Modelos Estadísticos para la Regresión y la Clasificación
Estadística Descriptiva
Clase 1: Estadística Descriptiva

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

7 de agosto de 2024

Estadística Descriptiva

La estadística se utiliza en todos los sectores de la actividad humana: encuestas, fiabilidad, control de calidad, ámbito médico, ecología... Siempre que tengamos datos podemos recurrir a la estadística. Estas han experimentado un gran auge con la llegada de los ordenadores de alto rendimiento y de nuevos métodos de modelización.

Estadística Descriptiva

La estadística se utiliza en todos los sectores de la actividad humana: encuestas, fiabilidad, control de calidad, ámbito médico, ecología... Siempre que tengamos datos podemos recurrir a la estadística. Estas han experimentado un gran auge con la llegada de los ordenadores de alto rendimiento y de nuevos métodos de modelización.
Existen dos ramas de la estadística:

Estadística Descriptiva

La estadística se utiliza en todos los sectores de la actividad humana: encuestas, fiabilidad, control de calidad, ámbito médico, ecología... Siempre que tengamos datos podemos recurrir a la estadística. Estas han experimentado un gran auge con la llegada de los ordenadores de alto rendimiento y de nuevos métodos de modelización.
Existen dos ramas de la estadística:

1. La estadística exploratoria.
   Consiste esencialmente en describir, representar, estructurar, sintetizar y resumir datos con distintas herramientas y representaciones gráficas. En esta rama se habla de estadística descriptiva, análisis de datos y clasificación.

Estadística Descriptiva

La estadística se utiliza en todos los sectores de la actividad humana: encuestas, fiabilidad, control de calidad, ámbito médico, ecología... Siempre que tengamos datos podemos recurrir a la estadística. Estas han experimentado un gran auge con la llegada de los ordenadores de alto rendimiento y de nuevos métodos de modelización.
Existen dos ramas de la estadística:

1. La estadística exploratoria.
   Consiste esencialmente en describir, representar, estructurar, sintetizar y resumir datos con distintas herramientas y representaciones gráficas. En esta rama se habla de estadística descriptiva, análisis de datos y clasificación.
2. La estadística inferencial.
   Consiste en comprender el origen de los datos, asociar un modelo a los datos y tomar decisiones. Esencialmente hablamos aquí de estimación, comprobación de hipótesis, pruebas, modelización y previsión. En general, el objetivo es extender las propiedades observadas en una muestra a toda la población. El cálculo de probabilidades desempeña un papel fundamental.

Conceptos básicos

- Población, muestra, observación
- Variable: característica de un individuo de la población. Puede tener dos naturaleza:
  - Variable Cuantitativa
    - continua: altura, peso, salario, precio, etc.
    - discreta: cantidad de aviones que llegan en un aeropuerto por hora, cantidad de clientes satisfechos, cantidad de éxitos,...
  - Variable Cualitativa
    - ordinal: nivel de satisfacción, nivel de alerta, etc.
    - nominal: color de pelo, sexo, preferencia política,etc
  - Modalidad: valor que toma una variable cualitativa
  - Estadístico: cantidad calculada sobre un conjunto de datos
  - Serie estadística: conjunto de observaciones de una o varias variables.

Conjuntos de datos

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
```

Figura: Datos iris

```
> airquality
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
4      18     313 11.5   62     5   4
5      NA      NA 14.3   56     5   5
```

Figura: Datos airquality

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Figura: Datos iris

```
> summary(airquality)
     Ozone           Solar.R          Wind             Temp           Month            Day
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0
 NA's   :37       NA's   :7
```

Figura: Datos airquality

Serie estadística cualitativa

Disponemos de $n = 96$ observaciones de una variable cualitativa X que representa el nivel de satisfacción de clientes en una escala del 1 al 5.

3	1	5	1	1	2	4	2	5	1	5	5	2	1	2	2
2	2	2	3	2	2	2	4	3	1	4	3	2	1	1	3
4	1	4	4	2	4	1	2	3	4	3	4	2	1	4	5
2	1	3	1	2	1	3	1	2	2	3	2	5	1	2	4
1	4	2	1	3	1	1	2	4	1	5	1	3	2	2	1
2	3	2	2	2	5	3	2	2	2	5	1	5	3	4	1

un resumen estadístico de esta serie:

	1	2	3	4	5
Cantidades	25	32	15	14	10
Proporciones	0.26	0.33	0.16	0,15	0,1
Cantidades cumuladas	25	57	72	86	96
Proporciones cumuladas	0.26	0.59	0.75	0.9	1

Las proporciones de cada modalidad constituye la distribución empírica de la variable observada.

Serie estadística cualitativa

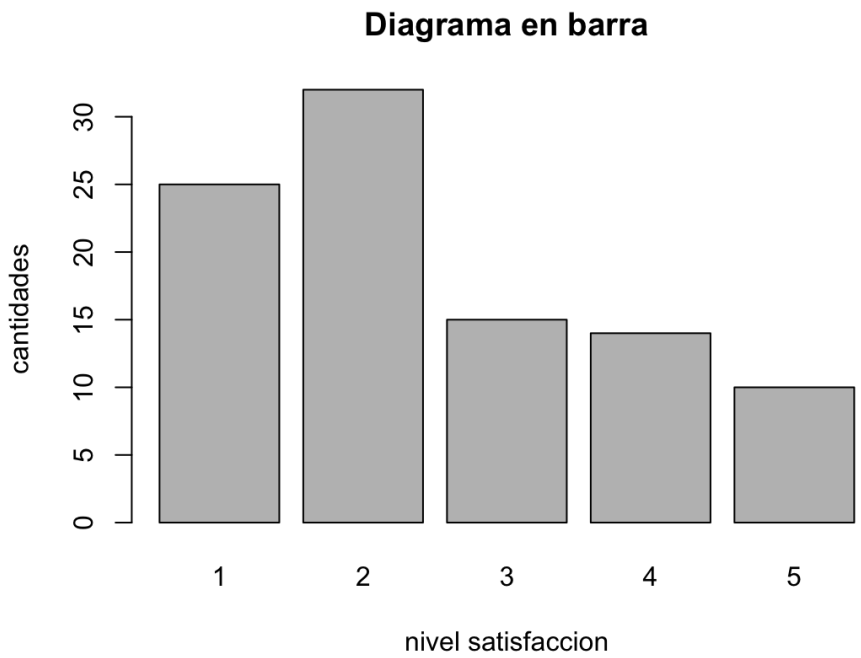


Figura: Barplot del ejemplo

Serie estadística continua

Si X es el puntaje de 120 estudiantes en una prueba de probabilidad:

97	63	81	9	27	65	2	59	25	94	21	85	38	32	36
90	79	57	15	76	44	53	34	51	74	31	88	29	28	27
97	81	3	58	66	4	84	59	50	77	50	14	83	72	11
65	39	84	41	1	40	81	70	40	25	41	26	29	23	55
73	36	87	94	73	66	91	20	21	86	59	95	91	88	64
64	85	80	82	7	25	71	58	32	90	92	47	1	64	5
85	26	92	60	5	74	61	52	64	33	77	67	69	89	26
24	52	51	92	87	70	6	61	43	40	36	59	62	92	61

Se pueden identificar algunas cantidades, en particular los 5 números importantes que nos servirán para hacer el diagrama de caja: el mínimo m, el primer cuartil Q_1, la mediana m, el tercer cuartil Q_3, el máximo M (se incluye además el promedio pr):

x_m	Q_1	m	$\overline{x}$	Q_3	x_M
1.00	32	59.00	54.52	79	97.00

Si $\hat{F}(x)$ es la proporción de individuos en la muestra con un valor menor o igual a x entonces:

- La mediana es el valor m de la muestra tal que $\hat{F}(m) = 0{,}5$
- Cuartiles: $\hat{F}(Q_1) = 0{,}25$, $\hat{F}(Q_2) = 0{,}5$, $\hat{F}(Q_3) = 0{,}75$
- Cuantil p: $\hat{F}(x_p) = p$

Serie estadística continua

Estas cantidades se pueden definir en termino de la noción de profundidad de un dato en la serie.
Si tenemos n mediciones $x_1, \ldots, x_n$ y los ordenamos de menor a mayor, la profundidad p de x_i es el lugar que ocupa este dato en esta secuencia ordenada $x_1^* \leq x_2^* \leq \ldots, x_n^*$.
Entonces

- La profundidad de la mediana es $p(m) = \lfloor \frac{n+1}{2} \rfloor$ o la mediada es $m = x^*_{\lfloor \frac{n+1}{2} \rfloor}$
- $Q_1 = \text{mín} \left\{ x_i : p(x_i) \geq \frac{n}{4} \right\}$
- $Q_3 = \text{mín} \left\{ x_i : p(x_i) \geq \frac{3n}{4} \right\}$

La mediana divide en dos partes "iguales" el conjunto de mediciones y el intervalo intercuartílico $[Q_1, Q_3]$ contiene aproximadamente el 50 % de los datos.

Serie estadística continua

- El valor indicado por la raya en la caja es la mediana
- Las extremidades de la caja corresponden a los cuartiles Q_1 y Q_3
- El rango intercuartílico es $R = Q_3 - Q_1$
- Las extremidades de los bigotes corresponden al valor más grande menor a $Q_3 + 1{,}5R$ y al valor más chico mayor a $Q_1 - 1{,}5R$
- Los puntos alineados con los bigotes pero fuera de ellos corresponden a los puntos atípicos o outliers

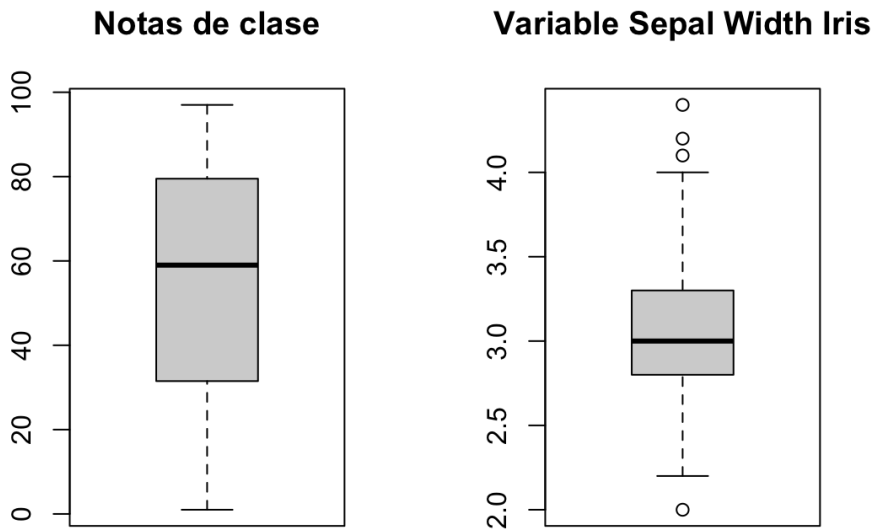
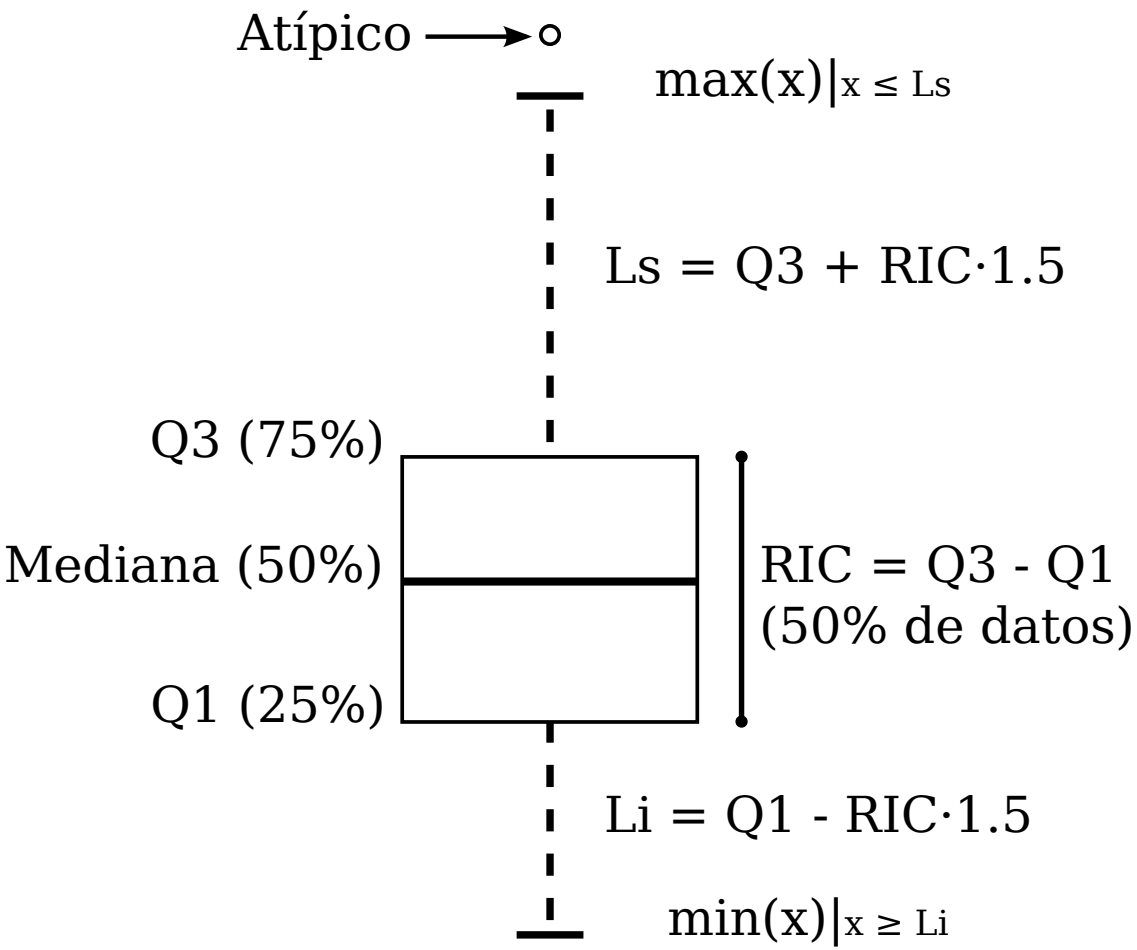
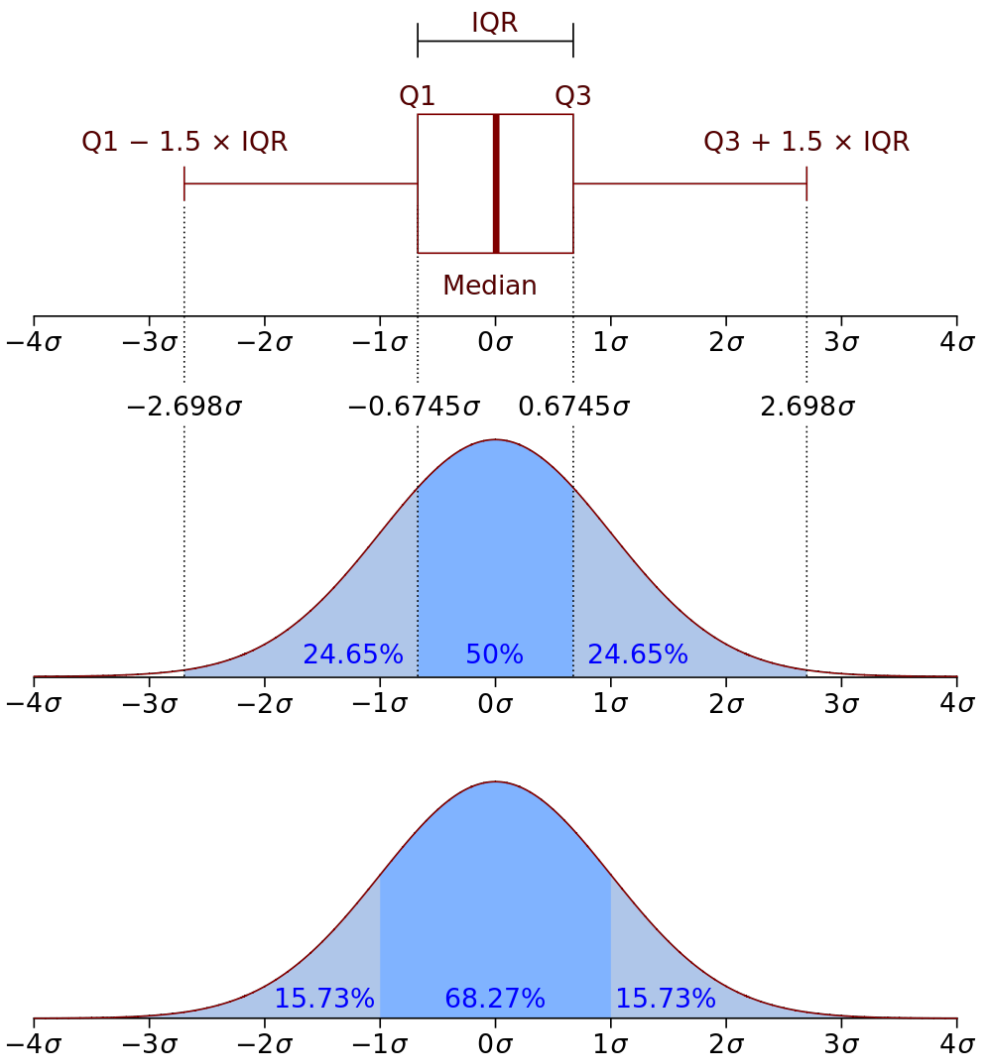


Figura: Boxplots

Diagrama de caja o boxplot



Comparación con la normal



En la distribución normal $N(\mu, \sigma^2)$, el 99.73 % del área debajo la campana está a menos de 3σ de distancia de μ.
El *IQR* es en este caso $1{,}349\sigma$. El valor $0{,}6745\sigma$ se conoce históricamente como *error probable*.

Fuente: Wikipedia

Histograma

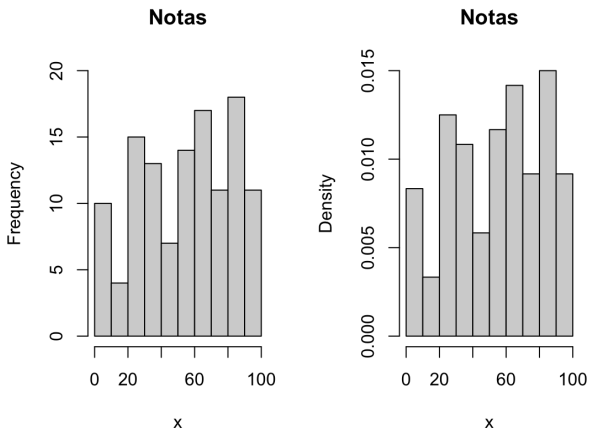


Figura: Datos iris

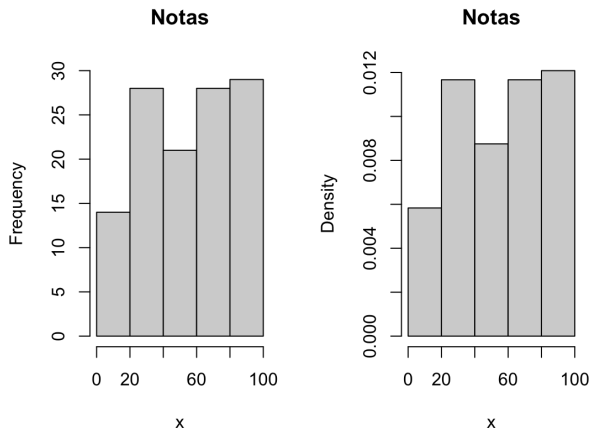
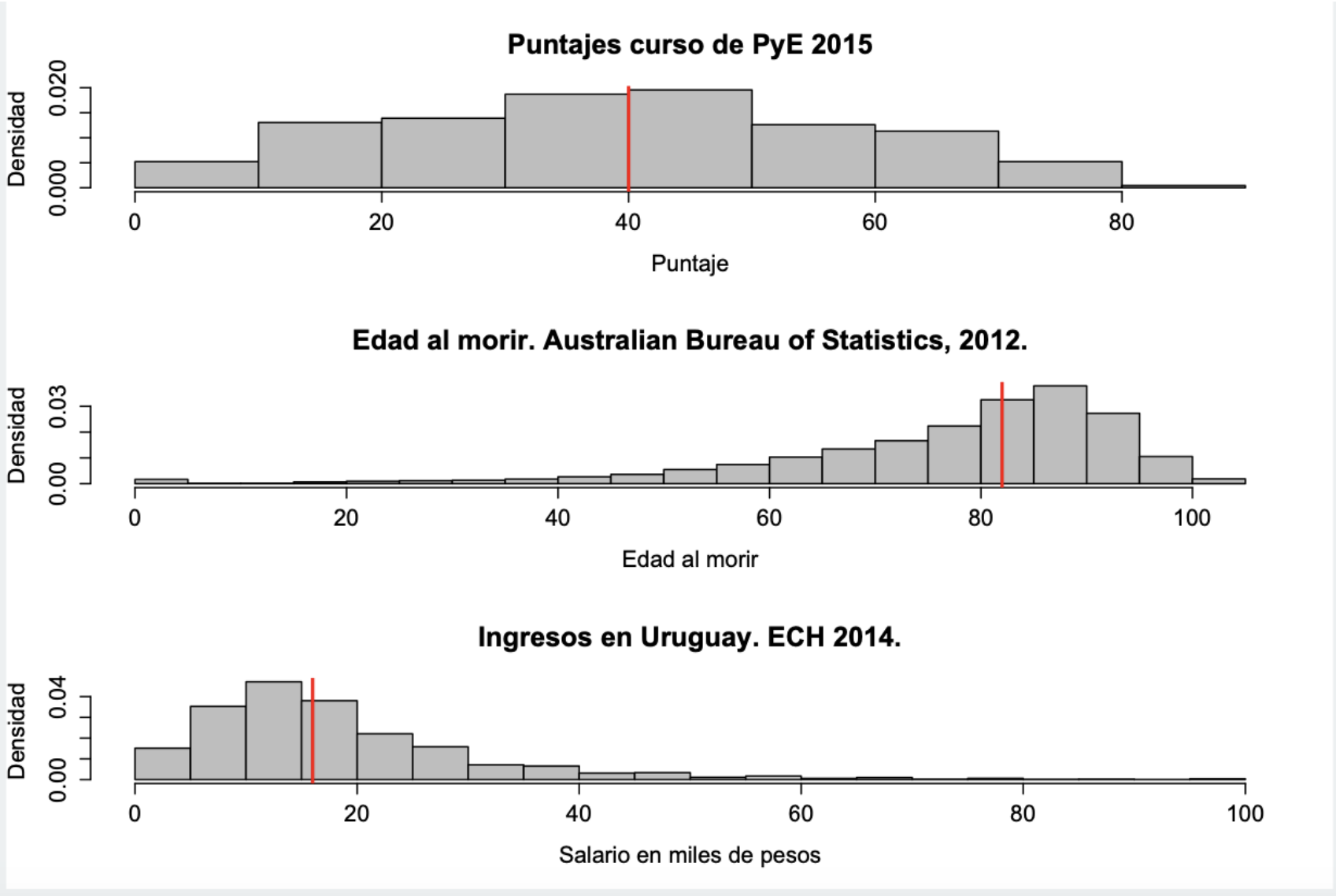


Figura: Datos airquality

	[0,20]	(20,40]	(40,60]	(60,80]	(80,100]
Cantidades	14	28	21	28	29
Proporciones	0.12	0.23	0.17	0,23	0,24
Cantidades cumuladas	14	42	63	91	120
Proporciones cumuladas	0.12	0.35	0.52	0.76	1

Serie estadística continua

Simetría



Simetría

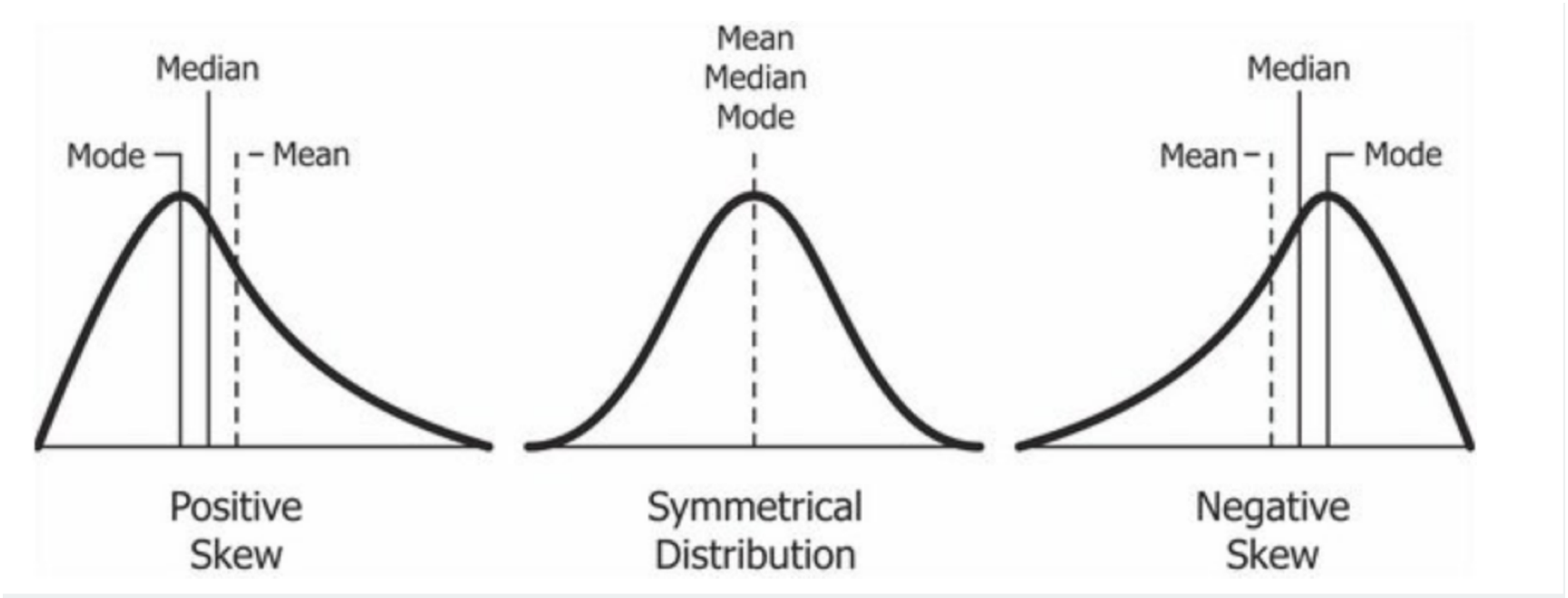
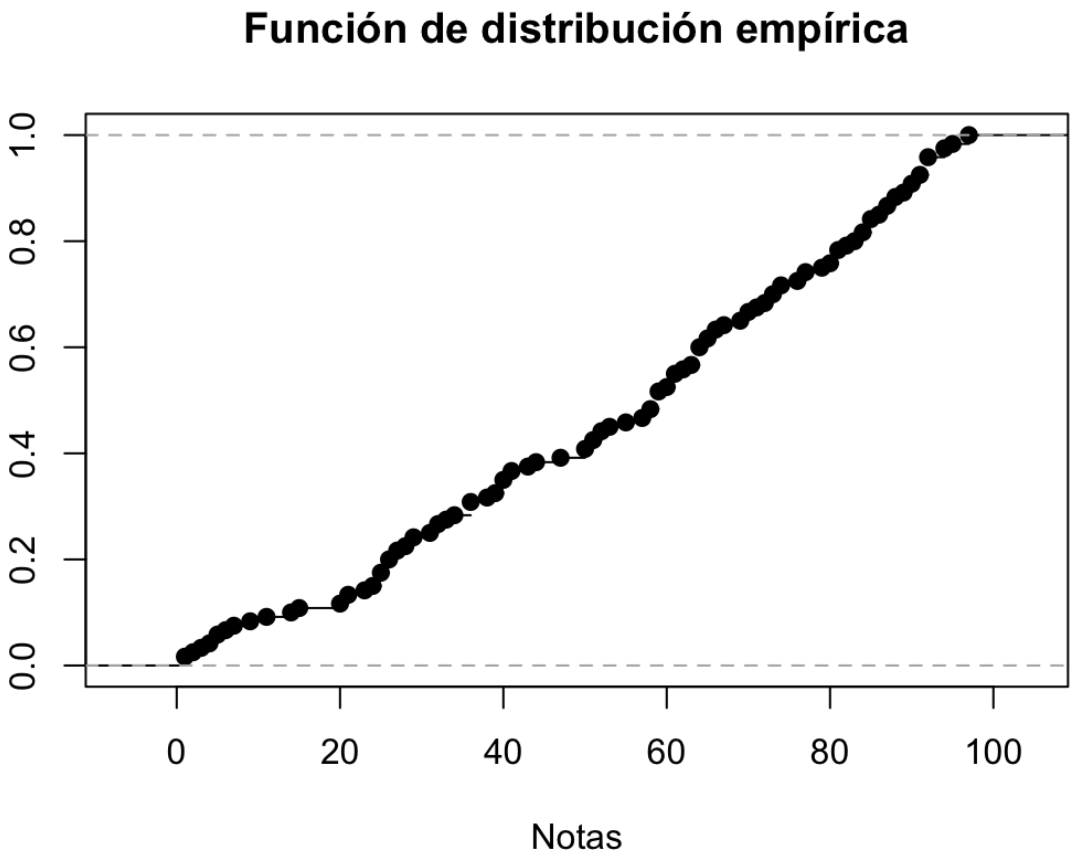


Figura: fuente: Wikipedia

Función de distribución empírica

$$F_n(t) = \sum_{i=1}^{n} \frac{\#\{\text{observaciones menores o iguales a t}\}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{i:x_i \leq t\}}$$



Serie estadística continua

Otros valores de interés. Si $x_1, \ldots, x_n$ es la muestra, consideramos también como indicadores:

- la media empírica: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- la varianza: $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- la varianza muestral: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- el desvío estándar: σ
- el mínimo x_m, el máximo x_M, el rango $x_M - x_m$
- la moda

Serie temporal

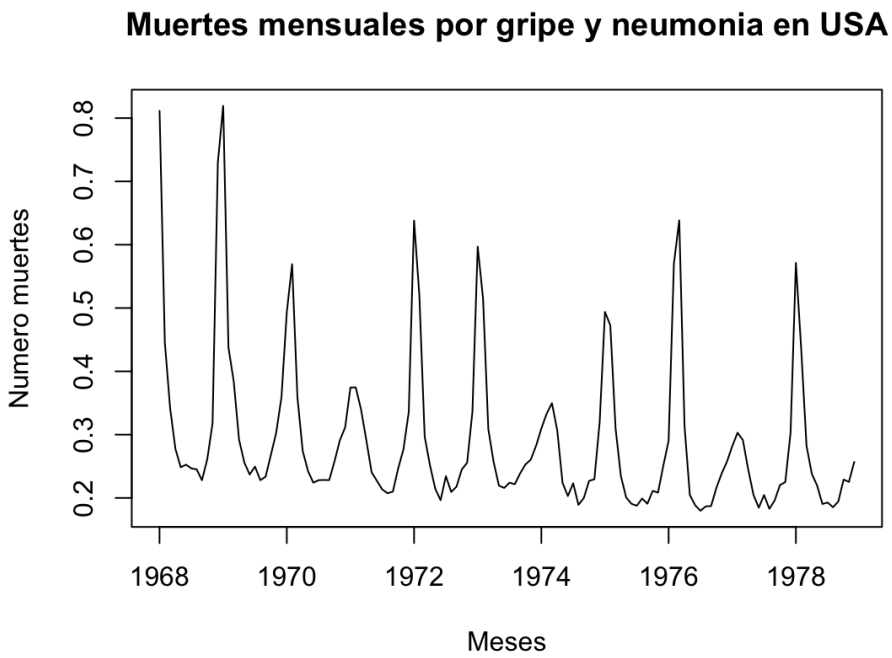


Figura: Gripe en USA

Primeros elementos de comparación de dos series

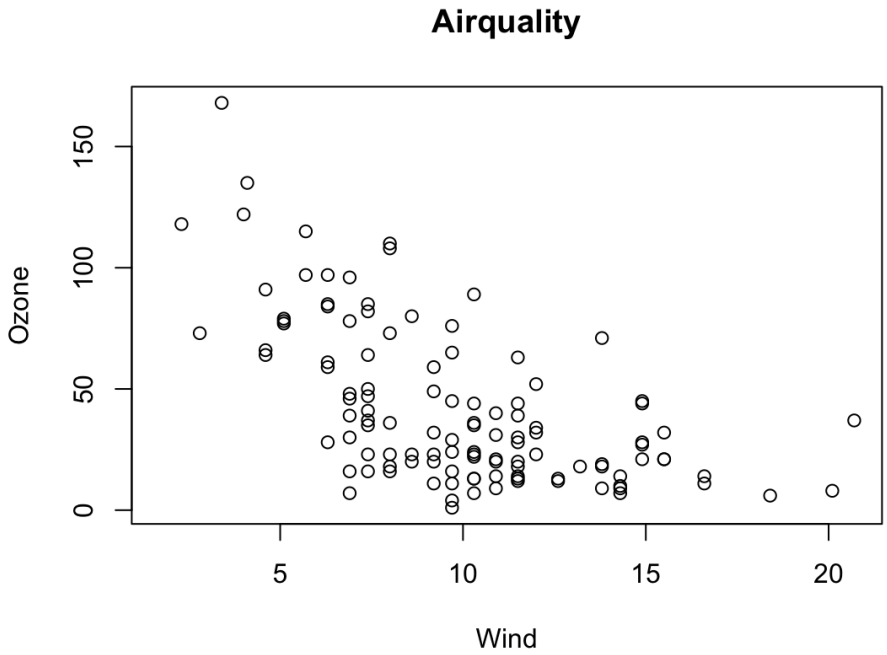
Consideramos sobre n individuos dos características cuantitativas

$$x_1, x_2, \ldots, x_n \quad y_1, y_2, \ldots, y_n$$

Se puede distinguir tres casos: las dos variables son cuantitativas, las dos variables son cualitativas, una variable es cualitativa y la otra es cuantitativa.

Primeros elementos de comparación de dos series

Dos variables continuas

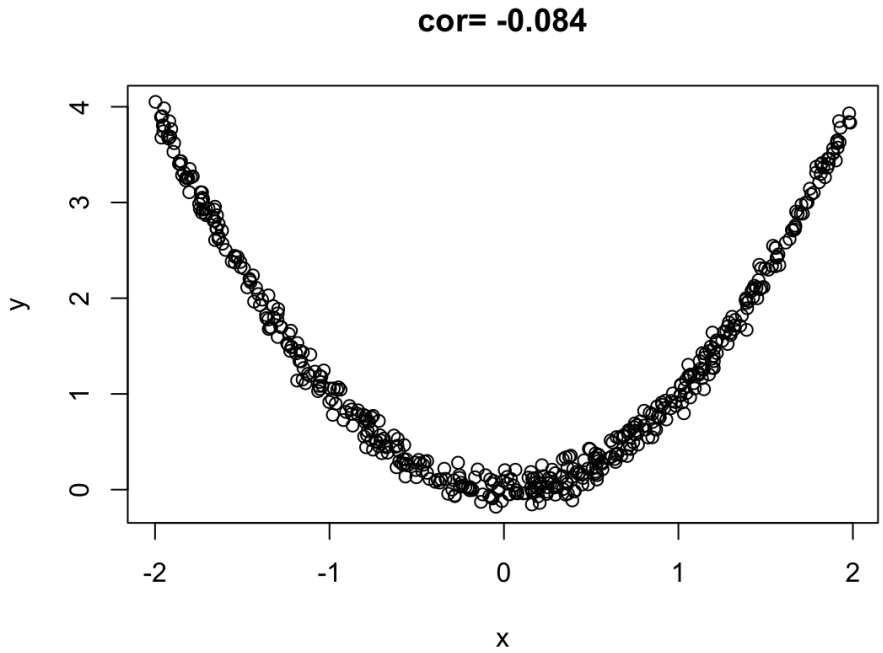


Coeficiente de correlación lineal

$$\rho_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

Primeros elementos de comparación de dos series

Dos variables continuas.
Dos problemas con la correlación:
1. La no correlación no significa independencia:



Primeros elementos de comparación de dos series

En 1971 el estadístico Frank Anscombe [1] (Yale) publicó cuatro conjuntos de datos con 11 observaciones y las mismas propiedades estadísticas:

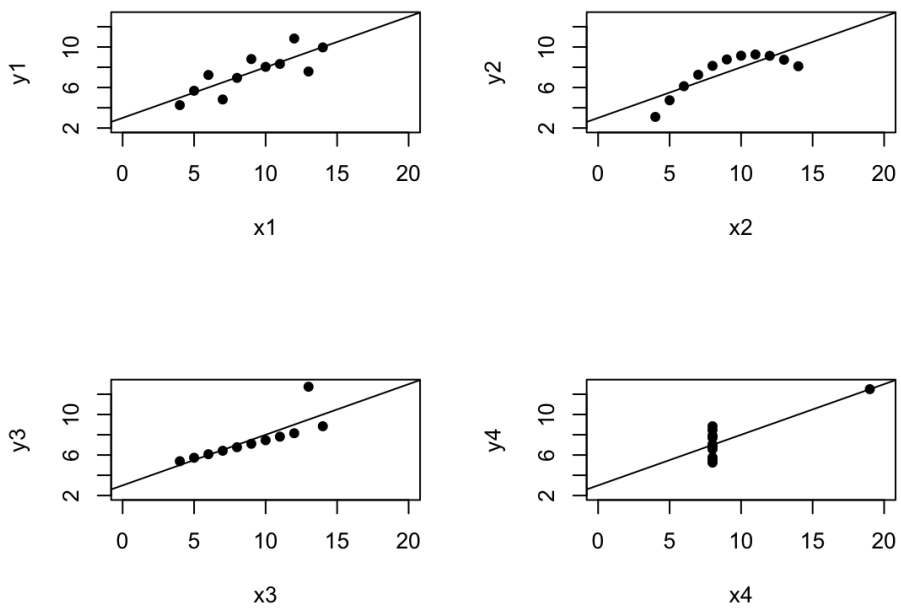


Figura: Cuarteto de Anscombe

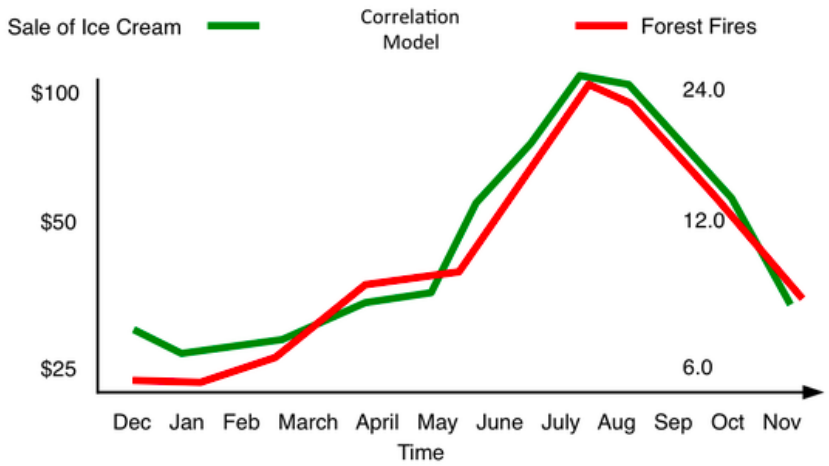
$\overline{x} = 9, \sigma_x^2 = 11, \overline{y} = 7{,}5, \sigma_y^2 = 4{,}12, \rho_{x,y} = 0{,}816$ y por lo tanto recta de regresión lineal

$$y = 3 + 0{,}5x$$

[1] https://tereom.github.io/est-computacional-2018/introduccion-a-visualizacion.html

Primeros Elementos de comparación de dos series

Dos variables continuas.
Dos problemas con la correlación:
2. Correlación no implica causalidad



Primeros elementos de comparación de dos series

Dos variables cualitativas
Tabla de contingencia
La variable X tiene modalidades $a_1, \ldots, a_K$ y la variable Y modalidades $b_1, \ldots, b_L$
El número n_{ij} es la cantidad de observaciones que tienen la modalidad a_i para X y la modalidad b_j para Y.

	b_1	b_2	$\cdots$	b_j	$\cdots$	b_L	
a_1	n_{11}	n_{12}	$\cdots$		$\cdots$	n_{1L}	$n_{1.}$
a_2	n_{21}	n_{22}	$\cdots$		$\cdots$	n_{2L}	$n_{2.}$
a_i				n_{ij}			$n_{i.}$
a_K	n_{K1}	n_{K2}	$\cdots$		$\cdots$	n_{KL}	$n_{K.}$
	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.L}$	n

Primeros elementos de comparación de dos series

Dos variables cualitativas

Tabla de contingencia de plebícito de una ley según preferencia política

	a favor	indiferente	en contra	total
democrato	138	83	64	285
republicano	64	67	84	215
total	202	150	148	500

Primeros elementos de comparación de dos series

Dos variables cualitativas

Tabla de contingencia de plebícito de una ley según preferencia política

	a favor	indiferente	en contra	total
democrato	138	83	64	285
republicano	64	67	84	215
total	202	150	148	500

Recordar que A y B son independientes si $P(A \cap B) = P(A)P(B)$

Primeros elementos de comparación de dos series

Dos variables cualitativas

Tabla de contingencia de plebícito de una ley según preferencia política

	a favor	indiferente	en contra	total
democrato	138	83	64	285
republicano	64	67	84	215
total	202	150	148	500

Recordar que A y B son independientes si $P(A \cap B) = P(A)P(B)$ $\mathbb{P}(a_1) = \frac{n_{1\cdot}}{n}$ y $\mathbb{P}(b_1) = \frac{n_{\cdot 1}}{n}$ y esperamos que $\mathbb{P}(a_1 \cap b_1) = \mathbb{P}(a_1)\mathbb{P}(b1) = \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n} = \frac{n_{1\cdot}n_{\cdot 1}}{n^2}$

La cantidad esperada es entonces $E_{11} = P(a_1 \cap b_1)n = \frac{n_{1\cdot}n_{\cdot 1}}{n}$

Primeros elementos de comparación de dos series

Dos variables cualitativas

Tabla de contingencia de plebícito de una ley según preferencia política

	a favor	indiferente	en contra	total
democrato	138	83	64	285
republicano	64	67	84	215
total	202	150	148	500

Recordar que A y B son independientes si $P(A \cap B) = P(A)P(B)$ $\mathbb{P}(a_1) = \frac{n_{1\cdot}}{n}$ y $\mathbb{P}(b_1) = \frac{n_{\cdot 1}}{n}$ y esperamos que $\mathbb{P}(a_1 \cap b_1) = \mathbb{P}(a_1)\mathbb{P}(b1) = \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n} = \frac{n_{1\cdot}n_{\cdot 1}}{n^2}$

La cantidad esperada es entonces $E_{11} = P(a_1 \cap b_1)n = \frac{n_{1\cdot}n_{\cdot 1}}{n}$

EL test de independencia χ^2 es

$$\begin{cases} (H_0): & X \text{ e } Y, \text{son independientes} \\ (H_1): & X \text{ e } Y \text{ no son independientes} \end{cases}$$

y tiene como estadística

$$\chi^2 = \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

donde $E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$ es la cantidad teórica esperada en el casillero (i, j) si las variables fueran independientes.

Es positivo y tiene distribución χ^2 con $(L-1)\times(K-1)$ grados de libertad. Más cerca está de 0, más las variables tienden a ser independientes.

Distribución χ^2

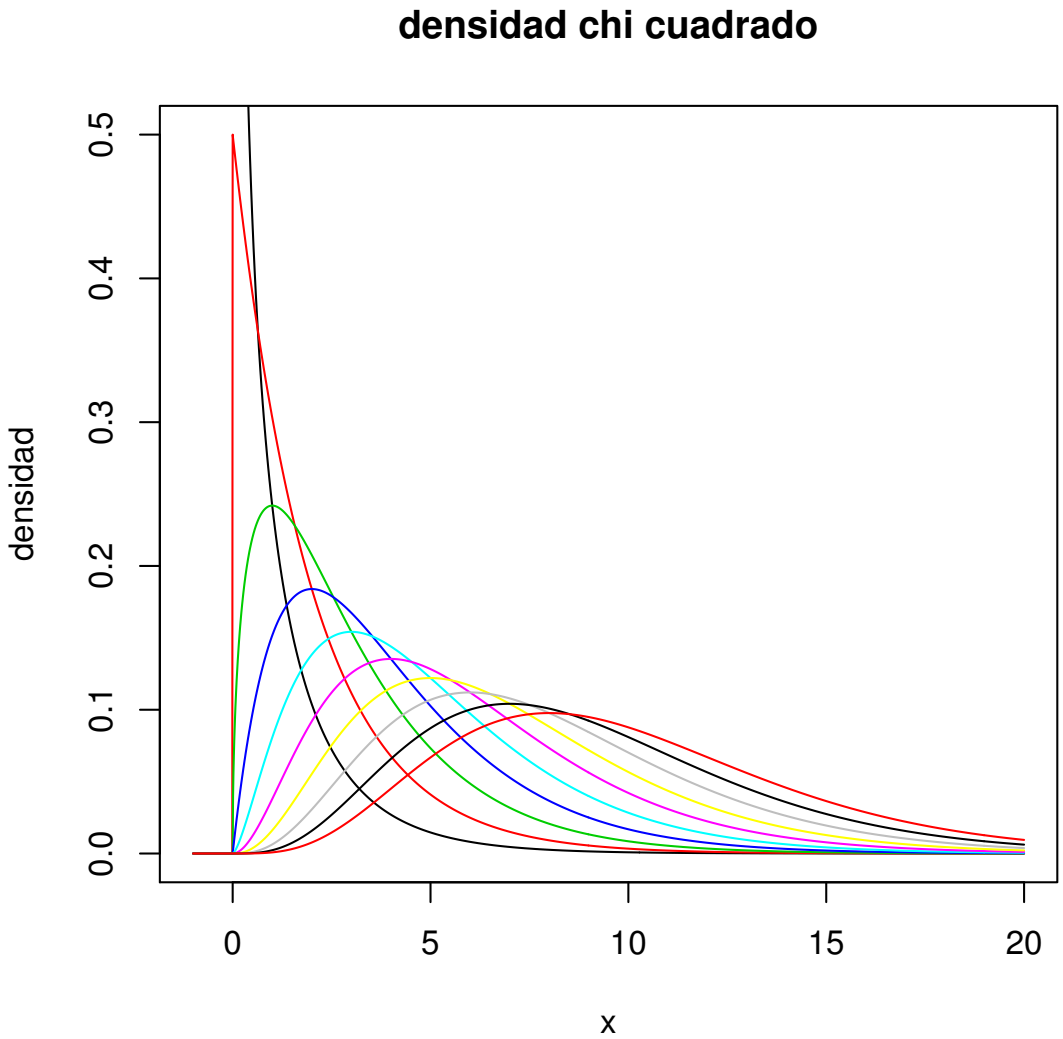


Figura: Densidad χ^2 con 1,2,...,10 grados de libertad

Primeros elementos de comparación de dos series

Dos variables cualitativas
Cantidades esperadas:

	a favor	indiferente	en contra	total
democrato	$\frac{285(202)}{500} = 115,14$	$\frac{285(150)}{500} = 85,5$	$\frac{285(148)}{500} = 84,36$	285
republicano	$\frac{215(202)}{500} = 86,86$	$\frac{215(150)}{500} = 64,5$	$\frac{215(148)}{500} = 63,64$	215
total	202	150	148	500

	a favor	indiferente	en contra	total
democrato	138 (115.4)	83 (85.5)	64 (84,36)	285
republicano	64 (86.86)	67 (64.50)	84 (63,64)	215
total	202	150	148	500

El cálculo de la estadística del test es:

$$\chi^2 = \frac{(138-115,4)^2}{115,14} + \frac{(83-85,5)^2}{85,5} + \frac{(64-86,86)^2}{86,86} + \frac{(64-86,86)^2}{86,86}$$

$$+\frac{(67-64,5)^2}{64,5} + \frac{(84-63,64)^2}{63,64} = 22,152$$

con grados de libertad $(2-1)(3-1) = 2$
Como $\mathbb{P}(\chi^2 > 22{,}152) = 0{,}001 < 0{,}05$ rechazamos la hipótesis nula de independencia.

Primeros elementos de comparación de dos series

Dos variables cualitativas (test χ^2)

```
> base=read.table("vacaciones.csv",sep=",",header=TRUE,row.names=1)
> k=chisq.test(base)
> k$observed
              Hotel Locación Res.Secund Padres Amigos Camping Grupo.Viaje Otros
Prod. Rurales   195       62          1    499     44     141          49    65
Jefes           700      354        229    959    185     292         119   140
Ejecutivo sup   961      471        633   1580    305     360         162   148
Ejecutivo prom  572      537        279   1689    206     748         155   112
Empleado        441      404        166   1079    178     434         178    92
Obrero          783     1114        387   4052    497    1464         525   387
Otras prof.     142      103        210   1133    132     181          46    59
Inactivos       741      332        327   1789    311     236         102   102
> k$expected
                 Hotel  Locación Res.Secund    Padres    Amigos   Camping Grupo.Viaje     Otros
Prod. Rurales  154.0899  114.7435   75.83873  434.2379  63.13099  131.0189    45.39451  37.54561
Jefes          434.5452  323.5853  213.87097 1224.5838 178.03417  369.4832   128.01596 105.88146
Ejecutivo sup  674.1433  502.0026  331.79446 1899.7909 276.19808  573.2076   198.60098 164.26204
Ejecutivo prom 627.1576  467.0146  308.66939 1767.3812 256.94791  533.2568   184.75910 152.81348
Empleado       433.6697  322.9333  213.44007 1222.1165 177.67547  368.7388   127.75804 105.66814
Obrero        1343.7632 1000.6369  661.36259 3786.8342 550.54287 1142.5691   395.86937 327.42189
Otras prof.    292.7124  217.9691  144.06487  824.8875 119.92497  248.8863    86.23238  71.32244
Inactivos      574.9188  428.1148  282.95891 1620.1680 235.54555  488.8394   169.36967 140.08494
> k

        Pearson's Chi-squared test

data:  base
X-squared = 2292.148, df = 49, p-value < 2.2e-16
```

En la primera tabla vemos los n_{ij} y en la segunda tabla los productos $n_{i.}n_{.j}/n$.
Conclusión: Hay una dependencia significativa entre las variables.