la homogeneidad

Diremos que una muestra es heterogénea cuando sus observaciones no provienen por el mismo modelo de distribución de probabilidad. Los contrastes de posición vistos anteriormente se puede considerar de homogeneidad dado que se está contrastando si la posición de las muestras es la misma o lo que es lo mismo si prescindiendo de la división entre grupos los datos pueden conformar una sola población homogénea. La heterogeneidad puede venir dada por muchos y diversos motivos. Por ejemplo, puede ser que la población sea heterogénea respecto a una posible agrupación de los datos o a una clasificación de variables cualitativas o que lo sea por la introducción de errores en el proceso de muestreo. En el segundo caso trataremos la homogeneidad en el sentido de ausencia de datos atípicos.

4.5.1. Test de homogeneidad en tablas de contingencia

Otro tipo de independencia es la que se puede obtener a partir de variables discretas. Diremos que dos variables cualitativas (o variables agrupadas) son independientes si la tabla de contingencia de las frecuencias observadas no presenta pautas de asociación entre alguna de las clases de estos atributos. En definitiva, se van a comparar las frecuencias observadas de la tabla de contingencia con las frecuencias esperadas bajo la hipótesis de independencia entre atributos. Sean entonces $A_1,\ldots,A_k$ las distintas clases de la primera variable y sean $B_1,\ldots,B_l$ las distintas clases de la segunda variable que constituyan una partición del espacio $\Omega_1 \times \Omega_2$. Estudiando el número de veces que se produce cada cruce entre la primera variable con la segunda (n_{ij}) tendremos una tabla de contingencia $k \times l$. Bajo la hipótesis de independencia el número esperado en cada celda es $e_{ij} = n \times n_{i\cdot} \times n_{\cdot j}/n^2$ donde $n_{i\cdot}$ y $n_{\cdot j}$ representan las frecuencias absolutas marginales. Por tanto, podemos construir un test similar al construido para bondad de ajuste mediante el estadístico:

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

que bajo la hipótesis de independencia tendrá una distribución χ^2 con $(k-1)\times(l-1)$ grados de libertad. Valores grandes de este estadístico favorecerán la hipótesis alternativa mientras que valores bajos favorecerán la hipótesis nula.

4.5.2. Test de valores atípicos

En la hipótesis de que los datos son normales, el test más usual para testear la presencia de datos atípicos es el conocido como test de Grubb que se basa en calcular el estadístico:

$$G = \text{máx}\left|\frac{x_i - \bar{x}}{\hat{s}}\right|$$

que se compara respecto a la tabla que se presenta a continuación:

n	3	4	5	6	7	8	9	10
$\alpha = 0,05$	1,15	1,48	1,78	1,89	2,02	2,13	2,21	2,29
n	11	12	13	14	15	20	25	30
$\alpha = 0,05$	2,36	2,41	2,46	2,51	2,55	2,71	2,82	2,91

Si se obtiene un valor G mayor que el de la tabla se supone que existe un dato atípico (al menos) y se procede eliminando de la muestra el dato que obtuvo el máximo del estadístico y siguiendo el proceso iterativamente. Usualmente este estadístico es útil para muestras pequeñas. Para muestras más grandes se suele utilizar el test de curtosis visto anteriormente que es capaz de descubrir la presencia simultánea de valores atípicos. Cuando se observa la presencia de datos atípicos el coeficiente de apuntamiento tiende a

hacerse significativamente más grande. El coeficiente de curtosis es bajo el supuesto de tener población normal N(3,24/n) y por tanto para testear la presencia de datos atípicos conviene realizar el test unilateral de que este coeficiente no es significativamente más grande.

4.6. Ejercicio resuelto

Ejercicio: El índice de masa corporal (IMC) se calcula dividiendo el peso en kg. entre la altura al cuadrado en metros. La siguiente tabla contiene las alturas, pesos e IMC de 10 hombres:

Altura	172	170	170	168	175	169	171	169	167	174
Peso	63	75	68	70	74	72	67	69	70	84
IMC	21,30	25,95	23,53	24,80	24,16	25,21	22,91	24,16	25,10	27,74

Determinar:

a) Si la distribución del índice de masa corporal es normal con media 24 y varianza 2,5.

b) ¿Existe algún dato atípico en la muestra?

c) ¿Son los datos independientes?

Solución:

a) Para resolver el primer apartado el único test apropiado es el de Kolmogorov-Smirnov ya que el test χ^2 de Pearson no se debe aplicar a menos de 25 datos y los otros tests de normalidad no contrastan específicamente los parámetros de la normal.

   Para ello, lo primero que debemos hacer el ordenar los datos, escribir en una segunda columna la función de distribución empírica y en la tercera columna la función de distribución de cada dato supuesto que sigue una N(24,$\sqrt{2,5}$). Para calcular esta última columna se debe estandarizar la variable y mirar en la tabla de la normal. El dato que está repetido (24,16) ocupa una fila ajustando el valor de la función de distribución empírica en el dato. A continuación, se añade otra columna con el valor del estadístico D_n en cada punto, siendo

$$D_n(h) = \text{máx}\left\{\left|F_n(x_{h-1}) - F(x_h)\right|, \left|F_n(x_h) - F(x_h)\right|\right\}.$$

   El estadístico de Kolmogorov-Smirnov será finalmente el máximo de esta columna

que compararemos con el valor crítico que aparece en las tablas.

Datos	$F_n(x)$	$F_0(x)$	$D_n(x_h)$
21,30	0,1	0,044	0,056
22,91	0,2	0,245	0,145
23,53	0,3	0,382	0,182
24,16	0,5	0,540	0,240
24,80	0,6	0,695	0,195
25,10	0,7	0,758	0,158
25,21	0,8	0,779	0,079
25,95	0,9	0,891	0,091
27,74	1,0	0,991	0,091

El máximo de la columna es 0,240 que comparado con el valor crítico de las tablas 0,410 concluye que se puede aceptar que los datos son normales con la media y varianza dada.

b) Para resolver el segundo apartado podemos aplicar el test de Grubb

$$G = \text{máx}\left|\frac{x_i - \overline{x}}{\widehat{s}}\right|,$$

que siempre obtiene su máximo en uno de los extremos de la muestra. Para aplicar el test necesitamos los datos de la media y la cuasidesviación típica que son respectivamente, 24,486 y 1,75. El test de Grubb en los extremos de la muestra resulta ser 1,92 y 1,96 que comparado con el valor crítico en las tablas (˜2,29) nos concluye que no hay datos atípicos en la muestra.

c) Para resolver el último apartado, el test más sencillo para calcular es el test de las rachas respecto a la mediana. Para ello simplemente tenemos que calcular la mediana para estos datos, que como son pares, debe ser la media de los dos centrales ordenados (mediana=(24,16+24,80)/2=24,48) y colocar un signo más o menos en la serie original en función de estar por encima o por debajo de la mediana. Calcular el número de rachas resulta simplemente contar en esa secuencia las veces que se produce cambio de signo+1. Esta secuencia se ve en la siguiente tabla y contando tenemos 8 rachas:

IMC	21,3	25,95	23,53	24,8	24,16	25,21	22,91	24,16	25,1	27,74
Signo	−	+	−	+	−	+	−	−	+	+

Debemos comparar ahora este valor con el valor crítico que aparece en las tablas. Ahora calcularíamos el intervalo de confianza para el número de rachas y concluriamos que intervalo de confianza estaría entre 6 y 12, y puesto que 8 está claramente dentro de ese intervalo aceptamos la hipótesis de independencia.

Capítulo 5

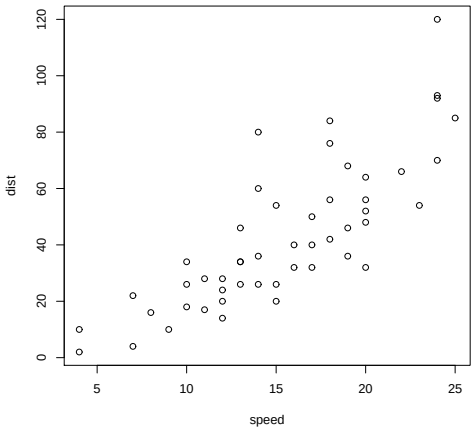
Modelos de regresión

5.1. Introducción

El término regresión procede de los estudios de Galton en biología a finales del siglo XIX. Galton estudió la dependencia de la estatura de los hijos (y) respecto a la de sus padres (x), encontrando lo que denominó una "regresión" a la media: los padres altos tienen en general hijos altos pero en promedio no tan altos como sus padres y los bajos tienen hijos bajos pero en promedio más altos que sus padres. El término en inglés que empleó, "regress", sirve desde entonces para nombrar todos aquellos modelos estadísticos que explican la dependencia de una variable (y) respecto a otra/s (x). El planteamiento de estos modelos se basa en calcular la esperanza condicionada de la variable y con respecto a la variable x, ya que esta esperanza condicionada (media condicionada) es la mejor predicción que podemos dar de la variable y conociendo la x. Si x no tiene información relevante de y, la esperanza condicionada de y respecto a x será simplemente la media de y.

El problema de regresión se concretará en obtener las relaciones entre las variables X e Y a partir de n pares de observaciones (x_1,y_1),..., (x_n,y_n). Como conjunto de ejemplo vamos a utilizar el conjunto *cars* disponible en el paquete base de R (*data(cars)*) que tiene dos variables *speed* y *dist* que son respectivamente velocidad del coche y distancia de frenado para pruebas hechas con varios modelos de coches.

La primera herramienta que utilizaremos para investigar la posible relación entre las variables será la conocida como nube de puntos que no es más que representar en un eje cartesiano estos pares de puntos. La nube de puntos de las variables indicadas aparece en la figura adjunta. De la observación de la nube de puntos podemos obtener la conclusión de que la velocidad tiene una cierta relación lineal con la distancia de frenado. Según crece la velocidad, crece la distancia

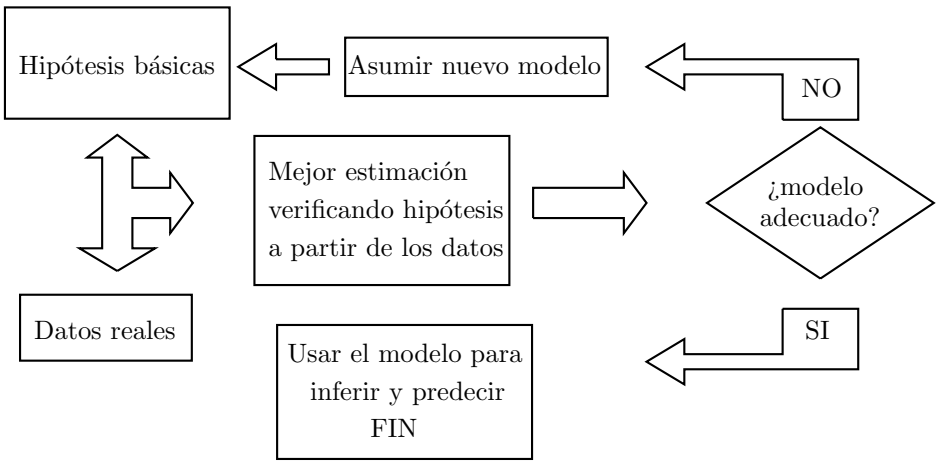


y de hecho si dibujamos una recta casi diagonal (entre (0,0) y (25,100)) de esquina inferior izquierda a la esquina superior derecha, esta podría ser el mejor ajuste que podríamos hacer de esta nube de puntos. Esta es precisamente la idea central de los modelos de regresión: encontrar una función f tal que, al representar $y = f(x)$, el ajuste con respecto a la nube de puntos sea bueno.

5.2. Planteamiento e hipótesis básicas

El planteamiento que se debe seguir para estimar un modelo de regresión puede verse en la siguiente figura.

Este proceso de decisión deja claro que el modelo de regresión va a ser función tanto de las hipótesis planteadas como de los datos que hemos medido. Esta doble dependencia hace que, por un lado, los estimadores sean variables aleatorias con una cierta distribución (por las hipótesis de partida) y, por otro, demos su valor más probable como estimación numérica (usando los datos reales).



5.2.1. Hipótesis básicas iniciales

1. Existencia.
   Para cualquier valor de x, y es una variable aleatoria que tiene media y varianza finita, es decir, la variable $y|x$ tiene momentos de orden 2 finitos. Esta propiedad nos asegura que lo que hacemos tiene sentido.

2. Independencia.
   Los valores de la variable y son estadísticamente independientes unos de otros.

3. Forma lineal . El valor medio de la variable $y|x$ es una función lineal de x. Es decir $E(y|x) = \beta_0 + \beta_1 x \Leftrightarrow y = \beta_0 + \beta_1 x + \varepsilon$ donde ε es una variable aleatoria de media (condicionada a x) cero.

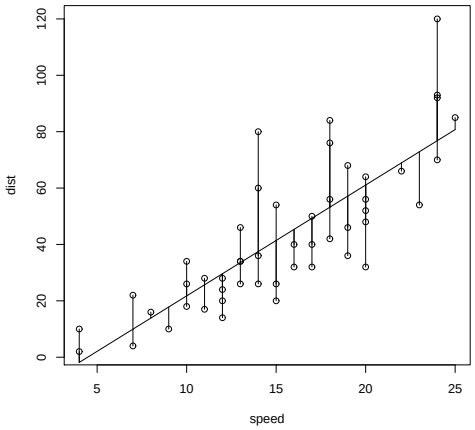
4. Homocedasticidad.
   La varianza de $y|x$ es constante para todo x, es decir $\sigma^2_{y|x} = \sigma^2$.

5. Normalidad. $\varepsilon \approx \mathrm{N}(0, \sigma)$.

Asumir las hipótesis anteriores simultáneamente nos lleva a que $y \approx \mathrm{N}(\beta_0 + \beta_1 x, \sigma)$.

Una vez determinadas las hipótesis que vamos a utilizar, suponemos que vamos a usar un conjunto de datos de tamaño n, que denotaremos por $\{x_i, y_i\}_{i=1,\ldots,n}$ y que son homogéneos respecto al estudio que se quiere realizar. Siguiendo el esquema antes planteado debemos unir las hipótesis con los datos para obtener la "mejor estimación". Ésta vendrá determinada por la medida de discrepancia entre el modelo teórico y los datos. La medida de discrepancia más clásica y utilizada es la que viene dada por $\sum e_i^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$ (mínimos cuadrados) y por tanto los mejores parámetros serán aquellos que hagan mínima esta cantidad.

En la figura adjunta se tiene el mismo ejemplo visto anteriormente donde se ha dibujado la mejor recta y los errores cometidos (las líneas verticales que van desde cada punto a la recta de regresión). Para este caso la mejor estimación es $\hat{\beta}_0 = -17{,}58$ y $\hat{\beta}_1 = 3{,}93$. Por supuesto, elegir otra medida de discrepancia nos llevaría a otros estimadores de los parámetros y a otra recta de regresión. A estos estimadores se les conoce como estimadores mínimo-cuadráticos.



5.3. Estimación

El procedimiento a seguir para calcular los estimadores mínimo-cuadráticos es, por tanto, encontrar aquellos que minimicen $\sum e_i^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$. Para minimizar respecto a los parámetros derivaremos respecto a ellos e igualamos a cero. Entonces nos queda:

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i); \quad \frac{\partial \sum e_i^2}{\partial \beta_1} = -2 \sum (y_i - \beta_0 - \beta_1 x_i)\, x_i.$$

De aquí se deduce que los estimadores deben cumplir las dos siguientes propiedades:

$$\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0; \quad \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0.$$

De la primera se deduce que $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ lo que significa que la recta de regresión siempre pasará por el punto $(\bar{x}, \bar{y})$.

Sustituyendo esta primera expresión en la segunda nos queda:

$$\sum \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \right) x_i = 0$$
$$\Rightarrow \sum \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \right) (x_i - \bar{x} + \bar{x}) = 0$$

Realizando los cálculos, obtenemos

$$\sum (y_i - \bar{y})(x_i - \bar{x}) + \underbrace{\bar{x} \sum (y_i - \bar{y})}_{=0} = \sum \left(\hat{\beta}_1 (x_i - \bar{x}) \right) (x_i - \bar{x}) + \underbrace{\bar{x} \sum \left(\hat{\beta}_1 (x_i - \bar{x}) \right)}_{=0},$$

donde tenemos términos iguales a cero por las propiedades de la media muestral. Si dividimos ahora las dos expresiones por n nos queda:

$$\frac{1}{n} \sum (y_i - \bar{y})(x_i - \bar{x}) = \hat{\beta}_1 \frac{1}{n} \sum (x_i - \bar{x})(x_i - \bar{x}) \Rightarrow S_{xy} = \hat{\beta}_1 S_x^2 \Rightarrow \hat{\beta}_1 = S_{xy} / S_x^2,$$

con lo que tenemos el estimador del parámetro pendiente. Para obtener el parámetro ordenada en el origen basta con sustituir en la primera expresión y resulta:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - S_{xy} \bar{x} / S_x^2.$$

Además, dado que estamos minimizando las distancias verticales de los puntos muestrales a la recta vemos que, en general, las rectas de regresión de $y|x$ y de $x|y$ no serán iguales. Se puede tener la tentación de razonar de la siguiente forma: "puesto que tengo la recta regresión $y|x$ despejo x y tengo la recta de regresión $x|y$". Esta afirmación es falsa ya que estamos minimizando distancias verticales y, por tanto, los mejores estimadores son aquellos que hacen menor esta distancia. La recta de regresión $x|y$ minimizaría distancias horizontales viendo de igual manera el gráfico que hemos visto. Si lo comprobamos haciendo cálculos (sólo para el parámetro pendiente) y despejamos en función de x la recta de regresión obtenida, tendríamos que el parámetro pendiente así despejado sería:

$$\hat{\beta}_1^{x|y} = \frac{1}{\hat{\beta}_1^{y|x}} = \frac{S_x^2}{S_{xy}}$$

y por otro lado, si hacemos la regresión como debe ser, tenemos que

$$\hat{\beta}_1^{x|y} = \frac{S_{xy}}{S_y^2}.$$

Igualando las dos expresiones tenemos que $\frac{S_x^2}{S_{xy}} \neq \frac{S_{xy}}{S_y^2}$, salvo en el caso en que la relación entre x e y sea perfecta, en cuyo caso $S_{xy}^2 = S_x^2 S_y^2$ y el coeficiente de correlación es exactamente igual a 1 ó -1.

5.3.1. Propiedades de los estimadores

El estimador de la pendiente de la recta puede escribirse de la siguiente forma:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{nS_x^2} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{nS_x^2}(y_i-\bar{y}) = \sum_{i=1}^{n} w_i^x(y_i-\bar{y}),$$

donde se ve claramente que tiene la forma de una suma de valores que sólo dependen de $w_i^x(y_i-\bar{y})$. De esta manera sabemos que si hemos supuesto que la variable y es normal (Hip. 5) entonces el estimador de la pendiente también será normal porque es combinación lineal de normales independientes. Por ello bastará simplemente con determinar su media y varianza.

$$E\left(\hat{\beta}_1\right) = E\left(\sum_{i=1}^{n} w_i^x(y_i-\bar{y})\right) = \sum_{i=1}^{n} w_i^x E\left((y_i-\bar{y})\right) \overset{hip,3}{=} \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{nS_x^2}\beta_1(x_i-\bar{x}) = \beta_1$$

y por tanto este estimador es insesgado. Análogamente si calculamos su varianza tenemos:

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left(\sum_{i=1}^{n} w_i^x(y_i-\bar{y})\right) \overset{hip,2}{=} \sum_{i=1}^{n}(w_i^x)^2\,\mathrm{Var}\left((y_i-\bar{y})\right)\\ &\overset{hip,4}{=} \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n^2S_x^4}\sigma^2 = \frac{\sigma^2}{nS_x^2}.\end{aligned}$$

Por tanto, de esta varianza podemos deducir que: i) será más pequeña cuantos más datos tengamos, ii) disminuirá con la varianza muestral de x y iii) aumentará proporcionalmente a la varianza original de los datos.

Haciendo lo mismo para el parámetro ordenada en el origen tenemos:

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x} = \frac{\sum y_i}{n} - \bar{x}\sum_{i=1}^{n} w_i^x y_i + \underbrace{\bar{x}\sum_{i=1}^{n} w_i^x \bar{y}}_{=0} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}w_i^x\right)y_i = \sum_{i=1}^{n} r_i^x y_i,$$

que también es normal por ser combinación lineal de normales independientes. Procediendo como en el caso anterior se calcula media y varianza y tenemos:

$$\begin{aligned}E\left(\hat{\beta}_0\right) &= E\left(\sum_{i=1}^{n} r_i^x y_i\right) = \sum_{i=1}^{n} r_i^x E\left(y_i\right) \overset{hip,3}{=} \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{nS_x^2}\right)(\beta_0+\beta_1 x_i)\\ &= \beta_0\sum\frac{1}{n} - \frac{\beta_0\bar{x}}{nS_x^2}\underbrace{\sum(x_i-\bar{x})}_{=0} + \beta_1\sum\frac{x_i}{n} - \frac{\beta_1\bar{x}}{nS_x^2}\underbrace{\sum(x_i-\bar{x})\,x_i}_{=nS_x^2}\\ &= \beta_0+\beta_1\bar{x}-\beta_1\bar{x} = \beta_0,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_0\right) &= \mathrm{Var}\left(\sum_{i=1}^{n} r_i^x y_i\right) \overset{hip,2}{=} \sum_{i=1}^{n} \left(r_i^x\right)^2 \mathrm{Var}\left(y_i\right) \overset{hip,4}{=} \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}\left(w_i^x\right)\right)^2 \sigma^2 \\ &= \sigma^2 \sum \left(\frac{1}{n^2} + \bar{x}^2\left(w_i^x\right)^2 - \frac{2\bar{x}\left(w_i^x\right)}{n}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right).\end{aligned}$$

En este caso, la varianza $\hat{\beta}_0$ es más peculiar ya que siempre es más grande que σ^2/n (el segundo sumando siempre es positivo) y solamente puede tomar este valor inferior cuando $\bar{x} = 0$. Fijémonos que si la media de x es cero entonces el estimador ordenada en el origen está estimando la media de y, y es bien conocida la propiedad de que la varianza de la media de n datos es σ^2/n.

También cabe preguntarse cuál sería la covarianza entre estos dos estimadores:

$$\begin{aligned}\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\sum_{i=1}^{n} r_i^x y_i, \sum_{i=1}^{n} w_i^x y_i\right) \overset{hip,2}{=} \sum_{i=1}^{n} r_i^x w_i^x \mathrm{Var}\left(y_i\right) \\ &\overset{hip,4}{=} \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i^x\right) w_i^x \sigma^2 \\ &= \sigma^2 \sum \left(\frac{(x_i - \bar{x})}{n^2 S_x^2} - \bar{x}\frac{(x_i - \bar{x})^2}{n^2 S_x^4}\right) = -\frac{\bar{x}\sigma^2}{nS_x^2}.\end{aligned}$$

De la expresión de la covarianza se deduce que los estimadores solamente serán independientes cuando tengamos que $\bar{x} = 0$. En este caso $\hat{\beta}_0 = \bar{y}$ que por supuesto no depende de la pendiente de la recta.

Una vez determinadas medias y varianzas de los estimadores y usando las hipótesis 2, 3, 4 y 5, podemos concluir que:

$$\hat{\beta}_0 \approx N\left(\beta_0, \sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)}\right) \quad \text{y} \quad \hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma}{\sqrt{n}S_x}\right),$$

lo que nos permitirá hacer inferencias sobre los parámetros.

En cualquiera de los dos casos tenemos una relación que involucra al estimador del parámetro con una distribución conocida en función de los parámetros teóricos del modelo. Para construir un intervalo de confianza al $(1-\alpha)\,\%$ se utilizan las relaciones siguientes:

$$\hat{\beta}_i \approx N\left(\beta_i, \sigma\left(\beta_i\right)\right) \Rightarrow \hat{\beta}_i - \beta_i \approx N\left(0, \sigma\left(\beta_i\right)\right) \Rightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma\left(\beta_i\right)} \approx N\left(0,1\right).$$

La varianza del estimador depende directamente de la varianza del error que, en general, será desconocida y por tanto necesitaremos estimarla. El estimador natural de la varianza del error podría ser simplemente la varianza muestral del error, es decir, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$. Sin embargo, si calculamos la esperanza de este estimador tenemos:

$$\begin{aligned} E\left(\hat{\sigma}^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\right) = E\left(\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\hat{e}_i^2}{\sigma^2}\right) = \frac{\sigma^2}{n}E\left(\sum_{i=1}^{n}\frac{\hat{e}_i^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n}E\left(\chi^2_{n-2}\right) = \frac{\sigma^2}{n}\left(n-2\right), \end{aligned}$$

donde el sumatorio que es de normales(0,1) (hip.5) e independientes (hip.2) es una χ^2 con tantos grados de libertad como sumandos independientes. Como se han consumido 2 grados de libertad correspondientes a los parámetros β, (las dos restricciones para minimizar la suma de cuadrados) la χ^2 tiene $n-2$ grados de libertad. Así tenemos que la esperanza de ese estimador NO es el parámetro teórico. Realmente lo único que debemos hacer para subsanar este inconveniente es ajustar los grados de libertad del estimador y por eso se suele considerar como mejor estimador de la varianza del error a

$$\hat{S}_R^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2,$$

cuya única diferencia es que ahora se divide por los grados de libertad adecuados. Así calculando ahora la esperanza de este estimador se tiene que: $E\left(\hat{S}_R^2\right) = \sigma^2$ y además

$$\frac{\left(n-2\right)\hat{S}_R^2}{\sigma^2} \approx \chi^2_{n-2}.$$

Esta última propiedad que nos servirá para hacer inferencias sobre la varianza del error.

Las propiedades más importantes de los estimadores insesgados que hemos visto se resumen en la siguiente tabla:

Parámetro	Estimador	Varianza	Distribución
β_0	$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$ (3)	$\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)$ (2), (4)	Normal (5)
β_1	$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$ (3)	$\frac{\sigma^2}{nS_x^2}$ (2), (4)	Normal (5)
σ^2	$\hat{S}_R^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$ (2), (3), (4)	$\frac{2}{n-2}\sigma^2$	χ^2_{n-2}

Si aplicamos estas propiedades al ejemplo que hemos visto tenemos:

	Estimador	D.T.	t	Sig.	Lím. inferior 95 %	Lím. superior 95 %
β_0	−17,5791	6,7584	−2,601	,0123	−31,168	−3,990
β_1	3,9324	0,4155	9,464	,000	3,105	4,759

La columna t refleja el estadístico pivote de suponer que el parámetro es cero y la columna de Sig. la probabilidad de que bajo la hipótesis nula se produzca un valor más

alto que el estadístico pivote (significación). Va a ser equivalente que dicha significación sea inferior a 0,05 y que el correspondiente intervalo de confianza al 95 % no incluya al cero. De la tabla podemos deducir que ninguno de los parámetros (a la vista de los datos) puede ser igual a cero. Particularmente es importante el correspondiente a β_1 porque este parámetro refleja la importancia que la variable x tiene sobre la predicción de y.

5.4. Contrastes de regresión y de las hipótesis

En este apartado analizaremos herramientas específicas para el análisis del modelo de regresión. Suponemos ahora que hemos estimado un modelo de regresión y trataremos de comprobar si este modelo que hemos estimado es consistente con las hipótesis que nos hemos planteado. El primer objetivo será verificar si el modelo propuesto no es trivial, es decir, si la variable x aporta información relevante sobre y. La forma de hacerlo es a través del procedimiento conocido como análisis de la varianza, viendo la descomposición de la variabilidad total, es decir:

$$\begin{aligned}\mathrm{VT} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \\ &+ \sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = \mathrm{VNE}+\mathrm{VE}+\sum_{i=1}^{n}\hat{e}_i(\hat{y}_i-\bar{y}) = \mathrm{VNE}+\mathrm{VE}\end{aligned}$$

Vemos que la variabilidad total se puede descomponer en dos sumandos positivos (son funciones al cuadrado): por un lado tenemos la variabilidad explicada (VE), que depende de la diferencia que resulta de explicar y con la información de x a no tener dicha información y por otro lado tenemos la variabilidad residual (VR) que representa lo que x no puede explicar de y. Cuanto más grande es uno, más pequeño debe ser otro. Esto significa que cuanto más explica x, de y menos queda por explicar. Así, la variabilidad total funciona como una gran tarta que se puede repartir en función de la información que se tenga de la variable a través de sus explicaciones. Si dividimos toda la expresión anterior por VT tenemos $1 = {}^{\mathrm{VR}}\!/_{VT} + {}^{\mathrm{VE}}\!/_{\mathrm{VT}}$ que, llamando coeficiente de determinación a $r^2 = {}^{\mathrm{VE}}\!/_{\mathrm{VT}}$, equivale a $1-r^2 = {}^{\mathrm{VR}}\!/_{\mathrm{VT}}$. El coeficiente de determinación tiene una interpretación sencilla ya que es la proporción de variabilidad total explicada y es un número entre 0 y 1. Además, en este caso es el cuadrado del conocido coeficiente de correlación $r = {}^{S_{xy}}\!/_{S_x S_y}$. En efecto,

$$\begin{aligned} r^2 &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \Big/ \sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta_1}^2(x_i-\bar{x})^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= \hat{\beta_1}^2 S_x^2/S_y^2 = S_{x,y}^2/S_x^2 S_y^2. \end{aligned}$$

De la descomposición obtenida se puede calcular un contraste de regresión de la manera que sigue: por un lado tenemos que $\mathrm{VR} = (n-2)\hat{S}_R^2$ es, por construcción, como hemos

visto antes, una χ^2 con $(n-2)$ grados de libertad. Sin embargo, el sumando VE sólo será una χ^2 con 1 grado de libertad si el modelo no explica, es decir, si la mejor estimación que podemos hacer de y_i es la media de y. Por tanto,

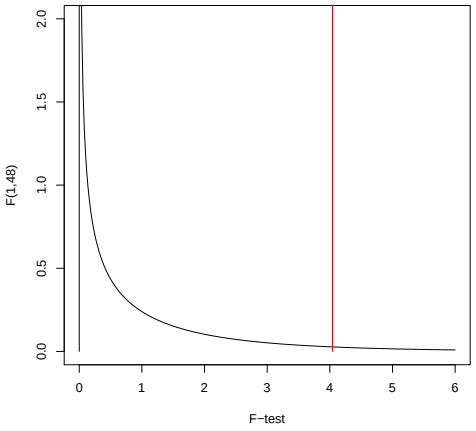
$$\frac{\text{VE}/1}{\text{VR}/n-2} = \frac{\text{VE}}{\hat{S}_R^2}$$

seguirá una distribución F con 1 y $n-2$ grados de libertad solamente cuando el numerador sea una χ^2 con 1 grado de libertad. Además, cuando nos alejemos de esta hipótesis y por tanto la variable x explique apreciablemente a la y, la cantidad VE tenderá a ser más grande de lo que debiera (y VR más pequeña de lo que debiera) y forzará a este último término a tomar valores grandes.

Por el contrario valores pequeños de VE favorecen la hipótesis de que la regresión planteada es inútil ya que la parte explicada no es suficiente.

Con todo esto en mente, para decidir si un modelo de regresión es o no útil se plantea lo siguiente: valores altos de ese cociente favorecen a la hipótesis alternativa (la regresión explica), valores bajos favorecen a la hipótesis nula (la regresión es inútil). Como siempre tomando el nivel de confianza del $(1-\alpha)\,\%$ (normalmente el $95\,\%$) determinaremos aquel punto a partir del cual rechazaremos la hipótesis nula como el cuantil $(1-\alpha)\,\%$ de la distribución de la F.

El punto para un caso de una F con 1 y 48 grados de libertad y confianza del $95\,\%$ (se corresponde con el ejemplo que venimos tratando) es 4,04.



Para este caso $\text{VE}\Big/\hat{S}_R^2 = 89{,}57$ que al ser mucho mayor que 4.04 rechaza claramente la hipótesis nula de que la variable regresora no tiene información de la variable dependiente. Lo habitual es escribir este contraste en forma de tabla ANOVA siguiendo las directrices habituales en el diseño de experimentos. La tabla ANOVA de regresión quedaría entonces conformada de la siguiente manera:

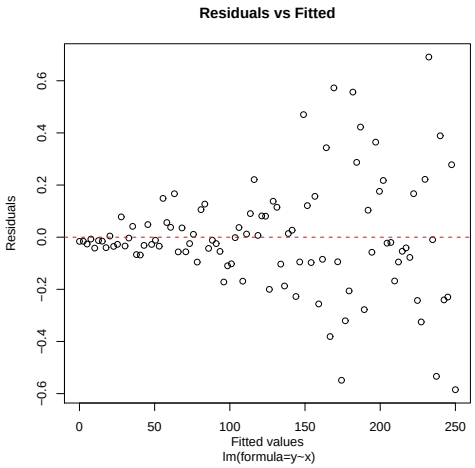
Fuente de variación	Suma de cuadrados	g.l.	Varianzas	Cociente	Crítico
Variación explicada	$\text{VE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$\hat{S}_e^2 = \text{VE}/1$	$\text{VE}\Big/\hat{S}_R^2$	$F_{1,n-2}$
Variación no explicada	$\text{VNE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$\hat{S}_R^2 = \text{VNE}/n-2$		
Variación total	$\text{VT} = \text{VNE} + \text{VE}$	$n-1$	$\hat{S}_Y^2 = \text{VT}/n-1$		

El trabajo que queda por delante es la comprobación de las hipótesis del modelo de regresión mediante el análisis de los residuos de la regresión. Si las hipótesis planteadas son ciertas la distribución de los residuos es normal de media cero y varianza $\hat{S}_R^2$. Esta

varianza debe ser constante y no depender de x ni de ninguna otra causa asignable. Además deben ser homogéneos en el sentido de que no se deben apreciar datos atípicos y deben ser independientes. También del estudio de los residuos podemos obtener conclusiones acerca de la calidad del modelo lineal.

Para contrastar normalidad e independencia hemos visto tests en el capítulo anterior que se pueden aplicar ahora para contrastar estas hipótesis. La hipótesis de homocedasticidad se puede detectar con un gráfico de los residuos frente a los valores previstos.

En el gráfico mostrado a la derecha se observa claramente como a medida que aumenta el valor ajustado la nube de puntos se hace más ancha evidenciando mayor variabilidad. Este sería un caso claro de heterocedasticidad. Si la hipótesis de homocedasticidad es cierta el gráfico debería mostrar una nube de puntos equidistante del centro (el cero) y de igual anchura en cualquier punto. También en este gráfico no deben encontrarse pautas respecto a la línea central. Si esto ocurriera tendríamos evidencia de que necesitamos un modelo más complejo para explicar los datos.



Para contrastar la presencia de datos atípicos, aunque se pueden utilizar los test vistos en el capítulo anterior, en este apartado se mirará con más cuidado los valores atípicos que resulten de la relación existente entre x e y. Es decir, aparte de que los datos sean atípicos por ser de una u otra población nos fijaremos especialmente en aquellos casos que resulten atípicos por la relación lineal establecida entre las variables. Hablaremos entonces de datos influyentes como aquellos que tienen excesiva importancia en la regresión y que podrían ser candidatos a ser datos atípicos. Veamos como analizar las observaciones influyentes. Dado que

$$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) = \sum_{j=1}^{n} \frac{y_j}{n} + \sum_{j=1}^{n} w_j^x y_j (x_i - \bar{x}) = \sum_{j=1}^{n} h_{ij} y_j$$

donde $(h_{ij}) = H$ es la matriz cuyos elementos son

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})}$$

y que medirían la influencia que ejerce la observación j-ésima en la predicción de la i-ésima. En particular, los elementos de la diagonal de esta matriz h_{ii} cumplen la propiedad de que son positivos y su suma es 2. Por tanto, la influencia media de cada observación debería ser $2/n$. Estos elementos de la diagonal serían la influencia de cada dato en la predicción de si mismo considerándose en la práctica que un valor es influyente si

$h_{ii} > 4/n$. También para detectar datos atípicos se puede analizar distintas funciones relacionadas con los residuos o con distancias relevantes como por ejemplo:

- Residuos: $\hat{e}_i = y_i - \hat{y}_i = (1 - h_{ii})y_i - \sum_{j \neq i} h_{ij} y_j \sim \mathrm{N}(0, S_R)$.
- Residuos estandarizados: $z_i = \frac{\hat{e}_i}{S_R \sqrt{1-h_{ii}}} \sim \mathrm{N}(0,1)$.
- Residuos estudentizados: $t_i = \frac{\hat{e}_i}{S_R^{(i)} \sqrt{1-h_{ii}}} \sim t_{n-3}$.
- Residuos eliminados: $\hat{e}_i^{(i)} = y_i - \hat{y}_i^{(i)}$.
- Residuos eliminados estudentizados: $t_i^{(i)} = \frac{\hat{e}_i^{(i)}}{S_R^{(i)}}$.
- Estadístico D de Cook: $D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_j^{(i)}\right)^2}{2S_R^2}$.
- Distancia de Mahalanobis: $M_i = \frac{(x_i - \bar{x})^2}{S_x^2}$.

5.5. Predicción

Hablaremos de dos tipos de predicción: estimar las medias de las distribuciones de y condicionadas para cada valor de la variable x y prever futuros valores de la variable respuesta o dependiente. Ambas predicciones se obtienen sustituyendo en la recta el valor de x y calculando el valor previsto. Sin embargo, veremos que la precisión de las dos estimaciones es diferente.

5.5.1. Predicción de la media condicionada a x

Estamos prediciendo la media condicionada (m_h) a un valor de la variable $x = x_h$. Como ya hemos mencionado el mejor estimador de la media es $\hat{y}_h = \bar{y} + \hat{\beta}_1(x_h - \bar{x})$ que cumple:

- $E(\hat{y}_h) = E\left(\bar{y} + \hat{\beta}_1(x_h - \bar{x})\right) = m_h$.
- $\mathrm{Var}(\hat{y}_h) = \mathrm{Var}\left(\bar{y} + \hat{\beta}_1(x_h - \bar{x})\right) = \frac{\sigma^2}{n} + \frac{\sigma^2(x_h - \bar{x})^2}{nS_x^2} = \frac{\sigma^2}{n}\left(1 + \frac{(x_h - \bar{x})^2}{S_x^2}\right) = \frac{\sigma^2}{n_h}$, donde

$$n_h = \frac{n}{1 + \frac{(x_h - \bar{x})^2}{S_x^2}}.$$

Al valor n_h se le llama número equivalente de observaciones. Por tanto y procediendo como en apartados anteriores

$$\frac{\hat{y}_h - m_h}{\hat{S}_R \Big/ \sqrt{n_h}} \sim t_{n-2},$$

lo que nos servirá para hacer intervalos de confianza para la media. Dado que n_h decrece con la distancia respecto a la media de la variable x el intervalo de confianza crecerá a medida que nos alejamos de la media.

5.5.2. Predicción de una nueva observación condicionada a x

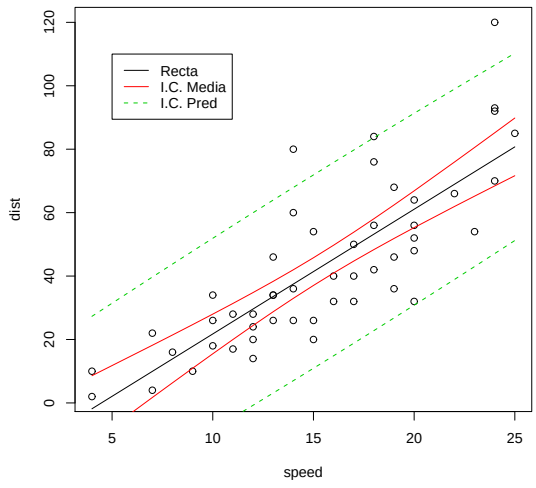
Como ya hemos mencionado, la predicción de una nueva observación se realiza también a través del estimador $\hat{y}_h = \bar{y} + \hat{\beta}_1(x_h - \bar{x})$. La motivación es que, como en el caso de la media condicionada, a falta de otra información la media es la mejor predicción para una población. Sin embargo, lo que resulta diferente es la precisión respecto a una nueva observación. En caso de predecir una nueva observación la varianza viene dada por:

$$\begin{aligned} E\left((y_h - \hat{y}_h)^2\right) &= E\left((y_h - m_h + m_h - \hat{y}_h)^2\right) = E\left((y_h - m_h)^2\right) + E\left((\hat{y}_h - m_h)^2\right) \\ &+ 2E\left((y_h - m_h)(\hat{y}_h - m_h)\right)_{=0} = \mathrm{Var}(y_h) + \mathrm{Var}(\hat{y}_h) = \sigma^2\left[1 + \frac{1}{n_h}\right]. \end{aligned}$$

Procediendo como en el apartado anterior, el intervalo de confianza para la predicción de una nueva observación vendrá dado por

$$\frac{\hat{y}_h - y_h}{\hat{S}_R\Big/\sqrt{1 + \frac{1}{n_h}}} \sim t_{n-2}.$$

En la siguiente figura se muestra para el ejemplo descrito los correspondientes intervalos de confianza para la media y para la predicción.



5.6. Ejercicio resuelto

Ejercicio: La siguiente tabla contiene las alturas y pesos de 10 hombres:

Altura	172	170	170	168	175	169	171	169	167	174
Peso	63	75	68	70	74	72	67	69	70	84

Se pide:

a) Dar la mejor predicción del peso de un hombre si su altura es 170 cm.

b) ¿Qué porcentaje de la variabilidad del peso es explicado por la altura?. ¿Influye la altura en la información que tenemos del peso?

c) Contrastar si el parámetro pendiente de la recta es 1.

d) Determinar el intervalo de confianza para una nueva observación con altura igual a 170 cm.

Solución:

a) Denotaremos por a la variable Altura y por p la variable Peso. Para resolver el primer apartado, necesitamos los datos básicos relevantes que son $n = 10$, $\overline{a} = 170{,}5$, $\overline{p} = 71{,}2$, $s_a^2 = 5{,}85$, $s_p^2 = 28{,}96$ y $s_{ap} = 5{,}2$. Resolviendo la recta de regresión tenemos:

$$\widehat{\beta}_1 = \frac{s_{ap}}{s_a^2} = \frac{5{,}2}{5{,}85} = 0{,}8888,$$
$$\widehat{\beta}_0 = \overline{p} - \widehat{\beta}_1 \overline{a} = 71{,}2 - 0{,}8888 \times 170{,}5 = -80{,}355.$$

Por lo tanto, la predicción para un hombre de 170 cm. de altura es

$$\widehat{y}_h = -80{,}355 + 0{,}8888 \times 170 = 70{,}755.$$

b) Para resolver el segundo apartado, el porcentaje de variabilidad viene dado por el coeficiente de determinación:

$$r^2 = \frac{s_{ap}^2}{s_a^2 s_p^2} = 0{,}1596,$$

y por tanto el 15,96 % de la variabilidad del peso es explicado por la altura. Para decidir si influye la altura en el peso debemos usar el estadístico $F = \frac{\widehat{s}_\epsilon^2}{\widehat{s}_R^2}$, que bajo la hipótesis de que la altura no influye en el peso sigue una distribución $F_{1,n-2}$. Por tanto, debemos calcular ambos términos del estadístico. Para calcular el numerador usaremos la siguiente fórmula:

$$VE = r^2 VT = r^2 n s_p^2 = 0{,}1596 \times 10 \times 28{,}96 = 46{,}22.$$

El denominador de ese estadístico es:

$$\widehat{s}_R^2 = \frac{VNE}{n-2} = \frac{\left(1-r^2\right) n s_p^2}{n-2} = \frac{0{,}8404 \times 10 \times 28{,}96}{8} = 30{,}42.$$

Por tanto, el valor del estadístico F es:

$$F = \frac{\widehat{s}_\epsilon^2}{\widehat{s}_R^2} = \frac{46{,}22}{30{,}42} = 1{,}52,$$

que debemos comparar contra el valor crítico de una $F_{1,8}$ que es 5,31. Al ser más grande permite aceptar la hipótesis de que la altura no tiene suficiente información sobre el peso para esta muestra.

c) Para resolver el tercer apartado, para contrastar si el parámetro pendiente puede valer 1 usamos el estadístico:

$$\frac{\widehat{\beta}_1 - \beta_1}{\frac{\widehat{s}_R}{\sqrt{n} s_a}} \sim t_{n-2},$$

de donde despejando tenemos:

$$\beta_1 \in \left[\widehat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\widehat{s}_R}{\sqrt{n} s_a}\right],$$

y si 1 pertenece a este intervalo de confianza entonces podremos aceptar la hipótesis de que el 1 puede ser el parámetro de la recta. Para calcular este intervalo necesito antes calcular:

$$\widehat{s}_R = \sqrt{30{,}42} = 5{,}52,$$

y así el intervalo es:

$$\beta_1 \in \left[0{,}8888 \pm 2{,}31 \frac{5{,}52}{\sqrt{10} \cdot 2{,}4187}\right] = \left[0{,}8888 \pm 1{,}6658\right] = \left[-0{,}777, 2{,}554\right],$$

que claramente incluye al 1 lo que supone la aceptación de la hipótesis.

d) Para resolver el último apartado, usaremos el estadístico:

$$\frac{\widehat{y}_h - y_h}{\widehat{s}_R \sqrt{1 + \frac{1}{n_h}}} \sim t_{n-2},$$

que despejando nos queda,

$$y_h \in \left[\widehat{y}_h \pm t_{n-2,\alpha/2} \widehat{s}_R \sqrt{1 + \frac{1}{n_h}}\right],$$

donde,

$$n_h = \frac{n}{1 + \frac{(x_h - \overline{a})^2}{s_a^2}} = \frac{10}{1 + \frac{(170 - 170{,}5)^2}{5{,}85}} = 9{,}59,$$

y $\widehat{y}_h = -80{,}355 + 0{,}8888 \times 170 = 70{,}755$. Por lo tanto el intervalo para una nueva observación con altura 170 cm es:

$$y_h \in \left[70{,}755 \pm 2{,}31 \times 5{,}516 \times \sqrt{1 + \frac{1}{9{,}59}}\right] = [70{,}755 \pm 13{,}39] = [57{,}36, 84{,}14]\,.$$