Introduction to Probability and Statistics

Michael P. Wiper,
Universidad Carlos III de Madrid

Course objectives

Starting from first principles, we shall firstly review the main properties of probability and random variables and their properties. In particular, we shall introduce the probability and moment generating functions. Secondly, we shall analyze the different methods of collecting, displaying and summarizing data samples. This course should provide the basic knowledge necessary for the first term course in Statistics.

Recommended reading

- CM Grinstead and JM Snell (1997). *Introduction to Probability*, AMS. Available from:

  http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/pdf.html

- Online Statistics: An Interactive Multimedia Course of Study is a good online course at:

  http://onlinestatbook.com/

- MP Wiper (2006). Here are some notes on probability from an elementary course.

  http://halweb.uc3m.es/esp/Personal/personas/mwiper/docencia/Spanish/Doctorado_EEMC/probability_class.pdf

Index

- Probability and random variables:
  - Mathematical probability and the Kolmogorov axioms.
  - Different interpretations of probability.
  - Conditional probability and Bayes theorem.
  - Random variables and their characteristics.
  - Generating functions.

- Descriptive statistics:
  - Sampling.
  - Different types of data.
  - Displaying a sample of data.
  - Sample moments.
  - Bivariate samples and regression.

Probability

Chance is a part of our everyday lives. Everyday we make judgements based on probability:

- There is a 90% chance Real Madrid will win tomorrow.
- There is a $1/6$ chance that a dice toss will be a 3.

Probability Theory was developed from the study of games of chance by Fermat and Pascal and is the mathematical study of randomness. This theory deals with the possible outcomes of an event and was put onto a firm mathematical basis by Kolmogorov.

The Kolmogorov axioms

Kolmogorov

For a *random experiment* with *sample space* $\mathbf{\Omega}$, then a probability measure P is a function such that

1. for any event $A \in \mathbf{\Omega}$, $P(A) \geq 0$.

2. $P(\mathbf{\Omega}) = 1$.

3. $P\left(\cup_{j \in J} A_j\right) = \sum_{j \in J} P(A_j)$ if $\{A_j : j \in J\}$ is a countable set of incompatible events.

Laws of probability

The basic laws of probability can be derived directly from set theory and the Kolmogorov axioms. For example, for any two events A and B, we have the addition law,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Laws of probability

The basic laws of probability can be derived directly from set theory and the Kolmogorov axioms. For example, for any two events A and B, we have the addition law,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof

$$\begin{aligned}
A &= (A \cap B) \cup (A \cap \bar{B}) \quad \text{so} \\
P(A) &= P(A \cap B) + P(A \cap \bar{B}) \quad \text{and similarly for } B. \text{ Also,} \\
A \cup B &= (A \cap \bar{B}) \cup (B \cap \bar{A}) \cup (A \cap B) \quad \text{so} \\
P(A \cup B) &= P(A \cap \bar{B}) + P(B \cap \bar{A}) + P(A \cap B) \\
&= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B).
\end{aligned}$$

Partitions

The previous example is easily extended when we have a sequence of events, $A_1, A_2, \ldots, A_n$, that form a *partition*, that is

$$\bigcup_{i=1}^{n} A_i = \Omega, \quad A_i \cap A_j = \phi \text{ for all } i \neq j.$$

In this case,

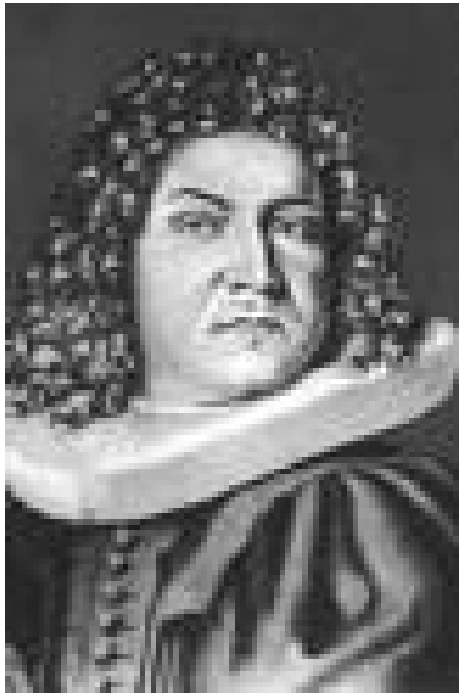
$$\begin{aligned} P\left(\cup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{j>i=1}^{n} P(A_i \cap A_j) + \sum_{k>j>i=1}^{n} P(A_i \cap A_j \cap A_k) + \ldots \\ &\quad +(-1)^n P(A_1 \cap A_2 \cap \ldots \cap A_n). \end{aligned}$$

Interpretations of probability

The Kolmogorov axioms provide a mathematical basis for probability but don't provide for a real life interpretation. Various ways of interpreting probability in real life situations have been proposed.

- The classical interpretation.
- Frequentist probability.
- Subjective probability.
- Other approaches; logical probability and propensities.

Classical probability



Bernoulli

This derives from the ideas of Jakob Bernoulli (1713) contained in the *principle of insufficient reason* (or *principle of indifference*) developed by Laplace (1812) which can be used to provide a way of assigning epistemic or subjective probabilities.

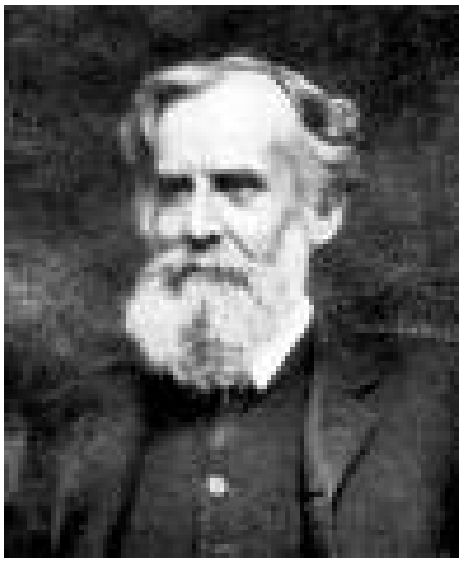
The principle of insufficient reason

If we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way.

Thus the probability of an event is the coefficient between the number of favourable cases and the total number of possible cases.

This is a very limited definition and cannot be easily applied in infinite dimensional or continuous sample spaces.

Frequentist probability



Venn

Von Mises

The idea comes from Venn (1876) and von Mises (1919).

Given a repeatable experiment, the probability of an event is defined to be the limit of the proportion of times that the event will occur when the number of repetitions of the experiment tends to infinity.

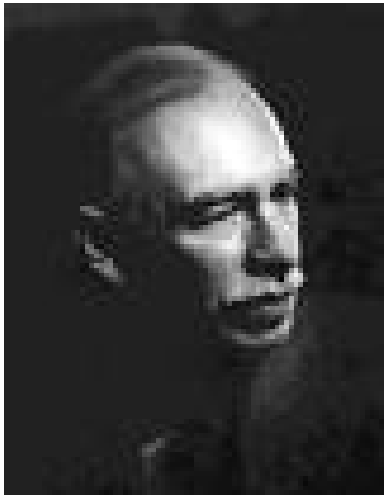
This is a restricted definition of probability. It is impossible to assign probabilities in non repeatable experiments.

Subjective probability

Ramsey

A different approach uses the concept of ones own probability as a subjective measure of ones own uncertainty about the occurrence of an event. Thus, we may all have different probabilities for the same event because we all have different experience and knowledge. This approach is more general than the other methods as we can now define probabilities for unrepeatable experiments. Subjective probability is studied in detail in Bayesian Statistics.

Other approaches



Keynes

- *Logical probability* was developed by Keynes (1921) and Carnap (1950) as an extension of the classical concept of probability. The (conditional) probability of a proposition H given evidence E is interpreted as the (unique) degree to which E logically entails H.

Popper

- Under the theory of *propensities* developed by Popper (1957), probability is an innate disposition or propensity for things to happen. Long run propensities seem to coincide with the frequentist definition of probability although it is not clear what individual propensities are, or whether they obey the probability calculus.

Conditional probability and independence

The probability of an event B conditional on an event A is defined as

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

This can be interpreted as the probability of B given that A occurs.

Two events A and B are called *independent* if $P(A \cap B) = P(A)P(B)$ or equivalently if $P(B|A) = P(B)$ or $P(A|B) = P(A)$.

The multiplication law

A restatement of the conditional probability formula is the *multiplication law*

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

Example 1

What is the probability of getting two cups in two draws from a Spanish pack of cards?

Write C_i for the event that draw i is a cup for $i = 1, 2$. Enumerating all the draws with two cups is not entirely trivial. However, the conditional probabilities are easy to calculate:

$$P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) = \frac{9}{39} \times \frac{10}{40} = \frac{3}{52}.$$

The multiplication law can be extended to more than two events. For example,

$$P(A \cap B \cap C) = P(C|A, B)P(B|A)P(A).$$

The birthday problem

Example 2

What is the probability that among n students in a classroom, at least two will have the same birthday?

http://webspace.ship.edu/deensley/mathdl/stats/Birthday.html

The birthday problem

Example 2

What is the probability that among n students in a classroom, at least two will have the same birthday?

http://webspace.ship.edu/deensley/mathdl/stats/Birthday.html

The solution is not obvious but can be solved using conditional probability. Let b_i be the birthday of student i, for $i = 1, \ldots, n$. Then it is easiest to calculate the probability that all birthdays are distinct

$$\begin{aligned} P(b_1 \neq b_2 \neq \ldots \neq b_n) &= P(b_n \notin \{b_1, \ldots, b_{n-1}\} | b_1 \neq b_2 \neq \ldots b_{n-1}) \times \\ & P(b_{n-1} \notin \{b_1, \ldots, b_{n-2}\} | b_1 \neq b_2 \neq \ldots b_{n-2}) \times \cdots \\ & \times P(b_3 \notin \{b_1, b_2\} | b_1 \neq b_2) P(b_1 \neq b_2) \end{aligned}$$

Now clearly,

$$P(b_1 \neq b_2) = \frac{364}{365}, \quad P(b_3 \notin \{b_1, b_2\} | b_1 \neq b_2) = \frac{363}{365}$$

and similarly

$$P(b_i \notin \{b_1, \ldots, b_{i-1}\} | b_1 \neq b_2 \neq \ldots b_{i-1}) = \frac{366 - i}{365}$$

for $i = 3, \ldots, n$.

Thus, the probability that at least two students have the same birthday is, for $n < 365$,

$$1 - \frac{364}{365} \times \cdots \times \frac{366 - n}{365} = \frac{365!}{365^n (365 - n)!}.$$

For $n = 23$, this probability is greater than 0.5 and for $n > 50$, it is virtually one.

The law of total probability

The simplest version of this rule is the following.

Theorem 1
For any two events A and B, then

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}).$$

We can also extend the law to the case where $A_1, \ldots, A_n$ form a partition. In this case, we have

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

Bayes theorem

Theorem 2

For any two events A and B, then

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

Supposing that $A_1, \ldots, A_n$ form a partition, using the law of total probability, we can write Bayes theorem as

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \quad \text{for } j = 1, \ldots, n.$$

The Monty Hall problem

Example 3

The following statement of the problem was given in a column by Marilyn vos Savant in a column in Parade magazine in 1990.

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Simulating the game

Have a look at the following web page.

http://www.stat.sc.edu/~west/javahtml/LetsMakeaDeal.html

Simulating the game

Have a look at the following web page.

http://www.stat.sc.edu/~west/javahtml/LetsMakeaDeal.html

Using Bayes theorem

http://en.wikipedia.org/wiki/Monty_Hall_problem

Random variables

A random variable generalizes the idea of probabilities for events. Formally, a random variable, X simply assigns a numerical value, x_i to each event, A_i, in the sample space, Ω. For mathematicians, we can write X in terms of a mapping, $X : \Omega \rightarrow \mathbb{R}$.

Random variables may be classified according to the values they take as

- discrete
- continuous
- mixed

Discrete variables

Discrete variables are those which take a discrete set range of values, say $\{x_1, x_2, \ldots\}$. For such variables, we can define the *cumulative distribution function*,

$$F_X(x) = P(X \leq x) = \sum_{i, x_i \leq x} P(X = x_i)$$

where $P(X = x)$ is the *probability function* or *mass function*.

For a discrete variable, the *mode* is defined to be the point, $\hat{x}$, with maximum probability, i.e. such that

$$P(X = x) < P(X = \hat{x}) \text{for all } x \neq \hat{x}.$$

Moments

For any discrete variable, X, we can define the mean of X to be

$$\mu_X = E[X] = \sum_i x_i P(X = x_i).$$

Recalling the frequency definition of probability, we can interpret the mean as the limiting value of the sample mean from this distribution. Thus, this is a measure of location.

In general we can define the expectation of any function, $g(X)$ as

$$E[g(X)] = \sum_i g(x_i) P(X = x_i).$$

In particular, the variance is defined as

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

and the standard deviation is simply $\sigma = \sqrt{\sigma^2}$. This is a measure of spread.

Chebyshev's inequality

It is interesting to analyze the probability of being close or far away from the mean of a distribution. *Chebyshev's inequality* provides loose bounds which are valid for any distribution with finite mean and variance.

Theorem 3

For any random variable, X, with finite mean, μ, and variance, σ^2, then for any $k > 0$,

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

Thus, we know that $P(\mu - \sqrt{2}\sigma \leq X \leq \mu + \sqrt{2}\sigma) \geq 0.5$ for any variable X.

Proof

$$
\begin{aligned}
P(|X-\mu| \geq k\sigma) &= P\left((X-\mu)^2 \geq k^2\sigma^2\right) \\
&= P\left(\left(\frac{X-\mu}{k\sigma}\right)^2 \geq 1\right) \\
&= E\left[I_{\left(\frac{X-\mu}{k\sigma}\right)^2 \geq 1}\right] \quad \text{where } I \text{ is an indicator function} \\
&\leq E\left[\left(\frac{X-\mu}{k\sigma}\right)^2\right] = \frac{1}{k^2} \quad \blacksquare
\end{aligned}
$$

Important discrete distributions

The binomial distribution

Let X be the number of heads in n independent tosses of a coin such that $P(\text{head}) = p$. Then X has a binomial distribution with parameters n and p and we write $X \sim \mathcal{BI}(n, p)$. The mass function is

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

The mean and variance of X are np and $np(1-p)$ respectively.

The geometric distribution

Suppose that Y is defined to be the number of tails observed before the first head occurs for the same coin. Then Y has a geometric distribution with parameter p, i.e. $Y \sim \mathcal{GE}(p)$ and

$$P(Y = y) = p(1-p)^y \quad \text{for } y = 0, 1, 2, \ldots$$

The mean any variance of X are $\frac{1-p}{p}$ and $\frac{1-p}{p^2}$ respectively.

The negative binomial distribution

A generalization of the geometric distribution is the negative binomial distribution. If we define Z to be the number of tails observed before the r'th head is observed, then $Z \sim \mathcal{NB}(r, p)$ and

$$P(Z = z) = \begin{pmatrix} r+z-1 \\ z \end{pmatrix} p^r (1-p)^z \quad \text{for } z = 0, 1, 2, \ldots$$

The mean and variance of X are $r\frac{1-p}{p}$ and $r\frac{1-p}{p^2}$ respectively.

The negative binomial distribution reduces to the geometric model for the case $r = 1$.

The hypergeometric distribution

Suppose that a pack of N cards contains R red cards and that we deal n cards without replacement. Let X be the number of red cards dealt. Then X has a hypergeometric distribution with parameters N, R, n, i.e. $X \sim \mathcal{HG}(N, R, n)$ and

$$P(X = x) = \frac{\begin{pmatrix} R \\ x \end{pmatrix} \begin{pmatrix} N - R \\ n - x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad \text{for } x = 0, 1, \ldots, n.$$

Example 4

In the Primitiva lottery, a contestant chooses 6 numbers from 1 to 49 and 6 numbers are drawn without replacement. The contestant wins the grand prize if all numbers match. The probability of winning is thus

$$P(X = x) = \frac{\begin{pmatrix} 6 \\ 6 \end{pmatrix} \begin{pmatrix} 43 \\ 0 \end{pmatrix}}{\begin{pmatrix} 49 \\ 6 \end{pmatrix}} = \frac{6!43!}{49!} = \frac{1}{13983816}.$$

What if N and R are large?

For large N and R, then the factorials in the hypergeometric probability expression are often hard to evaluate.

Example 5

Suppose that $N = 2000$ and $R = 500$ and $n = 20$ and that we wish to find $P(X = 5)$. Then the calculation of $2000!$ for example is very difficult.

What if N and R are large?

For large N and R, then the factorials in the hypergeometric probability expression are often hard to evaluate.

Example 5

Suppose that $N = 2000$ and $R = 500$ and $n = 20$ and that we wish to find $P(X = 5)$. Then the calculation of $2000!$ for example is very difficult.

Theorem 4

Let $X \sim \mathcal{HG}(N, R, n)$ and suppose that $R, N \to \infty$ and $R/N \to p$. Then

$$P(X = x) \to \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

Proof

$$
\begin{aligned}
P(X=x) &= \frac{\left(\begin{array}{c} R \\ x \end{array}\right)\left(\begin{array}{c} N-R \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)} = \frac{\left(\begin{array}{c} n \\ x \end{array}\right)\left(\begin{array}{c} N-n \\ R-x \end{array}\right)}{\left(\begin{array}{c} N \\ R \end{array}\right)} \\
&= \left(\begin{array}{c} n \\ x \end{array}\right) \frac{R!(N-R)!(N-n)!}{(R-x)!(N-R-n+x)!N!} \\
&\to \left(\begin{array}{c} n \\ x \end{array}\right) \frac{R^x(N-R)^{n-x}}{N^n} \to \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}
\end{aligned}
$$

In the example,$p = 500/2000 = 0.25$ and using a binomial approximation, $P(X=5) \approx \left(\begin{array}{c} 20 \\ 5 \end{array}\right) 0.25^5 0.75^{15} = 0.2023$. The exact answer, from Matlab is 0.2024.

The Poisson distribution

Assume that rare events occur on average at a rate λ per hour. Then we can often assume that the number of rare events X that occur in a time period of length t has a Poisson distribution with parameter (mean and variance) λt, i.e. $X \sim \mathcal{P}(\lambda t)$. Then

$$P(X = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

The Poisson distribution

Assume that rare events occur on average at a rate λ per hour. Then we can often assume that the number of rare events X that occur in a time period of length t has a Poisson distribution with parameter (mean and variance) λt, i.e. $X \sim \mathcal{P}(\lambda t)$. Then

$$P(X = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

Formally, the conditions for a Poisson distribution are

- The numbers of events occurring in non-overlapping intervals are independent for all intervals.
- The probability that a single event occurs in a sufficiently small interval of length h is $\lambda h + o(h)$.
- The probability of more than one event in such an interval is $o(h)$.

Continuous variables

Continuous variables are those which can take values in a continuum. For a continuous variable, X, we can still define the distribution function, $F_X(x) = P(X \leq x)$ but we cannot define a probability function $P(X = x)$. Instead, we have the density function

$$f_X(x) = \frac{dF(x)}{dx}.$$

Thus, the distribution function can be derived from the density as $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$. In a similar way, moments of continuous variables can be defined as integrals,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

and the mode is defined to be the point of maximum density.

For a continuous variable, another measure of location is the *median*, $\tilde{x}$, defined so that $F_X(\tilde{x}) = 0.5$.

Important continuous variables

The uniform distribution

This is the simplest continuous distribution. A random variable, X, is said to have a uniform distribution with parameters a and b if

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a < x < b.$$

In this case, we write $X \sim \mathcal{U}(a, b)$ and the mean and variance of X are $\frac{a+b}{2}$ and $\frac{(b-a)^2}{12}$ respectively.

The exponential distribution

Remember that the Poisson distribution models the number of rare events occurring at rate λ in a given time period. In this scenario, consider the distribution of the time between any two successive events. This is an exponential random variable, $Y \sim \mathcal{E}(\lambda)$, with density function

$$f_Y(y) = \lambda e^{-\lambda y} \quad \text{for } y > 0.$$

The mean and variance of X are $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ respectively.

The normal distribution

This is probably the most important continuous distribution. A random variable, X, is said to follow a normal distribution with mean and variance parameters μ and σ^2 if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad \text{for } -\infty < x < \infty.$$

In this case, we write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

- If X is normally distributed, then $a + bX$ is normally distributed. In particular, $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
- $P(|X-\mu| \geq \sigma) = 0.3174$, $P(|X-\mu| \geq 2\sigma) = 0.0456$, $P(|X-\mu| \geq 3\sigma) = 0.0026$.
- Any sum of normally distributed variables is also normally distributed.

Example 6

Let $X \sim \mathcal{N}(2, 4)$. Find $P(3 < X < 4)$.

$$\begin{aligned} P(3 < X < 4) &= P\left(\frac{3-2}{\sqrt{4}} < \frac{X-2}{\sqrt{4}} < \frac{4-2}{\sqrt{4}}\right) \\ &= P(0.5 < Z < 1) \quad \text{where } Z \sim \mathcal{N}(0, 1) \\ &= P(Z < 1) - P(Z < 0.5) = 0.8413 - 0.6915 \\ &= 0.1499 \end{aligned}$$

The central limit theorem

One of the main reasons for the importance of the normal distribution is that it can be shown to approximate many real life situations due to the central limit theorem.

Theorem 5
Given a random sample of size $X_1, \ldots, X_n$ from some distribution, then under certain conditions, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows a normal distribution.

Proof See later. ∎

For an illustration of the CLT, see

http://cnx.rice.edu/content/m11186/latest/

Mixed variables

Occasionally it is possible to encounter variables which are partially discrete and partially continuous. For example, the time spent waiting for service by a customer arriving in a queue may be zero with positive probability (as the queue may be empty) and otherwise takes some positive value in $(0, \infty)$.

The probability generating function

For a discrete random variable, X, taking values in some subset of the non-negative integers, then the probability generating function, $G_X(s)$ is defined as

$$G_X(s) = E\left[s^X\right] = \sum_{x=0}^{\infty} P(X = x)s^x.$$

This function has a number of useful properties:

- $G(0) = P(X = 0)$ and more generally, $P(X = x) = \frac{1}{x!}\frac{d^x G(s)}{ds^x}\,|_{s=0}$.
- $G(1) = 1$, $E[X] = \frac{dG(1)}{ds}$ and more generally, the k'th factorial moment, $E[X(X-1)\cdots(X-k+1)]$, is

$$E\left[\frac{X!}{(X-k)!}\right] = \left.\frac{d^k G(s)}{ds^k}\right|_{s=1}$$

- The variance of X is

$$V[X] = G''(1) + G'(1) - G'(1)^2.$$

Example 7

Consider a negative binomial variable, $X \sim \mathcal{NB}(r, p)$.

$$\begin{aligned} P(X = x) &= \binom{r+x-1}{x} p^r (1-p)^x \quad \text{for } z = 0, 1, 2, \ldots \\ E[s^X] &= \sum_{x=0}^{\infty} s^x \binom{r+x-1}{x} p^r (1-p)^x \\ &= p^r \sum_{x=0}^{\infty} \binom{r+x-1}{x} \{(1-p)s\}^x = \left(\frac{p}{1-(1-p)s}\right)^r \end{aligned}$$

$$
\begin{aligned}
\frac{dE}{ds} &= \frac{rp^r(1-p)}{(1-(1-p)s)^{r+1}} \\
\left.\frac{dE}{ds}\right|_{s=1} &= r\frac{1-p}{p} = E[X] \\
\frac{d^2E}{ds^2} &= \frac{r(r+1)p^r(1-p)^2}{(1-(1-p)s)^{r+2}} \\
\left.\frac{d^2E}{ds^2}\right|_{s=1} &= r(r+1)\left(\frac{1-p}{p}\right)^2 = E[X(X-1)] \\
V[X] &= r(r+1)\left(\frac{1-p}{p}\right)^2 + r\frac{1-p}{p} - \left(r\frac{1-p}{p}\right)^2 \\
&= r\frac{1-p}{p^2}.
\end{aligned}
$$

The probability generating function for a sum of independent variables

Suppose that $X_1, \ldots, X_n$ are *independent* with generating functions $G_i(s)$ for $s = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then

$$\begin{aligned} G_Y(s) &= E\left[s^Y\right] \\ &= E\left[s^{\sum_{i=1}^{n} X_i}\right] \\ &= \prod_{i=1}^{n} E\left[s^{X_i}\right] \quad \text{by independence} \\ &= \prod_{i=1}^{n} G_i(s) \end{aligned}$$

Furthermore, if $X_1, \ldots, X_n$ are identically distributed, with common generating function $G_X(s)$, then

$$G_Y(s) = G_X(s)^n.$$

Example 8

Suppose that $X_1, \ldots, X_n$ are Bernoulli trials so that

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = 0) = 1 - p \quad \text{for } i = 1, \ldots, n$$

Then, the probability generating function for any X_i is $G_X(s) = 1 - p + sp$. Now consider a binomial random variable, $Y = \sum_{i=1}^{n} X_i$. Then

$$G_Y(s) = (1 - p + sp)^n$$

is the binomial probability generating function.

Another useful property of pgfs

If N is a discrete variable taking values on the non-negative integers and with pgf $G_N(s)$ and if $X_1, \ldots, X_N$ is a sequence of independent and identically distributed variables with pgf $G_X(s)$, then if $Y = \sum_{i=1}^{N} X_i$, we have

$$\begin{aligned} G_Y(s) &= E\left[s^{\sum_{i=1}^{N} X_i}\right] \\ &= E\left[E\left[s^{\sum_{i=1}^{N} X_i} \mid N\right]\right] \\ &= E\left[G_X(s)^N\right] \\ &= G_N(G_X(s)) \end{aligned}$$

This result is useful in the study of *branching processes*. See the course in Stochastic Processes.

The moment generating function

For any variable, X, the moment generating function of X is defined to be

$$M_X(s) = E\left[e^{sX}\right].$$

This generates the moments of X as we have

$$\begin{aligned} M_X(s) &= E\left[\sum_{i=1}^{\infty} \frac{(sX)^i}{i!}\right] \\ \left.\frac{d^i M_X(s)}{ds^i}\right|_{s=0} &= E\left[X^i\right] \end{aligned}$$

Example 9

Suppose that $X \sim \mathcal{G}(\alpha, \beta)$. Then

$$
\begin{aligned}
f_X(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0 \\
M_X(s) &= \int_0^\infty e^{sx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx \\
&= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-s)x}\, dx \\
&= \left(\frac{\beta}{\beta - s}\right)^\alpha \\
\frac{dM}{ds} &= \frac{\alpha\beta^\alpha}{(\beta - s)^{\alpha-1}} \\
\left.\frac{dM}{ds}\right|_{s=0} &= \frac{\alpha}{\beta} = E[X]
\end{aligned}
$$

Example 10

Suppose that $X \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}
M_X(s) &= \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[x^2 - 2s\right]\right)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[x^2 - 2s + s^2 - s^2\right]\right)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[(x-s)^2 - s^2\right]\right)\, dx \\
&= e^{\frac{s^2}{2}}.
\end{aligned}$$

The moment generating function of a sum of independent variables

Suppose we have a sequence of independent variables, $X_1, X_2, \ldots, X_n$ with mgfs $M_1(s), \ldots, M_n(s)$. Then, if $Y = \sum_{i=1}^{n} X_i$, it is easy to see that

$$M_Y(s) = \prod_{i=1}^{n} M_i(s)$$

and if the variables are identically distributed with common mgf $M_X(s)$, then

$$M_Y(s) = M_X(s)^n.$$

Example 11

Suppose that $X_i \sim \mathcal{E}(\lambda)$ for $i = 1, \ldots, n$ are independent. Then

$$\begin{aligned} M_X(s) &= \int_0^\infty e^{sx} \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda - s)x}\, dx \\ &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

Therefore the mgf of $Y = \sum_{i=1}^n X_i$ is given by

$$M_Y(s) = \left(\frac{\lambda}{\lambda - s} \right)^n$$

which we can recognize as the mgf of a gamma distribution, $Y \sim \mathcal{G}(n, \lambda)$.

Proof of the central limit theorem

For any variable, Y, with zero mean and unit variance and *such that all moments exist*, then the moment generating function is

$$M_Y(s) = E[e^{sY}] = 1 + \frac{s^2}{2} + o(s^2).$$

Now assume that $X_1, \ldots, X_n$ are a random sample from a distribution with mean μ and variance σ^2. Then, we can define the standardized variables, $Y_i = \frac{X_i - \mu}{\sigma}$, which have mean 0 and variance 1 for $i = 1, \ldots, n$ and then

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$$

Now, suppose that $M_Y(s)$ is the mgf of Y_i, for $i = 1, \ldots, n$. Then

$$M_{Z_n}(s) = M_Y\left(s/\sqrt{n}\right)^n$$

and therefore,

$$M_{Z_n}(s) = \left(1 + \frac{s^2}{2n} + o(s^2/n)\right)^n \to e^{\frac{s^2}{2}}$$

which is the mgf of a normally distributed random variable.

and therefore,

$$M_{Z_n}(s) = \left(1 + \frac{s^2}{2n} + o(s^2/n)\right)^n \to e^{\frac{s^2}{2}}$$

which is the mgf of a normally distributed random variable.

To make this result valid for variables that do not necessarily possess all their moments, then we can use essentially the same arguments but defining the characteristic function $C_X(s) = E[e^{isX}]$ instead of the moment generating function.

Multivariate distributions

It is straightforward to extend the concept of a random variable to the multivariate case. Full details are included in the course on Multivariate Analysis.

For two discrete variables, X and Y, we can define the joint probability function at $(X = x, Y = y)$ to be $P(X = x, Y = y)$ and in the continuous case, we similarly define a joint density function $f_{X,Y}(x, y)$ such that

$$\begin{aligned}
\sum_x \sum_y P(X = x, Y = y) &= 1 \\
\sum_y P(X = x, Y = y) &= P(X = x) \\
\sum_x P(X = x, Y = y) &= P(Y = y)
\end{aligned}$$

and similarly for the continuous case.

Conditional distributions

The conditional distribution of Y given $X = x$ is defined to be

$$f_{Y|x}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Two variables are said to be *independent* if for all x, y, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ or equivalently if $f_{Y|x}(y|x) = f_Y(y)$ or $f_{X|y}(x|y) = f_X(x)$.

We can also define the conditional expectation of $Y|x$ to be $E[Y|x] = \int y f_{Y|x}(y|x)\,dx$.

Covariance and correlation

It is useful to obtain a measure of the degree of relation between the two variables. Such a measure is the *correlation*.

We can define the expectation of any function, $g(X, Y)$, in a similar way to the univariate case,

$$E[g(X,Y)] = \int \int g(x,y) f_{X,Y}(x,y)\, dx\, dy.$$

In particular, the *covariance* is defined as

$$\sigma_{X,Y} = Cov[X,Y] = E[XY] - E[X]E[Y].$$

Obviously, the units of the covariance are the product of the units of X and Y. A scale free measure is the *correlation*,

$$\rho_{X,Y} = Corr[X,Y] = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Properties of the correlation are as follows:

- $-1 \leq \rho_{X,Y} \leq 1$
- $\rho_{X,Y} = 0$ if X and Y are independent. (This is not necessarily true in reverse!)
- $\rho_{X_Y} = 1$ if there is an exact, positive relation between X and Y so that $Y = a + bX$ where $b > 0$.
- $\rho_{X_Y} = -1$ if there is an exact, negative relation between X and Y so that $Y = a + bX$ where $b < 0$.

Theorem 6

For two variables, X and Y, then

$$\begin{aligned} E[Y] &= E[E[Y|X]] \\ V[Y] &= E[V[Y|X]] + V[E[Y|X]] \end{aligned}$$

Proof

$$\begin{aligned} E[E[Y|X]] &= E\left[\int y f_{Y|X}(y|X)\,dy\right] = \int f_X(x) \int y f_{Y|X}(y|X)\,dy\,dx \\ &= \int y \int f_{Y|X}(y|x) f_X(x)\,dx\,dy \\ &= \int y \int f_{X,Y}(x,y)\,dx\,dy \\ &= \int y f_Y(y)\,dy = E[Y] \quad \blacksquare \end{aligned}$$

Example 12

A random variable X has a beta distribution, $X \sim \mathcal{B}(\alpha, \beta)$, if

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 < x < 1.$$

The mean of X is $E[X] = \frac{\alpha}{\alpha+\beta}$.

Suppose now that we toss a coin with probability $P(\text{heads}) = X$ a total of n times and that we require the distribution of the number of heads, Y.

This is the beta-binomial distribution which is quite complicated:

$$
\begin{aligned}
P(Y=y) &= \int_0^1 P(Y=y|X=x) f_X(x)\,dx \\
&= \int_0^1 \left(\begin{array}{c} n \\ y \end{array} \right) x^y (1-x)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \left(\begin{array}{c} n \\ y \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+y-1}(1-x)^{\beta+n-y-1}\,dx \\
&= \left(\begin{array}{c} n \\ y \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}
\end{aligned}
$$

for $y = 0, 1, \ldots, n$.

We could try to calculate the mean of Y directly using the above probability function. However, this would be very complicated. There is a much easier way.

We could try to calculate the mean of Y directly using the above probability function. However, this would be very complicated. There is a much easier way.

$$\begin{aligned} E[Y] &= E[E[Y|X]] \\ &= E[nX] \quad \text{because } Y|X \sim \mathcal{BI}(n, X) \\ &= n\frac{\alpha}{\alpha+\beta}. \end{aligned}$$

Statistics

Statistics is the science of data analysis. This is concerned with

- how to generate suitable samples of data
- how to summarize samples of data to illustrate their important features
- how to make inference about populations given sample data.

Sampling

In statistical problems we usually wish to study the characteristics of some *population*. However, it is usually impossible to measure the values of the variables of interest for all members of the population. This implies the use of a *sample*.

There are many possible ways of selecting a sample. Non random approaches include:

- Convenience sampling
- Volunteer sampling
- Quota sampling

Such approaches can suffer from induced *biases*.

Random sampling

A better approach is *random sampling*. For a population of elements, say $e_1, \ldots, e_N$, then a simple random sample of size n selects every possible n-tuple of elements with equal probability. Unrepresentative samples can be selected by this approach, but is no *a priori* bias which means that this is likely.

When the population is large or heterogeneous, other random sampling approaches may be preferred. For example:

- Systematic sampling,
- Stratified sampling
- Cluster sampling
- Multi stage sampling

Sampling theory is studied in more detail in Quantitative Methods.

Descriptive statistics

Given a data sample, it is important to develop methods to summarize the important features of the data both visually and numerically. Different approaches should be used for different types of data.

- Categorical data:
  - Nominal data,
  - Ordinal data.
- Numerical data:
  - Discrete data,
  - Continuous data.

Categorical data

Categorical data are those that take values in different categories, e.g. blood types, favourite colours, etc. These data may be *nominal*, when the different categories have no inherent sense of order or *ordinal*, when the categories are naturally ordered.

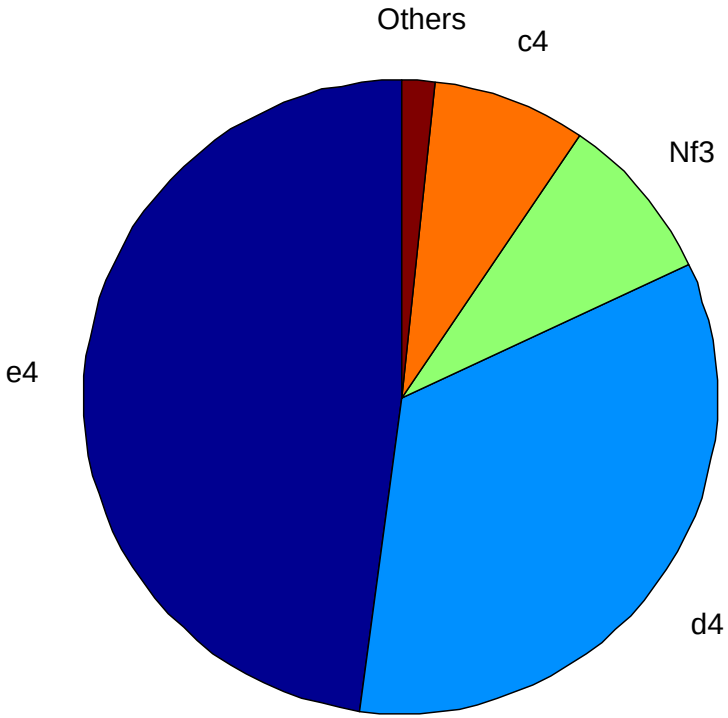
Example 13

The following table gives the frequencies of the different first moves in a chess game found on 20/02/1996 using the search engine of http://www.chessgames.com/.

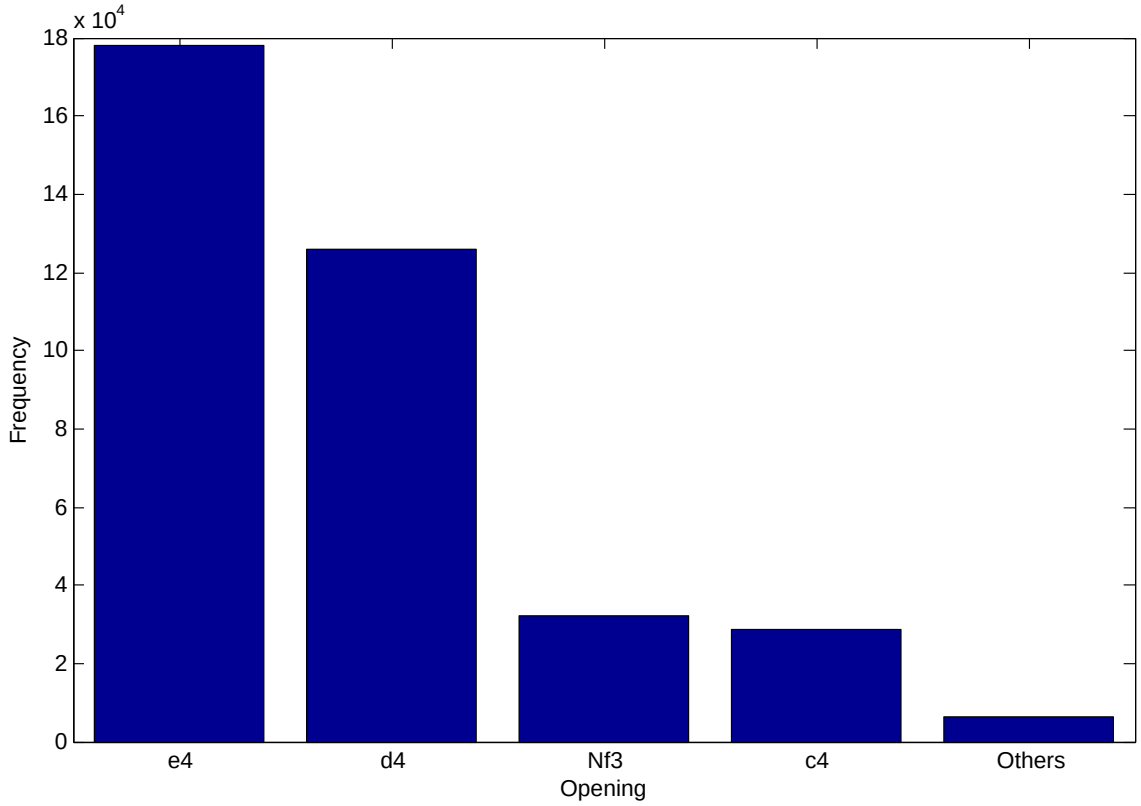
Opening	Frequency	Relative frequency
$e4$	178130	0.4794
$d4$	125919	0.3389
$Nf3$	32206	0.0867
$c4$	28796	0.0776
$Others$	6480	0.0174
$Total$	371531	1.0000

This is an example of a sample of nominal data. The frequency table has been augmented with the relative frequencies or proportions in each class. We can see immediately that the most popular opening or $modal\ class$ is $e4$, played in nearly half the games.

A nice way of visualizing the data is via a $pie\ chart$. This could be augmented with the frequencies or relative frequencies in each class.

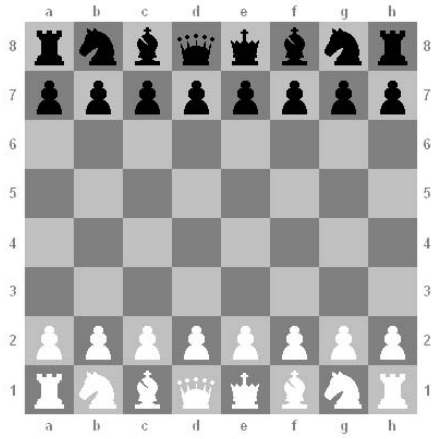


An alternative display is a bar chart which can be constructed using frequencies or relative frequencies.

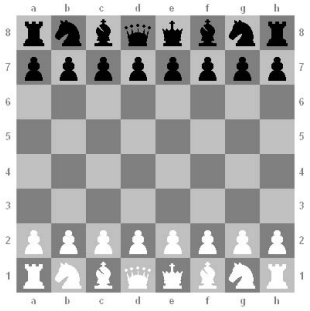


When the data are categorical, it is usual to order them from highest to lowest frequency. With ordinal data, it is more sensible to use the ordering of the classes.

A final approach which is good to look at but not so easy to interpret is the *pictogram*. The area of each image is proportional to the frequency.



$e4$



$d4$

$Nf3$

$c4$

Others

Measuring the relation between two categorical variables

Often we may record the values of two (or more) categorical variables. In such cases we are interested in whether or not there is any relation between the variables. To do this, we can construct a *contingency table*.

Example 14

The following data given in Morrell (1999) come from a South African study of single birth children. At birth in 1990 it was recorded whether or not the mothers received medical aid and later, in 1995 the researchers attempted to trace the children. Those children found were included in the five year group for further study.

	Children not traced	Five-Year Group	
Had Medical Aid	195	46	
No Medical Aid	979	370	
			1590

CH Morrell (1999). Simpson's Paradox: An Example From a Longitudinal Study in South Africa. *Journal of Statistics Education*, **7**.

Analysis of a contingency table

In order to analyze the contingency table it is useful to first calculate the marginal totals.

	Children not traced	Five-Year Group	
Had Medical Aid	195	46	241
No Medical Aid	979	370	1349
	1174	416	1590

and then to convert the original data into percentages.

	Children not traced	Five-Year Group	
Had Medical Aid	.123	.029	.152
No Medical Aid	.615	.133	.848
	.738	.262	1

Then it is also possible to calculate conditional frequencies. For example, the proportion of children not traced who received medical aid is

$$195/1174 = .123/.738 = .166.$$

Finally, we may often wish to assess whether there exists any relation between the two variables. In order to do this we can assess how many data we would expect to see in each cell, assuming the marginal totals if the data really were *independent*.

	Children not traced	Five-Year Group	
Had Medical Aid	177.95	63.05	241
No Medical Aid	996.05	352.95	1349
	1174	416	1590

Comparing these expected totals with the original frequencies, we could set up a formal statistical (χ^2) test for independence.

Simpson's paradox

Sometimes we can observe apparently paradoxical results when a population which contains heterogeneous groups is subdivided. The following example of the so called *Simpson's paradox* comes from the same study.

http://www.amstat.org/publications/jse/secure/v7n3/datasets.morrell.cfm

Numerical data

When data are naturally numerical, we can use both graphical and numerical approaches to summarize their important characteristics. For discrete data, we can frequency tables and bar charts in a similar way to the categorical case.

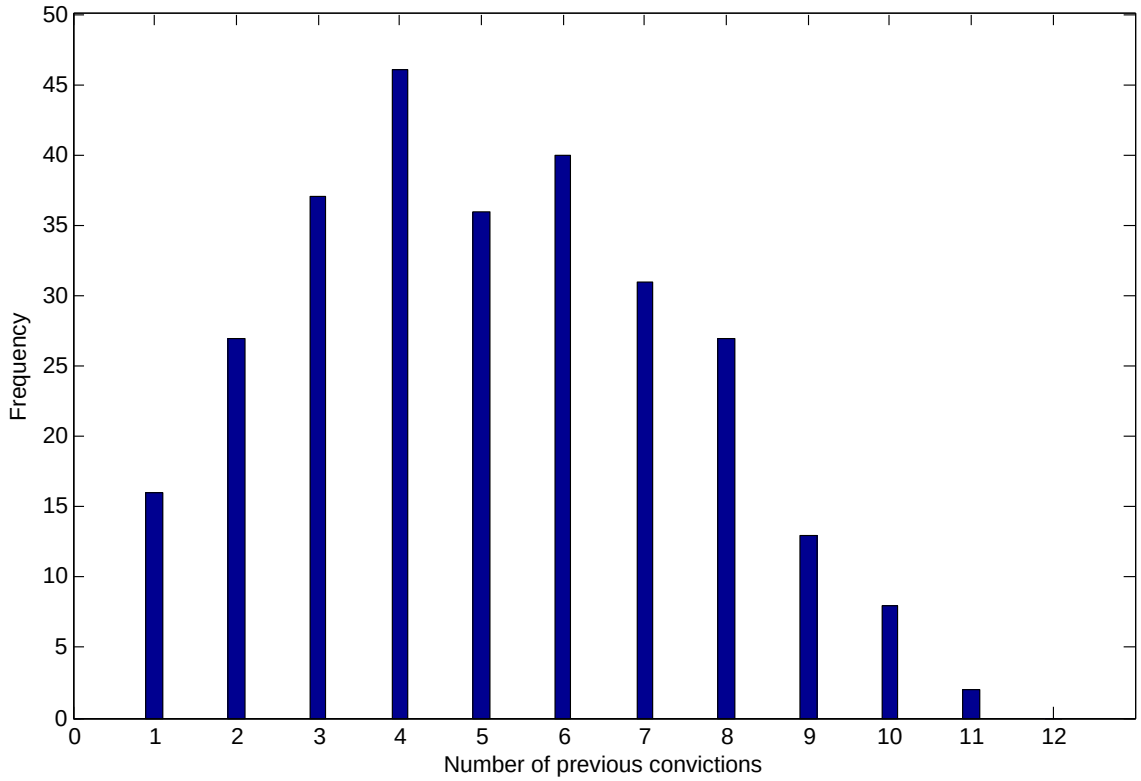
Example 15

The table reports the number of previous convictions for 283 adult males arrested for felonies in the USA taken from Holland et al (1981).

# Previous convictions	Frequency	Rel. freq.	Cum. freq.	Cum. rel. freq.
0	0	0.0000	0	0.0000
1	16	0.0565	16	0.0565
2	27	0.0954	43	0.1519
3	37	0.1307	80	0.2827
4	46	0.1625	126	0.4452
5	36	0.1272	162	0.5724
6	40	0.1413	202	0.7138
7	31	0.1095	233	0.8233
8	27	0.0954	260	0.9187
9	13	0.0459	273	0.9647
10	8	0.0283	281	0.9929
11	2	0.0071	283	1.0000
> 11	0	0.0000	283	1.0000

TR Holland, M Levi & GE Beckett (1981). Associations Between Violent And Nonviolent Criminality: A Canonical Contingency-Table Analysis. *Multivariate Behavioral Research*, **16**, 237–241.

Note that we have augmented the table with relative frequencies and cumulative frequencies. We can construct a bar chart of frequencies as earlier but we could also use cumulative or relative frequencies.



Note that the bars are much thinner in this case. Also, we can see that the distribution is slightly *positively skewed* or skewed to the right.

Continuous data and histograms

With continuous data, we should use histograms instead of bar charts. The main difficulty is in choosing the number of classes. We can see the effects of choosing different bar widths in the following web page.

http://www.shodor.org/interactivate/activities/histogram/

An empirical rule is to choose around $\sqrt{n}$ classes where n is the number of data. Similar rules are used by the main statistical packages.

It is also possible to illustrate the differences between two groups of individuals using histograms. Here, we should use the same classes for both groups.

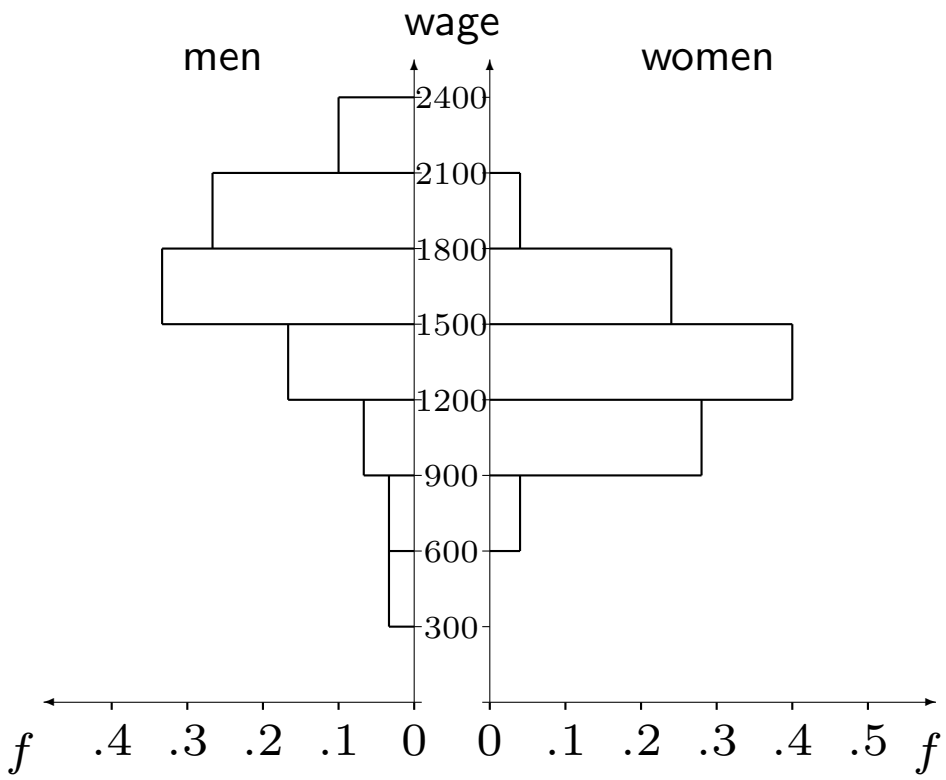
Example 16

The table summarizes the hourly wage levels of 30 Spanish men and 25 Spanish women, (with at least secondary education) who work $>$ 15 hours per week.

	M		W	
Interval	n_i	f_i	n_i	f_i
$[300, 600)$	1	.033	0	0
$[600, 900)$	1	.033	1	.04
$[900, 1200)$	2	.067	7	.28
$[1200, 1500)$	5	.167	10	.4
$[1500, 1800)$	10	.333	6	.24
$[1800, 2100)$	8	.267	1	.04
$[2100, 2400)$	3	.100	0	0
> 2400	0	0	0	0
	30	1	25	1

J Dolado and V LLorens (2004). Gender Wage Gaps by Education in Spain: Glass Floors vs. Glass Ceilings, *CEPR DP.*, **4203**.

http://www.eco.uc3m.es/temp/dollorens2.pdf



The male average wage is a little higher and the distribution of the male wages is more disperse and asymmetric.

Histograms with intervals of different widths

In this case, the histogram is constructed so that the area of each bar is proportional to the number of data.

Example 17

The following data are the results of a questionnaire to marijuana users concerning the weekly consumption of marijuana.

g / week	Frequency
$[0, 3)$	94
$[3, 11)$	269
$[11, 18)$	70
$[18, 25)$	48
$[25, 32)$	31
$[32, 39)$	10
$[39, 46)$	5
$[46, 74)$	2
> 74	0

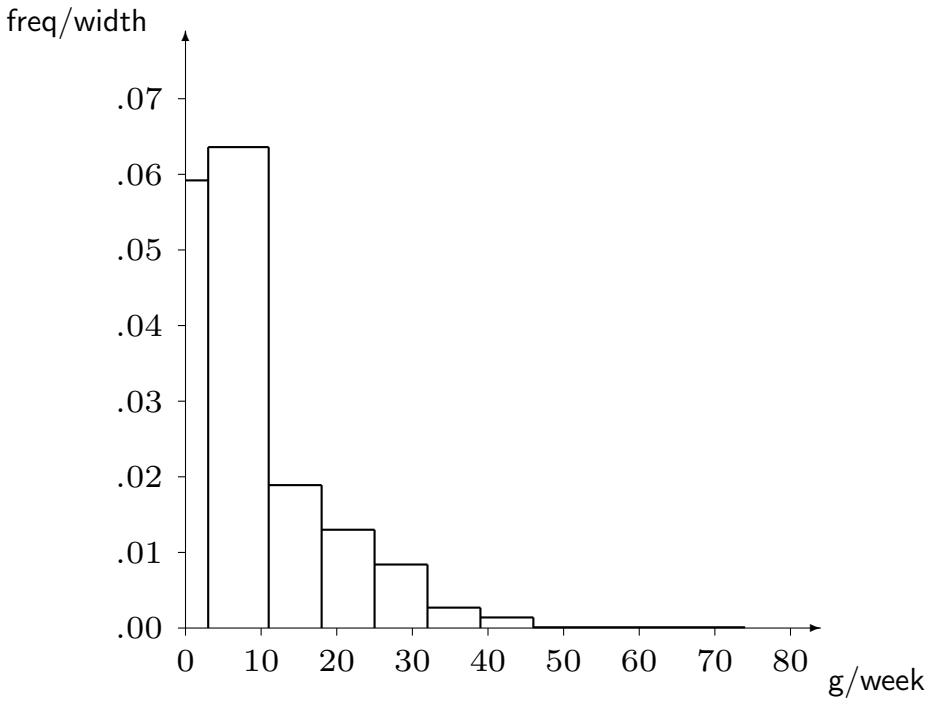
Landrigan et al (1983). Paraquat and marijuana: epidemiologic risk assessment. *Amer. J. Public Health*, **73**, 784-788

We augment the table with relative frequencies and bar widths and heights.

g / week	width	n_i	f_i	height
$[0, 3)$	3	94	.178	.0592
$[3, 11)$	8	269	.509	.0636
$[11, 18)$	7	70	.132	.0189
$[18, 25)$	7	48	.091	.0130
$[25, 32)$	7	31	.059	.0084
$[32, 39)$	7	10	.019	.0027
$[39, 46)$	7	5	.009	.0014
$[46, 74)$	28	2	.004	.0001
> 74	0	0	0	0
Total		529	1	

We use the formula

$$\textbf{height} = \textbf{frequency}/\textbf{interval width}$$



The distribution is very skewed to the right.

Other graphical methods

- *the frequency polygon*. A histogram is constructed and the bars are lines are used to join each bar at the centre. Usually the histogram is then removed. This simulates the probability density function.

- *the cumulative frequency polygon*. As above but using a cumulative frequency histogram and joining at the end of each bar.

- *the stem and leaf plot*. This is like a histogram but retaining the original numbers.

Sample moments

For a sample, $x_1, \ldots, x_n$ of numerical data, then the *sample mean* is defined as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the *sample variance* is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The sample standard deviation is $s = \sqrt{s^2}$.

The sample mean may be interpreted as an estimator of the population mean. It is easiest to see this if we consider grouped data say $x_1, \ldots, x_k$ where x_j is observed n_j times in total and $\sum_{i=1}^{k} n_i = n$. Then, the sample mean is

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j x_j = \sum_{j=1}^{k} f_j x_j$$

where f_j is the proportion of times that x_j was observed.

When $n \to \infty$, then (using the frequency definition of probability), we know that $f_j \to P(X = x_j)$ and so $\bar{x} \to \mu_X$, the true population mean.

Sometimes the sample variance is defined with a denominator of n instead of $n - 1$. However, in this case, it is a *biased* estimator of σ^2.

Problems with outliers, the median and interquartile range

The mean and standard deviation are good estimators of location and spread of the data if there are no outliers or if the sample is reasonably symmetric. Otherwise, it is better to use the *sample median* and *interquartile range*.

Assume that the sample data are ordered so that $x_1 \leq x_2 \leq \ldots \leq x_n$. Then the sample median is defined to be

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2} & \text{if } n \text{ is even.} \end{cases}$$

For example, if we have a sample $1, 2, 6, 7, 8, 9, 11$, the median is 7 and for the sample $1, 2, 6, 7, 8, 9, 11, 12$ then the median is 7.5. We can think of the median as dividing the sample in two.

We can also define the quartiles in a similar way. The lower quartile is $Q_1 = x_{\frac{n+1}{4}}$ and the upper quartile may be defined as $Q_3 = x_{\frac{3(n+1)}{4}}$ where if the fraction is not a whole number, the value should be derived by interpolation. Thus, for the sample $1, 2, 6, 7, 8, 9, 11$, then $Q_1 = 2$ and $Q_3 = 9$. For the sample $1, 2, 6, 7, 8, 9, 11, 12$, we have $\frac{n+1}{4} = 2.25$ so that

$$Q_1 = 2 + 0.25(6 - 2) = 3$$

and $\frac{3(n+1)}{4} = 6.75$ so

$$Q_3 = 9 + 0.75(11 - 9) = 10.5.$$

A nice visual summary of a data sample using the median, quartiles and range of the data is the so called *box and whisker plot* or boxplot.

http://en.wikipedia.org/wiki/Box_plot

Correlation and regression

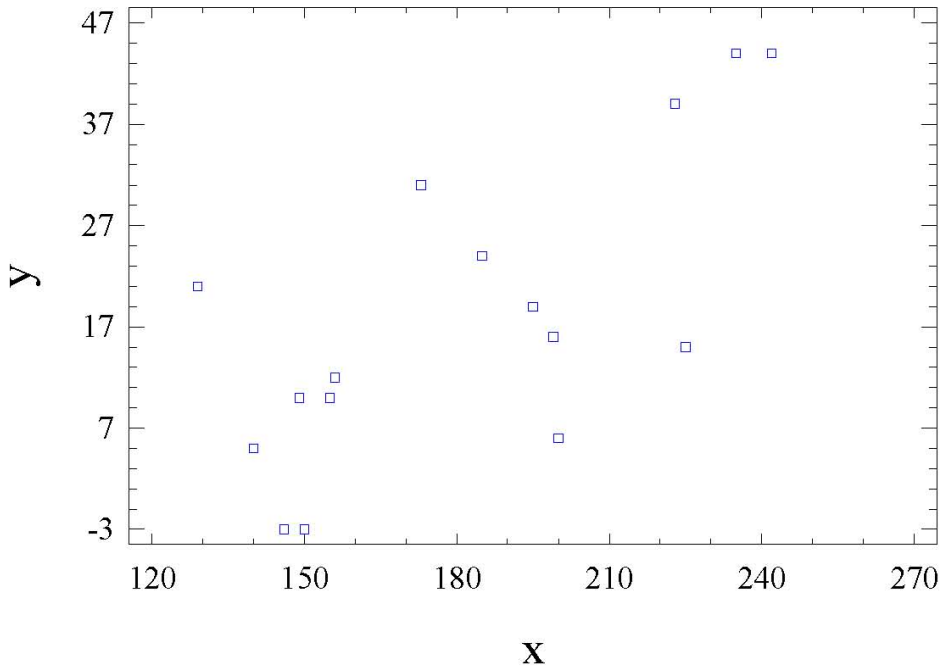
Often we are interested in modeling the extent of a linear relationship between two data samples.

Example 18

In a study on the treatment of diabetes, the researchers measured patients weight losses, y, and their initial weights on diagnosis, x to see if weight loss was influenced by initial weight.

X	225	235	173	223	200	199	129	242
Y	15	44	31	39	6	16	21	44
X	140	156	146	195	155	185	150	149
Y	5	12	-3	19	10	24	-3	10

In order to assess the relationship, it is useful to plot these data as a scatter plot.



We can see that there is a positive relationship between initial weight and weight loss.

The sample covariance

In such cases, we can measure the extent of the relationship using the sample correlation. Given a sample, $(x_1, y_1), \ldots, (x_n, y_n)$, then the *sample covariance* is defined to be

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

In our example, we have

$$\begin{aligned}
\bar{x} &= \frac{1}{16}(225 + 235 + \ldots + 149) \\
&= 181.375 \\
\bar{y} &= \frac{1}{16}(15 + 44 + \ldots + 10) \\
&= 18.125 \\
s_{xy} &= \frac{1}{16}\{(225 - 181.375)(15 - 18.125) + \\
&\quad (235 - 181.375)(44 - 18.125) + \ldots + \\
&\quad (149 - 181.375)(10 - 18.125)\} \approx 361.64
\end{aligned}$$

The sample correlation

The *sample correlation* is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

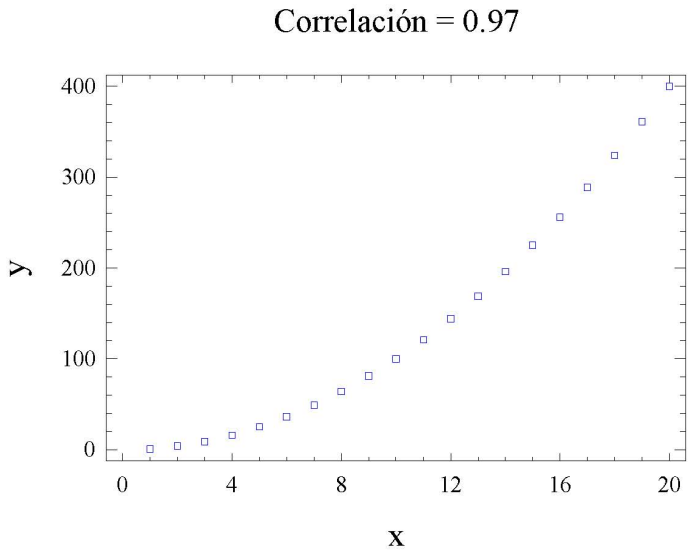
where s_x and s_y are the two standard deviations. This has properties similar to the population correlation.

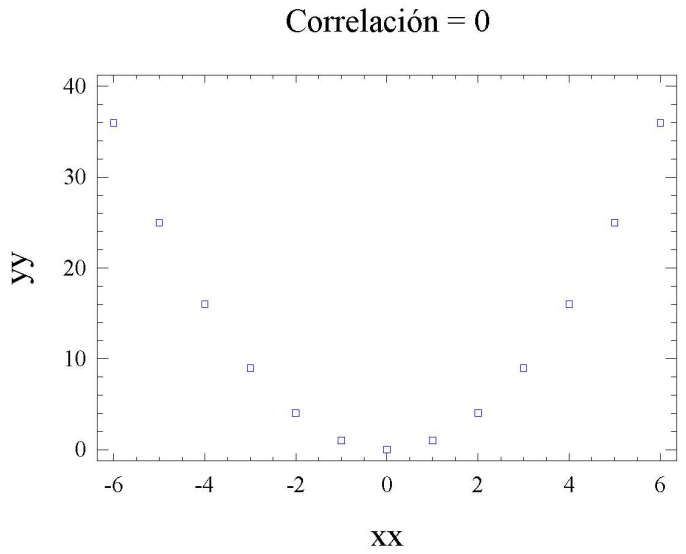
- $-1 \leq r_{xy} \leq 1$.
- $r_{xy} = 1$ if $y = a + bx$ and $r_{xy} = -1$ if $y = a - bx$ for some $b > 0$.
- If there is no relationship between the two variables, then the correlation is (approximately) zero.

In our example, we find that $s_x^2 \approx 1261.98$ and $s_y^2 \approx 211.23$ so that $s_x \approx 35.52$ and $s_y \approx 14.53$. which implies that $r_{xy} = \frac{361.64}{35.52 \times 14.53} \approx 0.70$ indicating a strong, positive relationship between the two variables.

Correlation only measures linear relationships!

High or low correlations can often be misleading.





In both cases, the variables have strong, non-linear relationships. Thus, whenever we are using correlation or building *regression models*, it is always important to plot the data first.

Spurious correlation

Correlation is often associated with causation. If X and Y are highly correlated, it is often assumed that X causes Y or Y causes X.

Spurious correlation

Correlation is often associated with causation. If X and Y are highly correlated, it is often assumed that X causes Y or Y causes X.

Example 19

Springfield had just spent millions of dollars creating a highly sophisticated "Bear Patrol" in response to the sighting of a single bear the week before.

Homer: Not a bear in sight. The "Bear Patrol" is working like a a charm
Lisa: That's specious reasoning, Dad.
Homer:[uncomprehendingly] Thanks, honey.
Lisa: By your logic, I could claim that this rock keeps tigers away.
Homer: Hmm. How does it work? Lisa:
It doesn't work. (pause) It's just a stupid rock!
Homer: Uh-huh.
Lisa: But I don't see any tigers around, do you?
Homer: (pause) Lisa, I want to buy your rock.

Much Apu about nothing. The Simpsons series 7.

Example 20

1988 US census data showed that numbers of churches in a city was highly correlated with the number of violent crimes. Does this imply that having more churches means that there will be more crimes or that having more crime means that more churches are built?

Example 20

1988 US census data showed that numbers of churches in a city was highly correlated with the number of violent crimes. Does this imply that having more churches means that there will be more crimes or that having more crime means that more churches are built?

Both variables are highly correlated to population. The correlation between them is spurious.

Regression

An model representing an approximately linear relation between x and y is

$$y = \alpha + \beta x + \epsilon$$

where ϵ is a prediction error.

In this formulation, y is the *dependent variable* whose value is modeled as depending on the value of x, the *independent variable* .

How should we fit such a model to the data sample?

Least squares

Gauss

We wish to find the line which best fits the sample data $(x_1, y_1), \ldots, (x_n, y_n)$. In order to do this, we should choose the line, $y = a + bx$, which in some way minimizes the prediction errors or *residuals*,

$$e_i = y_i - (a + bx_i) \quad \text{for } i = 1, \ldots, n.$$

A minimum criterion would be that $\sum_{i=1}^{n} e_i = 0$. However, many lines satisfy this, for example $y = \bar{y}$. Thus, we need a stronger constraint.

The standard way of doing this is to choose to minimize the sum of squared errors, $E(a, b) = \sum_{i=1}^{n} e_i^2$.

Theorem 7

For a sample $(x_1, y_1), \ldots, (x_n, y_n)$, the line of form $y = a + bx$ which minimizes the sum of squared errors, $E[a, b] = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ is such that

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Proof Suppose that we fit the line $y = a + bx$. We want to minimize the value of $E(a, b)$. We can recall that at the minimum,

$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0.$$

Now, $E = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ and therefore

$$\begin{aligned}
\frac{\partial E}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) \quad \text{and at the minimum} \\
0 &= -2\sum_{i=1}^{n}(y_i - a - bx_i) \\
&= -2\left(n\bar{y} - na - nb\bar{x}\right) \\
a &= \bar{y} - b\bar{x}
\end{aligned}$$

$$\frac{\partial E}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) \quad \text{and at the minimum,}$$

$$0 = -2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i(a + bx_i)\right)$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i(a + bx_i)$$

$$= \sum_{i=1}^{n} x_i(\bar{y} - b\bar{x} + bx_i) \quad \text{substituting for } a$$

$$= n\bar{x}\bar{y} + b\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{ns_{xy}}{ns_{\bar{x}}^2} = \frac{s_{xy}}{s_{\bar{x}}^2}$$

We will fit the regression line to the data of our example on the weights of diabetics. We have seen earlier that $\bar{x} = 181.375$, $\bar{y} = 18.125$, $s_{xy} = 361.64$, $s_x^2 = 1261.98$ and $s_y^2 = 211.23$.

Thus, if we wish to predict the values of y (reduction in weight) in terms of x (original weight), the least squares regression line is

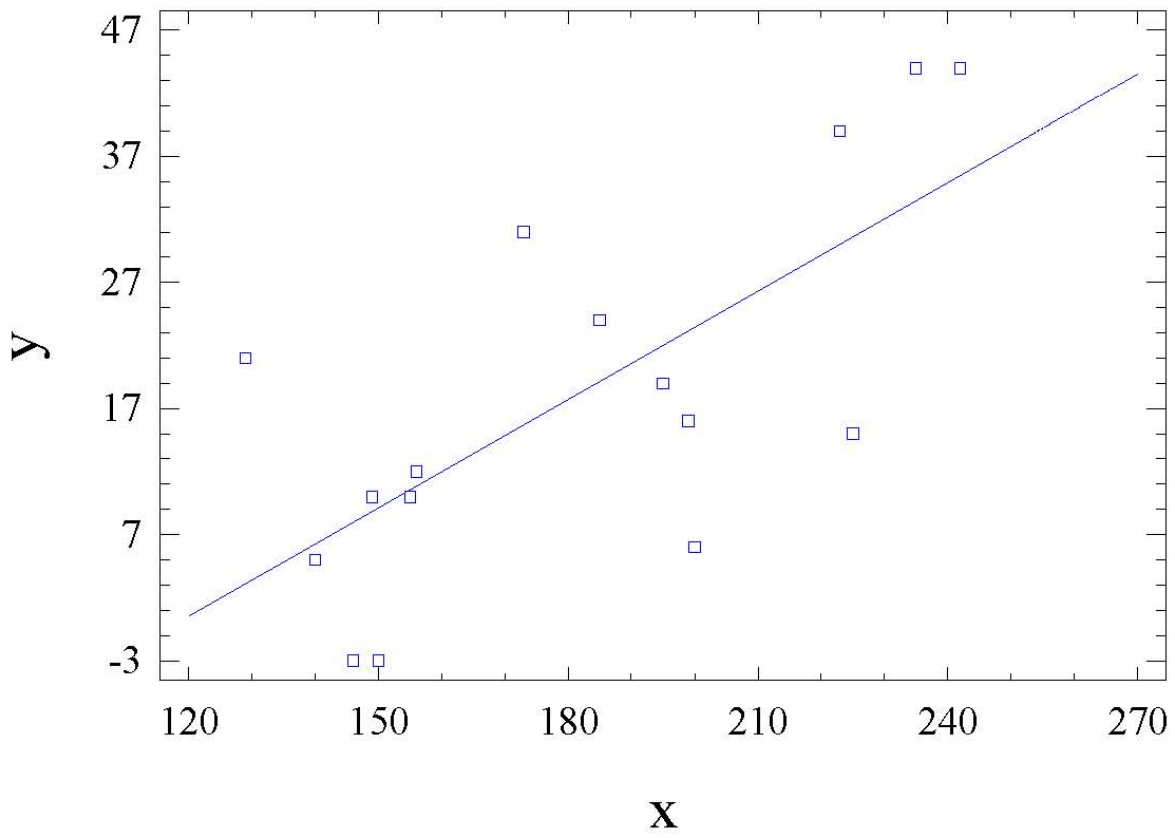
$$y = a + bx$$

where

$$\begin{aligned} b &= \frac{361.64}{1261.98} \approx 0.287 \\ a &= 18.125 - 0.287 \times 181.375 \approx -33.85 \end{aligned}$$

The following diagram shows the fitted regression line.

Diagrama de dispersión con la recta de regresión añadida



We can use this line to predict the weight loss of a diabetic given their initial weight. Thus, for a diabetic who weighed 220 pounds on diagnosis, we would predict that their weight loss would be around

$$\hat{y} = -33.85 + 0.287 \times 220 = 29.29 \text{ lbs.}$$

Note that we should be careful when making predictions outside the range of the data. For example the linear predicted weight gain for a 100 lb patient would be around 5 lbs but it is not clear that the linear relationship still holds at such low values.

Residual analysis

Los residuals or prediction errors are the differences $e_i = y_i - (a + bx_i)$. It is useful to see whether the average prediction error is small or large. Thus, we can define the *residual variance*

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} e_i^2$$

and the *residual standard deviation*, $s_e = \sqrt{s_e^2}$.

In our example, we have $e_1 = 15 - (-33.85 + 0.287 \times 225)$, $e_2 = 44 - (-33.85 + 0.287 \times 235)$ etc. and after some calculation, the residual sum of squares can be shown to be $s_e^2 \approx 123$. Calculating the results this way is very slow. There is a faster method.

Theorem 8

$$\begin{aligned} \bar{e} &= 0 \\ s_r^2 &= s_y^2\left(1 - r_{xy}^2\right) \end{aligned}$$

Proof

$$\begin{aligned} \bar{e} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\bar{y} - b\bar{x} + bx_i)) \quad \text{by definition of } a \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y}) - b\sum_{i=1}^{n}(x_i - \bar{x})\right) \\ &= 0 \end{aligned}$$

$$
\begin{aligned}
s_e^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-(a+bx_i)\right)^2 \\
&= \frac{1}{n-1}\left(y_i-(\bar{y}-b\bar{x}+bx_i)\right)^2 \quad \text{by definition of } a \\
&= \frac{1}{n-1}\left((y_i-\bar{y})-b(x_i-\bar{x})\right)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2-2b\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+b^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \\
&= s_y^2-2bs_{xy}+b^2s_x^2 = s_y^2-2\frac{s_{xy}}{s_x^2}s_{xy}+\left(\frac{s_{xy}}{s_x^2}\right)^2 s_x^2 \quad \text{by definition of } b \\
&= s_y^2-\frac{s_{xy}^2}{s_x^2} = s_y^2\left(1-\frac{s_{xy}^2}{s_x^2s_y^2}\right) \\
&= s_y^2\left(1-\left(\frac{s_{xy}}{s_xs_y}\right)^2\right) = s_y^2\left(1-r_{xy}^2\right) \qquad \blacksquare
\end{aligned}
$$

Interpretation

This result shows that

$$\frac{s_r^2}{s_y^2} = (1 - r_{xy}^2).$$

Consider the problem of estimating y. If we only observe $y, \ldots, y_n$, then our best estimate is $\bar{y}$ and the variance of our data is s_y^2.

Given the x data, then our best estimate is the regression line and the residual variance is s_r^2.

Thus, the percentage reduction in variance due to fitting the regression line is

$$R^2 = (1 - r_{xy}^2) \times 100\%$$

In our example, $r_{xy} \approx 0.7$ so $R^2 = (1 - 0.49) \times 100\% = 51\%$.

Graphing the residuals

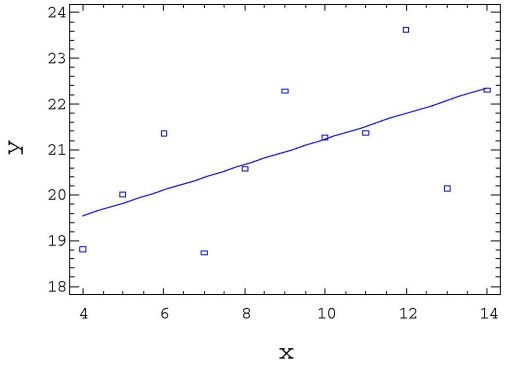
As we have seen earlier, the correlation between two variables can be high when there is a strong non-linear relation.

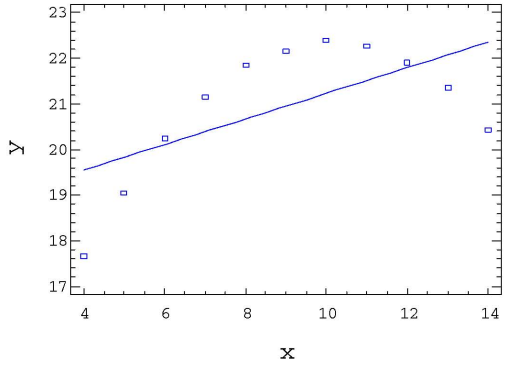
Whenever we fit a linear regression model, it is important to use residual plots in order to check the adequacy of the fit.

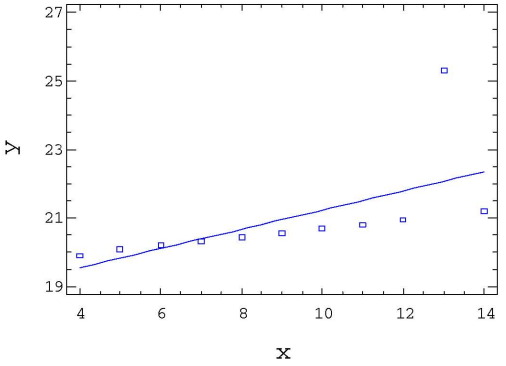
The regression line for the following five groups of data, from Basset et al (1986) is the same, that is

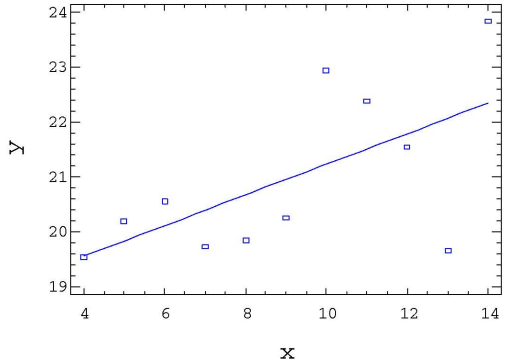
$$y = 18.43 + 0.28x$$

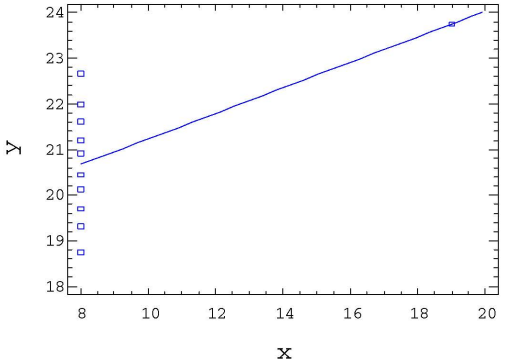
Bassett, E. et al (1986). *Statistics: Problems and Solutions*. London: Edward Arnold


y
x
24
23
22
21
20
19
18
4
6
8
10
12
14


y
x
23
22
21
20
19
18
17
4
6
8
10
12
14

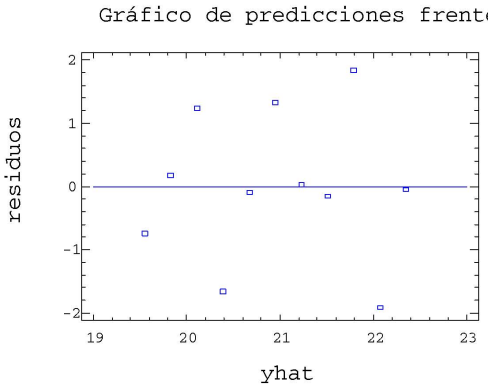

y
x
27
25
23
21
19
4
6
8
10
12
14

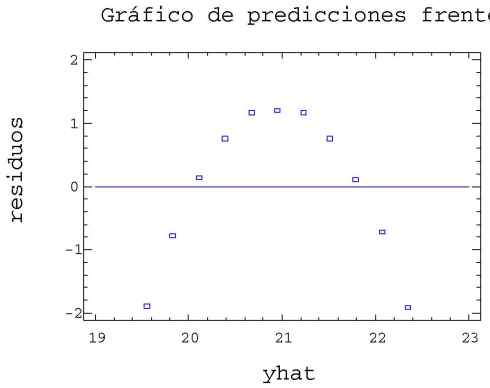

y
x
24
23
22
21
20
19
4
6
8
10
12
14

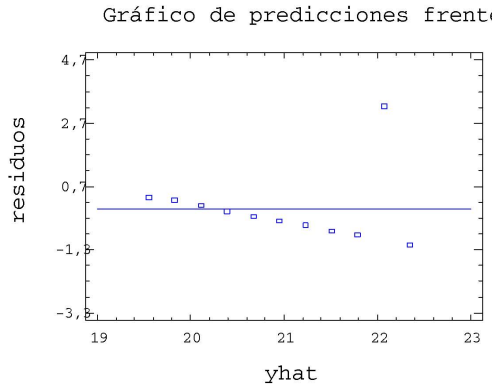

y
x
24
23
22
21
20
19
18
8
10
12
14
16
18
20

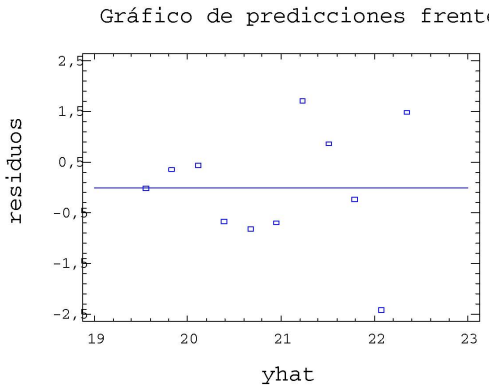
- The first case is a standard regression.
- In the second case, we have a non-linear fit.
- In the third case, we can see the influence of an outlier.
- The fourth case is a regression but ...
- In the final case, we see that one point is very influential.

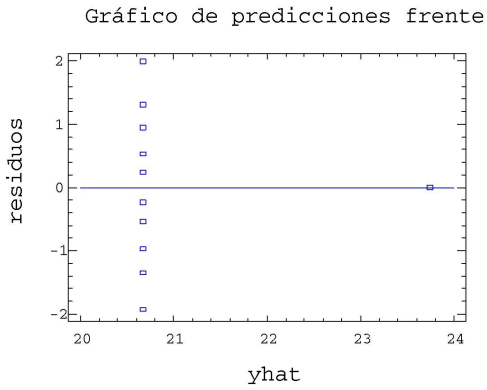
Now we can observe the residuals.











In case 4 we see the residuals increasing as y increases.

Two regression lines

So far, we have used the least squares technique to fit the line $y = a + bx$ where $a = \bar{y} - b\bar{x}$ and $b = \frac{s_{xy}}{s_x^2}$.

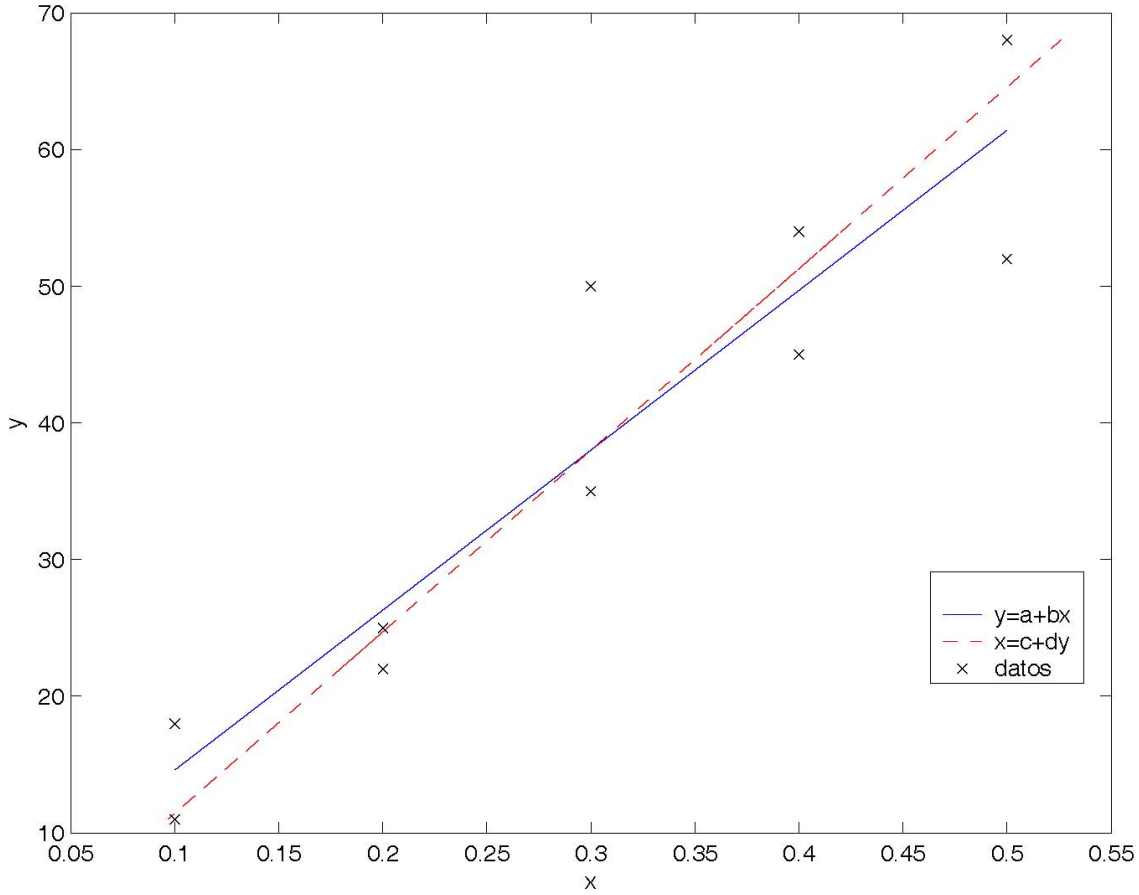
We could also rewrite the linear equation in terms of x and try to fit $x = c + dy$. Then, via least squares, we have that $c = \bar{x} - d\bar{y}$ and $d = \frac{s_{xy}}{s_y^2}$.

We might expect that these would be the same lines, but rewriting

$$y = a + bx \Rightarrow x = -\frac{a}{b} + \frac{1}{b}y \neq c + dy$$

It is important to notice that the least squares technique minimizes the prediction errors in one direction only.

The following example shows data on the extension of a cable, y relative to force applied, x and the fit of both regression lines.



Regression and normality

Suppose that we have a statistical model

$$Y = \alpha + \beta x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then, if data come from this model, it can be shown that the least squares fit method coincides with the *maximum likelihood* approach to estimating the parameters.

You will study this in more detail in the course on Regression Models.