



A Short Course in Basic Statistics

Ian Schindler

November 5, 2017

Creative commons license share and share alike  

1 Descriptive Statistics

1.1 Presenting statistical data

Definition 1 • *A statistical population is a set Ω of objects of study. We may assume that $\text{card}(\Omega) < \infty$. Therefore there is a bijection $i : \Omega \mapsto \llbracket 1, n \rrbracket$ for some $n \in \mathbb{N}$. This leads to using Ω and $\llbracket 1, n \rrbracket$ interchangeably.*

- *A random variable is a function $X : \Omega \mapsto \omega$.*

The goal of descriptive statistics is to summarise the range of random variables.

1.1.1 Raw data

Let $X : \Omega = \llbracket 1, n \rrbracket \mapsto \omega$. If $\omega \subset \mathbb{R}$, the value of X at $i \in \llbracket 1, n \rrbracket$, $X(i)$ will be denoted x_i .

- *Continuous:* $\omega = I \subset \mathbb{R}$.
- *Discrete:* $\omega \subset \mathbb{R}$ and $\text{card}(\omega)$ is “small”.
- *Categorical:* ω is nominal.

1.1.2 A table of a simple statistical series

For a discrete variable with K categories, $\{x_i\}_{i=1}^n$, we define:

- The class x_k where $x_k < x_{k+1}$ for $k \in \llbracket 1, K-1 \rrbracket$.
- The number of elements in the class k : n_k with $n = \sum_{k=1}^K n_k$.
- The frequency: $f_k = n_k/n$ with $\sum_{k=1}^K f_k = 1$.

- A discrete variable can be represented in a table:

$$(k, x_k, n_k)_{k=1}^K. \quad (1)$$

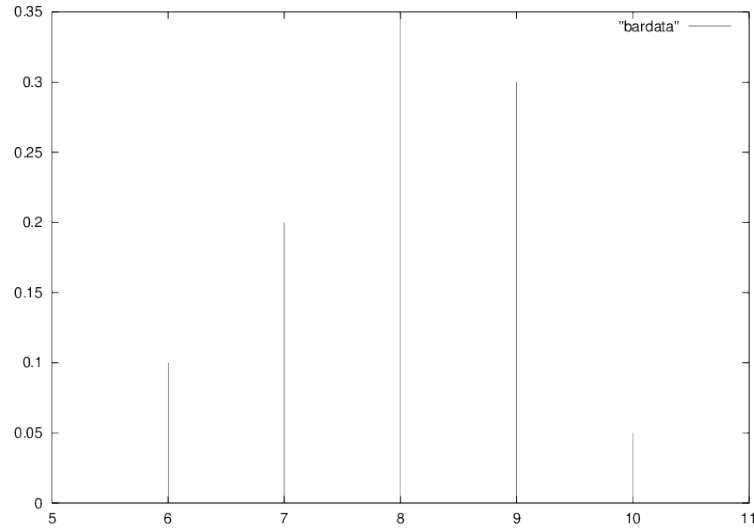
- More completely:

$$(k, x_k, n_k, f_k)_{k=1}^K \quad (2)$$

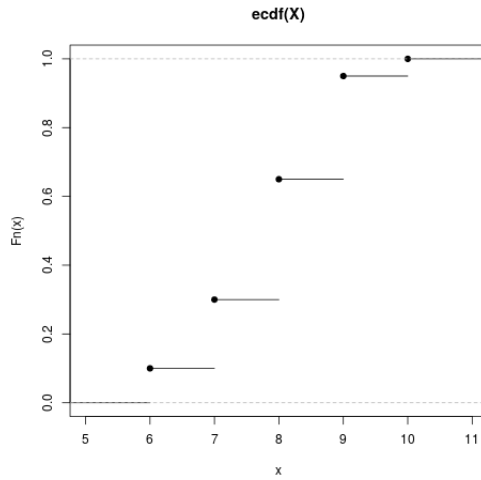
The cumulative number of elements is $s_k = \sum_{l=1}^k n_l$ with $s_K = n$. The cumulative frequency $\Phi_k = \sum_{l=1}^k f_l$. Below is a table representing a simple statistical series where Ω is 40 children and $X : \Omega \mapsto \mathbb{R}$ is their age in years.

k	x_k	n_k	s_k	f_k	Φ_k
1	6	4	4	0,1	0,10
2	7	8	12	0,2	0,30
3	8	14	26	0,35	0,65
4	9	12	38	0,30	0,95
5	10	2	40	0,05	1

The table can be represented by a bar graph:



The *empirical distribution function* $\Phi(x) : \mathbb{R} \mapsto [0, 1]$ is defined by $\Phi(x) :=$ the proportion of the range of $X \leq x$. Thus Φ is monotone increasing with $\Phi(-\infty) = 0$, $\Phi(+\infty) = 1$, $\Phi(x_k) = \Phi_k$ between the values of the range of X $\Phi(x)$ is constant. The function $\Phi(x)$ can be deduced from the above bar graph of X .



The *Stem and Leaf* diagram, introduced by J.W. Tukey (1977), gives more information than a barplot.

The decimal point is at the |

```

6 | 0000
6 |
7 | 00000000
7 |
8 | 0000000000000000
8 |
9 | 00000000000000
9 |
10 | 00

```

If the variable X is continuous rather than discrete, classes are prescribed so that it can be treated as if it were discrete. The number of classes, and the bounds on the classes are left to the software package, or chosen to make the graphics look good. A rule of thumb is to have \sqrt{n} classes. The data can then be presented in a table of the form:

$$[k, [b_{k-1}, b_k[, n_k]_{k=1, K} \text{ such that } \sum_{k=1}^K n_k = n$$

We define the amplitude of the class as $a_k := b_k - b_{k-1}$. To simplify, we will only consider constant amplitudes.

Rather than a barplot representing the frequencies of X , we use a histogram. The height of the classe in the histogram usually corresponds to the number in the class. The frequency of the class can also be used.

The definition of the empirical distribution function does not change. One can think of the empirical distribution function as the integral of the histogram of X (using frequencies) just as a distribution function is the integral of the probability density function in probability.

1.2 Two variables

Suppose that $X : \Omega \mapsto \omega_1$ and $Y : \Omega \mapsto \omega_2$, that is, using the equivalence $\Omega \equiv \llbracket 1, n \rrbracket$ we have couples $(x_i, y_i)_{i=1}^n$. Let $[X_1, \dots, X_j, \dots, X_J]$ and $[Y_1, \dots, Y_j, \dots, Y_J]$ designate the states of X and Y respectively (quantitative or nominal).

Let n_{jk} be the number of elements with state X_j and state Y_k .

X^Y	Y_1	\cdots	Y_k	\cdots	Y_K	
X_1	n_{11}	\cdots	n_{1k}	\cdots	n_{1K}	$n_{1.}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_j	n_{j1}	\cdots	n_{jk}	\cdots	n_{jK}	$n_{j.}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_J	n_{J1}	\cdots	n_{Jk}	\cdots	n_{JK}	$n_{J.}$
	$n_{.1}$	\cdots	$n_{.k}$	\cdots	$n_{.K}$	n

Table 1: Contingency table

Where $n_{j.} = \sum_k n_{jk}$, $n_{.k} = \sum_j n_{jk}$, and $n = \sum_j n_{j.} = \sum_k n_{.k}$.

The contingency table can be normalised to a frequency table by dividing each number by n to obtain: $f_{jk} = \frac{n_{jk}}{n}$. The conditional frequencies are used to establish the independence of the variable X and Y .

- *Conditional row profile* for row j : $\frac{n_{jk}}{n_{j.}}$
- *Conditional column profile* for column k : $\frac{n_{jk}}{n_{.k}}$.

1.3 Summary statistics of a quantitative variable X

1.3.1 Measures of the center

Definition 2 • *The mode:= the class or center of the class with the highest frequency.*

- *The median* : Q_2 is defined by $\Phi(Q_2) \geq .5$ and for all $x < Q_2$, $\Phi(x) < .5$.
- *The arithmetic mean or average*: $\bar{x} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n x_i$.
- *The geometric mean*: $\text{GM} \stackrel{\text{def}}{=} (\prod_{i=1}^n x_i)^{1/n}$.

Note that if $y_i = x_i + a$, then $\bar{y} = \bar{x} + a$.

1.3.2 Measures of dispersion

- *The range*: $R = x_{\max} - x_{\min}$.
- *The interquartile range*: $\text{IQR} \stackrel{\text{def}}{=} Q_3 - Q_1$.
- *The variance*:

$$\begin{aligned}
 \text{Var}X &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) - \bar{x}^2.
 \end{aligned}$$

To estimate the variance an unbiased estimator is used:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- The k th moment: $\mu_k := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$.
- *Standard deviation*:

$$\sigma_X = \sqrt{\text{Var } X} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

- The *coefficient of variation*: $c_X = \frac{\sigma_X}{\bar{x}}$.
- etc.

Note that $\text{Var}(aX + b) = a^2 \text{Var } X$.

2 Graphics

Graphics should clarify thus be readable. This means a graph should be titled and scales should be clear.

2.0.3 Nominative variables

- Column graph
- Bar graph
- Pie chart.

2.0.4 Discrete quantitative variable

- Bar graph.
- Empirical distribution function.

2.1 Continuous quantitative variable

- Histogramme
- Empirical distribution function.

2.1.1 Two quantitative variables

- Scatterplot.

2.2 Link between two variables

The scalar product of two vectors X and Y in \mathbb{R}^n is defined as $(X, Y) = \sum_{i=1}^n x_i y_i$. Note that $\|X\|^2 = (X, X)$. We also have $(X, Y) = \|X\| \|Y\| \cos \theta$ where θ is the angle between X , and Y .

2.2.1 Two quantitative variables X et Y

The data $(i, x(i), y(i))_{i=1}^n$ can be represented as vectors X and $Y \in \mathbb{R}^n$. We also define $\bar{X} \stackrel{\text{def}}{=} \bar{x}(1, \dots, 1)$. We have $\text{Var } X = \frac{1}{n}(X - \bar{X}, X - \bar{X}) = \frac{1}{n}\|X - \bar{X}\|^2$. We compute

- The *covariance*:

$$\begin{aligned} \text{Cov } (X, Y) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \left[\frac{1}{n} \sum_{i=1}^n x_i y_i \right] - \bar{x} \bar{y} \\ &= \frac{1}{n} (X - \bar{X}, Y - \bar{Y}). \end{aligned}$$

Property: $\text{Var } (X + Y) = \text{Var } X + \text{Var } Y + 2\text{Cov}(X, Y)$.

- The *Bravais Pearson correlation coefficient*.

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \cos \theta.$$

Where θ is the angle between Y and \hat{Y} (see below).

- $R^2 = \|\hat{Y} - \bar{Y}\|^2 / \|Y - \bar{Y}\|^2 = \cos^2 \theta$, Where θ is as above.

Linear regression or *least squares* regression consists of projecting $Y - \bar{Y}$ onto the span of $X - \bar{X}$. In other words we solve the problem:

$$\min_{a \in \mathbb{R}} \|Y - \bar{Y} - a(X - \bar{X})\|^2. \quad (3)$$

to obtain the projection of $Y - \bar{Y}$ onto the space spanned by $X - \bar{X}$. The projection $\hat{a}(X - \bar{X})$ is called $\hat{Y} - \bar{Y}$. Equivalently we suppose $y_i = f(x_i) = ax_i + b + \epsilon_i$ and compute \hat{a} and \hat{b} such that $\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (\hat{a}x_i + \hat{b})^2$ is minimal. One obtains

$$\begin{aligned} \hat{a} &= \frac{\text{Cov } (X, Y)}{\text{Var } X} \\ \hat{b} &= \bar{y} - \hat{a}\bar{x}. \end{aligned}$$

Remark 1 *The equation*

$$\hat{Y} = \hat{a}(x - \bar{x}) + \bar{y}$$

Defines the linear regression line which goes through the point (\bar{x}, \bar{y}) .

From the contingency table one can calculate:

- The mean:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^I n_{i.} x_i.$$

- La variance:

$$\sigma^2(X) := \frac{1}{n} \sum_{i=1}^I n_{i.} (x_i - \bar{x})^2.$$

- The *conditional means* :

$$\bar{x}_j := \frac{1}{n_{.j}} \sum_{i=1}^I n_{ij} x_i.$$

- The *conditional variances*:

$$\sigma_j^2(X) := \frac{1}{n_{.j}} \sum_{i=1}^I n_{ij} (x_i - \bar{x}_j)^2.$$

- The *variance of the conditional means of X* or *explained variance*:

$$\sigma^2(\bar{x}_j) := \frac{1}{n} \sum_{j=1}^J n_{.j} (\bar{x}_j - \bar{x})^2.$$

- The *average conditional variance of X* or *residual variance*:

$$\overline{\sigma_j^2(x)} := \frac{1}{n} \sum_{j=1}^J n_{.j} \sigma_j^2(X).$$

- The *correlation quotients*: $\eta_{X/Y}^2$ and $\eta_{Y/X}$:

$$\begin{aligned} \eta_{X/Y}^2 : &= \frac{\sigma^2(\bar{x}_j)}{\sigma^2(X)} \\ \eta_{Y/X}^2 : &= \frac{\sigma^2(\bar{y}_i)}{\sigma^2(Y)} \end{aligned}$$

The correlation quotients are used to test for a functional relationship (not necessarily linear) between X and Y .

2.2.2 Time Series

Indices:

- *Elementary indices*: $I_{\frac{n}{0}} = P_n/P_0$.

Properties:

- circularity: $I_{\frac{n}{0}} = I_{\frac{n}{n'}} \times I_{\frac{n'}{0}}$
- reversibility: $I_{\frac{0}{n}} = 1/I_{\frac{n}{0}}$.
- Chains: $I_{\frac{n}{0}} = I_{\frac{n}{n-1}} \times I_{\frac{n-1}{n-2}} \times \cdots \times I_{\frac{1}{0}}$.

- *Synthetic indices*: Let $\{B_i\}_{i=1}^N$, be N goods, with respective prices: $\{p_0^i\}_{i=1}^N$ at $t = t_0$ and $\{p_n^i\}_{i=1}^N$ at $t = t_n$ bought with respective quantities $\{q_0^i\}$ at $t = t_0$ and $\{q_n^i\}$ at $t = t_n$.

- Simple average index:

$$I_{\frac{n}{0}} = \frac{\sum_{i=1}^N p_n^i}{\sum_{i=1}^N p_0^i}$$

- Average of indices:

$$I_{\frac{n}{0}} = \frac{1}{N} \sum_{i=1}^N I_{\frac{n}{0}}^i$$

- The Laspeyre index:

$$L_{\frac{n}{0}} := \frac{\sum_{i=1}^N p_n^i q_0^i}{\sum_{i=1}^N p_0^i q_0^i} = \sum_i \rho_i \frac{p_n^i}{p_0^i}.$$

- The Paasche index:

$$P_{\frac{n}{0}} := \frac{\sum_{i=1}^N p_n^i q_n^i}{\sum_{i=1}^N p_0^i q_n^i} = \frac{1}{\sum_i \rho_i \frac{p_0^i}{p_n^i}}.$$

A simple time series is a data set of the form: $(t, x_t)_{t=1}^T$. Let

$$T_t := \text{the trend} \tag{4}$$

$$s_t := \text{seasonal variations} \tag{5}$$

$$c_t := \text{cyclical variations} \tag{6}$$

$$\epsilon_t := \text{random variations.} \tag{7}$$

We consider the model:

$$x_t = T_t + s_t + c_t + \epsilon_t.$$

In certain cases

$$x_t = T_t s_t c_t \epsilon_t$$

In which case we consider $\log x_t$ to reduce to the first case. We suppose s_t to be periodic with 0 average, that is, there is a k such that

$$\begin{aligned} s_t &= s_{t+k} \quad \forall t \\ \sum_{i=0}^k s_{t+i} &= 0 \quad \forall t. \end{aligned}$$

The *moving averages* k :

$$y_t := \frac{x_{t-p} + x_{t-p+1} + \cdots + x_t + \cdots + x_{t+p}}{2p+1} \quad (8)$$

for $k = 2p + 1$,

$$y_t := \frac{x_{t-p/2} + x_{t-p/2+1} + \cdots + x_{t+p/2-1} + x_{t+p/2}}{2p} \quad (9)$$

for $k = 2p$.

Forecasting:

- Step 1: linear regression of y_t as a function of t ,

$$\hat{y}_t = \hat{a}t + \hat{b}.$$

- Step 2: estimate seasonal values,

$$\hat{s}_t = x_t - \hat{T}_t = x_t - \hat{y}_t$$

- Step 3: because we have estimated s_t several times, we take the average

$$\bar{\hat{s}}_t := \frac{1}{I} \sum_{i=1}^I \hat{s}_{t+ik}.$$

Thus

$$\hat{x}_t = \hat{y}_t + \bar{\hat{s}}_t.$$

2.2.3 Two nominal variables

From the contingency table (cf 1.4) we may test whether two variables X and Y are independent or not with the χ^2 test of Independence's:

$$\chi^2 = \sum_{l=1}^L \sum_{c=1}^C \frac{(n_{lc} - E_{lc})^2}{E_{lc}}$$

where

$$\begin{aligned} n_{lc} &:= \text{the observed number} \\ E_{lc} &:= \frac{n_{l.} n_{.c}}{n}. \end{aligned}$$

Note that E_{lc} is the expected number if the variables are independent.

Remark 2 1. If X and Y have no interdependence, then one expects that $\chi^2 = 0$.

2. χ^2 depends on n , L , and C and $\lim_{n \rightarrow \infty} \chi^2$ is a known distribution.

If $C + L \geq 2$, we have the following bounds on χ^2 :

$$0 \leq \chi^2 \leq n[\min(L, C) - 1].$$

For large n , we approximate χ^2 with the limit distribution to test the hypothesis:

H_0 : X and Y are independent.

H_0 is rejected if the χ^2 statistic is sufficiently large. More precisely, we fix our tolerance of error, usually at either 5% or 1%. A bound is computed such that the probability of χ^2 exceeding the bound under H_0 is our tolerated error. If this bound is exceeded by our computed statistic H_0 is rejected.