Applied Statistics and Data Analysis

3. A review of inference concepts
a. Statistical models

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine

paolo.vidoni@uniud.it

Based mainly on Chapter 3 of the course textbook *Statistical models*

Table of contents

1. Summary and introduction
2. Probability distributions and random variables
3. Basic statistical models
4. Simulation of random samples
5. Model assumptions

Summary

- **Introduction to statistical models** (*for individual revising*)
- **Probability distributions and random variables** (*for individual revising*)
- **Basic statistical models** (*for individual revising*)
- **Simulation of random samples**
- **Model assumptions**

Basics of statistical models

Based also on Chapter 2 of *Core Statistics* by S.N. Wood

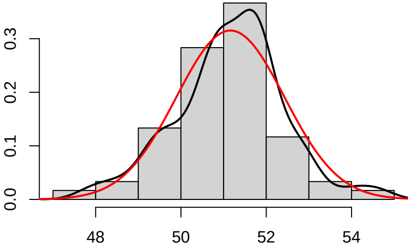
- Inferential statistics is about extracting information from data: specifically, information about the “system” that generated the data or about the population from which the sample data are obtained.
- Most data contain a component of **random variability**: replications of the data gathering process several times would give somewhat different data on each occasion.
- In many physical sciences, **deterministic models** are often adequate, as data variability may be small. Outside such cases, and nearly always in the social sciences, variability is a serious issue, and models have to incorporate it, leading to **statistical models**.
- Statistical models involve (**families**) of **probability distributions**, with the aim of providing an adequate description of the data generating system or of the interest phenomenon.

- If the model elements (for example, models parameters) were known then an adequate model could generate data that resembled the observed data, including reproducing its variability under replication.
- The purpose of statistical inference is to use the statistical model to go in the reverse direction: to infer the values of the model unknowns that are consistent with observed data.
- Statistical models for some data are chosen based on previous **experience** with similar data, subject area **knowledge** and careful use of **EDA findings**.
- Statistical models often combine a **deterministic component** and a **random component**, that is an inherently unpredictable component.
- The random component is often called **noise** or **error** (but there's typically nothing wrong with it), and sometimes the deterministic part is called **signal**.

Example: temperatures

Data set y with a 60 year record of `mean annual temperatures` (°F) in New Haven, Connecticut, from 1912 to 1971.

Numerical summaries: $\bar{y} = 51.16$, $y_{0.5} = 51.20$, $s^2 = 1.60$, $\gamma = -0.07$, $\beta = 3.38$.

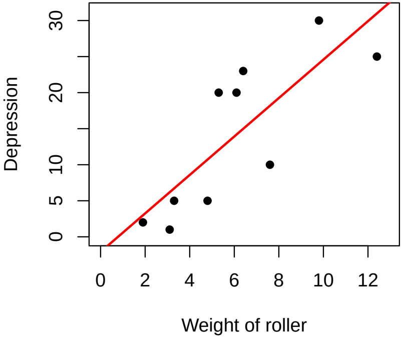


Graphical and numerical summaries suggest a simple model which considers y as independent observations from a normal distribution, although the tails seem heavier than those of the "bell curve".

Example: roller data

Data from an experiment where different `weights` (t) of roller were rolled over different part of a lawn, and the `depression` (mm) measured.

Graphical summary: scatterplot of the data, with the least squares line



The assumed model has a linear form for the *deterministic part* and an additive *error term*

$$\texttt{depression} = \alpha + \beta \cdot \texttt{weight} + \varepsilon$$

It is called **linear regression model**. Here α and β are **model parameters**, namely constants which must be estimated using the data.

Subscripts allow identification of the individual points: given observations $(x_1, y_1), \ldots, (x_n, y_n)$

$$y_i = \alpha + \beta\, x_i + \varepsilon_i,\, i = 1, \ldots, n$$

Using the least squares method, parameter estimates are $\widehat{\alpha} = -2.087$, $\widehat{\beta} = 2.667$ and the **fitted values** are defined by

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}\, x_i,\, i = 1, \ldots, n$$

whereas the observed **residuals** are given by

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta}\, x_i,\, i = 1, \ldots, n$$

- The focus of interest can be cast in terms of **interpretation of model parameters** or **prediction**.
- A crucial parameter is β, namely the rate of increase of depression with increasing roller weight.
- Predictions are given by fitted values $\widehat{y}_i$. One may also predict the depression corresponding to out-of-sample roller weights, though some care would be required.
- The model treats the pattern of change of depression with roller weight as a deterministic or *fixed effect term*.
- The measured values of depression incorporate also a *random term* that reflects:
  - variation from a part of the lawn to another;
  - differences in handling the roller ;
  - measurement error.

  It is assumed that its elements are **uncorrelated**: size and sign of one element does not provide any information on the other elements.
- Data from **multiple lawns** are essential if one wants to generalize results to other lawns.

A brief note on the least squares line

- It is useful to remember that the **least squares** line has coefficients $\widehat{\alpha}, \widehat{\beta}$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2$$

- It takes some simple linear algebra to show that the two coefficients solve a simple linear system giving

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}, \quad \widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- At times, **weighted least squares** are useful. Some fixed weights w_i are introduced, and the quantity to be minimized is then

$$\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2\, w_i$$

Table of contents

1. Summary and introduction
2. **Probability distributions and random variables**
3. Basic statistical models
4. Simulation of random samples
5. Model assumptions

Random variables

- The concepts of randomness and probability are central to Statistics and the view of data as coming from a probability distribution is fundamental to understanding statistical methods.
- **Random variables** (r.v.'s) are building blocks for statistical models, and in particular of their random component.
- A r.v. takes a different (numerical) value, at random, each time it is observed. It is possible to make probability statements about the values likely to occur, that is to specify its **probability distribution**.
- The **distribution function** of a r.v. X is the function $F(x)$ such that
$$F(x) = P(X \leq x)$$
It gives the probability that the value of X will be less than or equal to $x \in \mathbf{R}$.
- From $F(x)$ it is possible to define the potential values for X, which belong to the **support** $\mathcal{S}$ of X, and the probability of events related to X, such as $X = a$, $X > a$, $a < X \leq b$, with $a < b \in \mathbf{R}$.

Discrete and continuous random variables

Discrete r.v.'s take a discrete set of values (finite or countable) and they are suitable for finite or count data.

They are described by the **probability (mass) function**

$$f(x) = P(X = x)$$

Clearly, $f(x) \in [0, 1]$ and, for the potential values of X, that is x_i, $i \in I \subseteq \mathbf{N}$, $f(x_i) > 0$ and $\sum_{i \in I} f(x_i) = 1$.

Continuous r.v.'s take values in a continuous set and the probability of taking any particular value is zero.

They are described by the (**probability**) **density function** $f(x)$ such that

$$P(a \leq X \leq b) = \int_a^b f(x)dx,\ a < b \in \mathbf{R}$$

Clearly, $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x)dx = 1$, $\int_{-\infty}^{b} f(x)dx = F(b)$, so that $F'(x) = f(x)$, when the first derivative $F'(x)$ exists.

Mean, variance and quantiles

Instead of considering the distribution of a r.v. X completely, for many purposes its first two moments suffice.

In particular, the **expected value (mean)** $\mu = E(X)$ of a discrete or continuous r.v. X, given respectively by

$$E(X) = \sum_{i \in I} x_i f(x_i), \quad E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

and the **variance** $\sigma^2 = V(X) = E[(X-\mu)^2]$, with σ the associated **standard error**; index of skewness and kurtosis may be defined similarly.

The α-**quantile** x_α of X, with $\alpha \in (0,1)$, is a value that X will be less than or equal to, with probability α.

The **median** of X corresponds to $x_{0.5}$ whereas the **quartiles** and the **percentiles** are obtained with the corresponding choices for α.

The transformed r.v. $Z = (X-\mu)/\sigma$ is called **standardized r.v.**, since $E(Z) = 0$ and $V(Z) = 1$.

Random vectors

- Little can usually be learned, on the interest phenomenon, from single observations.
- Useful statistical analysis requires multiple observations, viewed as a realization of a **random vector** (**multivariate random variable**).
- A random vector $(X_1, \ldots, X_n)$ takes values $(x_1, \ldots, x_n) \in \mathbf{R}^n$, namely numerical n-dimensional vectors, according to a joint probability distribution.
- The probability distribution is defined by the **joint distribution function**

  $$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

  or, equivalently, by the **multivariate** (**joint**) version of the **density** (**probability**) **function** $f(x_1, \ldots, x_n)$.
- Each marginal component X_i, $i = 1, \ldots, n$, corresponds to a r.v. with **marginal density** (**probability**) **function** $f_i(x_i)$.

The two following situations greatly simplify statistical analysis:

- the component r.v.'s X_i, $i = 1, \ldots, n$, are **independent** (the realization of one does not affect the probability distribution of the others) and then

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

- the component r.v.'s X_i, $i = 1, \ldots, n$, are **independent and identically distributed** (i.i.d.), so that each component follow the same distribution with density (probability) function $g(x)$ and

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i).$$

Bivariate random variables

The two-dimensional case suffices to illustrate most of the concepts required for higher dimensions.

Let us consider a **continuous bivariate random variable** (X, Y) with density function $f(x, y)$; the results for the **discrete case** are simply obtained by substituting summation for integration.

- The **marginal density** of X is

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

and similarly for $f(y)$.

- The **conditional density** of X given $Y = y$ is

$$f(x|y) = \frac{f(x, y)}{f(y)}$$

assuming $f(y) > 0$, and similarly for $f(y|x)$.

- **Bayes theorem** (which leads to a whole school of statistical methods)
  $$f(x|y) = \frac{f(x)f(y|x)}{f(y)}$$
  assuming $f(y) > 0$.
- The component r.v.'s X and Y are **independent** if and only if $f(x,y) = f(x)f(y)$.
- The **conditional expectation (mean)** of X given $Y = y$ is
  $$E(X|Y = y) = \int_{-\infty}^{+\infty} xf(x|y)dx$$
  and similarly for $E(Y|X = x)$; analogous definition for the conditional variance of X given $Y = y$.
- The **covariance** of (X, Y) is
  $$Cov(X,Y) = \sigma_{XY} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\{x - E(X)\}\{y - E(Y)\}f(x,y)dxdy$$

- If X and Y are are independent, then $Cov(X,Y) = 0$; the reverse does not hold (a relevant exception concerns the multivariate normal distribution).
- The **Pearson correlation coefficient** of (X,Y), useful for describing linear dependencies, is

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

- The first and second order moments of (X,Y) are summarized by the **mean vector** $\mu = (\mu_X, \mu_Y) = (E(X), E(Y))$ and the **variance-covariance matrix**

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} V(X) & Cov(X,Y) \\ Cov(Y,X) & V(Y) \end{pmatrix}$$

The Σ matrix is symmetric, since $Cov(X,Y) = Cov(Y,X)$, and positive semidefinite.

Statistics

- A (**sample**) **statistic** is a function (summary) of a set of r.v.'s and it is itself a r.v.
- The probability distribution of a sample statistic is called **sampling distribution** and its form depends on the joint distribution of the initial random vector.
- Given a random vector $(X_1, \dots, X_n)$, well-known examples of statistics are the **sample mean** and the (corrected) **sample variance** defined, respectively, as

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

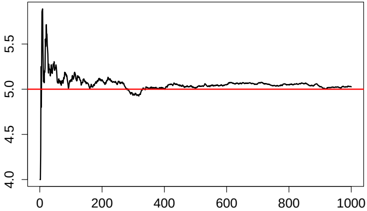
  The uncorrected sample variance is obtained by substituting the degrees of freedom $n-1$ with n.
- Further examples of statistics are the sample median, the sample quantiles, the sample MAD, the sample covariance and the sample correlation coefficient.

Some useful results are listed below.

- Whenever $X_1, \ldots, X_n$ are uncorrelated (independent) r.v.'s, having the same marginal mean μ and variance σ^2 (e.g. this happens for identically distributed r.v.'s):
  - $E(\sum_{i=1}^n X_i) = n\mu$, $V(\sum_{i=1}^n X_i) = n\sigma^2$;
  - $E(\bar{X}) = \mu$, $V(\bar{X}) = \sigma^2/n$;
  - $E(S^2) = \sigma^2$, while for the uncorrected version the expectation is $\sigma^2(n-1)/n < \sigma^2$.
- **Weak law of large numbers**: if $X_1, \ldots, X_n$ are i.i.d. r.v.'s, $\bar{X}$ converges in probability to μ, as $n \to +\infty$ (in symbols, $\bar{X} \xrightarrow{p} \mu$); that is, as n increases, the distribution of $\bar{X}$ is more and more concentrated around the marginal mean μ.
- A similar result holds for the (corrected and the uncorrected) sample variance: $S^2 \xrightarrow{p} \sigma^2$, as $n \to +\infty$.

As a simple application of the weak law of large numbers, let $X_1, \dots, X_n$ be i.i.d. $Po(\lambda)$ distributed r.v.'s with $\lambda = 5$.

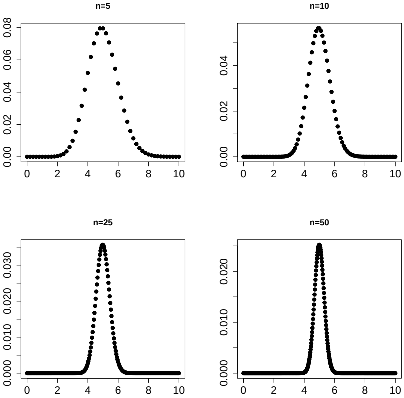
A sequence of observed values for the sample mean $\bar{X} = \sum_{i=1}^n X_i/n$, with $n = 1, \dots, 1000$, is given below



The sample path shows that, as n increases, the observed values of $\bar{X}$ tend to be more concentrated around $\mu = \lambda = 5$.

Indeed, since the sample sum $\sum_{i=1}^n X_i$ follows a $Po(n\lambda)$ distribution, it is easy to specify the distribution of the sample mean $\bar{X} = \sum_{i=1}^n X_i/n$.

The following figure describes the probability function of $\bar{X}$ for $n = 5, 10, 25, 50$



The mean value is always $\mu = 5$, while the variability lessens as n increases.

Table of contents

1. Summary and introduction
2. Probability distributions and random variables
3. **Basic statistical models**
4. Simulation of random samples
5. Model assumptions

Discrete uniform distribution

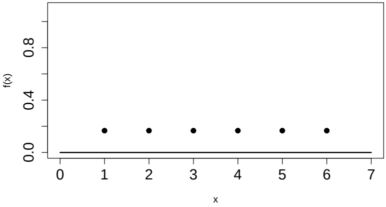
A discrete uniform distribution describes an experiment where a finite number of values are equally likely to be observed.

A discrete r.v. X follows a **discrete uniform distribution** with values $x_1, \dots, x_n \in \mathbf{R}$, $n \in \mathbf{N}^+$, abbreviated as $X \sim Ud(x_1, \dots, x_n)$, if $\mathcal{S} = \{x_1, \dots, x_n\}$ and

$$f(x_1) = \cdots = f(x_n) = \frac{1}{n}$$

Indeed, $E(X) = \sum_{i=1}^{n} x_i/n$, $V(X) = \sum_{i=1}^{n}\{x_i - E(X)\}^2/n$.

The probability function $f(x)$, for $n = 6$ and $x_i = i$, $i = 1, \dots, 6$, is

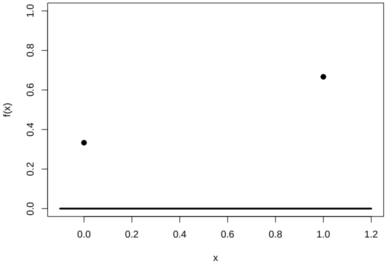


Bernoulli distribution

A discrete r.v. X follows a **Bernoulli distribution** with parameter $p \in (0,1)$, abbreviated as $X \sim Ber(p)$, if it describes an experiment where the possible outcomes are "success" (or 1) and "failure" (or 0); success may occur with probability p.

$\mathcal{S} = \{0,1\}$ and $f(1) = P(X=1) = p$, $f(0) = P(X=0) = 1-p$; indeed, $E(X) = p$ and $V(X) = p(1-p)$.

The probability function $f(x)$ for $p = 1/3$ is described below



Binomial distribution

A discrete r.v. X follows a **binomial distribution** with parameters $n \in \mathbf{N}$, $p \in (0,1)$, abbreviated as $X \sim Bi(n,p)$, if it describes the number of successes in n independent Bernoulli experiments with the same success probability p.

$\mathcal{S} = \{0, \ldots, n\}$ and

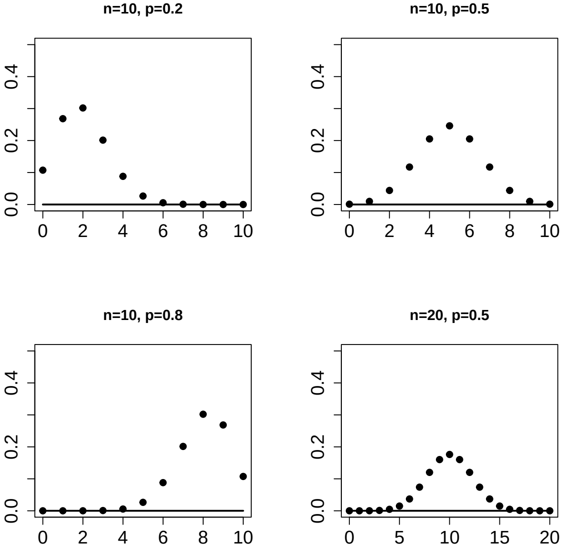
$$f(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

X may be viewed as the sum of n of independent $Ber(p)$ r.v.'s; note that $Bi(1,p)$ corresponds to $Ber(p)$.

Indeed, $E(X) = np$ and $V(X) = np(1-p)$.

It is easy to see that the proportion of successes X/n is such that $E(X/n) = p$ and $V(X/n) = p(1-p)/n$.

Probability function $f(x)$ for different n and p values.



Poisson distribution

The Poisson distribution is often used to model the number of events that occur in a certain time interval or in a prescribed spatial region, e.g. numbers of defects observed in manufactured products, number of visits to a website by an individual user.

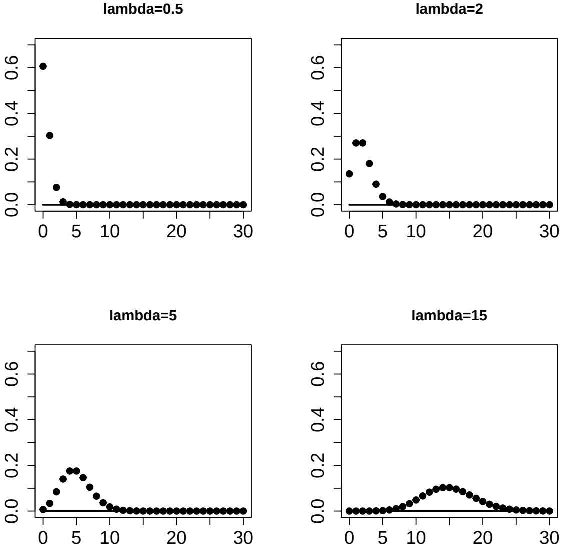
A discrete r.v. X follows a **Poisson distribution** with parameter $\lambda > 0$, abbreviated as $X \sim Po(\lambda)$, if $\mathcal{S} = \mathbf{N}$ and

$$f(x) = \begin{cases} \lambda^x e^{-\lambda}/x! & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = \lambda$ and $V(X) = \lambda$, thus the mean and the variance are both equal to the parameter λ.

Moreover, the probability distribution of a $Bi(n, p)$ r.v., with $n \geq 50$ and $p \leq 1/25$, is close to that of a Poisson distributed r.v. with $\lambda = np$.

Probability function $f(x)$ for different λ values.



Geometric distribution

The geometric distribution is similar to the binomial distribution but it records the number of trials for the first success in a sequence of independent Bernoulli experiments with the same success probability p.

A discrete r.v. X follows a **geometric distribution** with parameter $p \in (0, 1)$, abbreviated as $X \sim Ge(p)$, if $\mathcal{S} = \mathbf{N}^+$ and

$$f(x) = \begin{cases} (1-p)^{x-1}p & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = 1/p$ and $V(X) = (1-p)/p^2$.

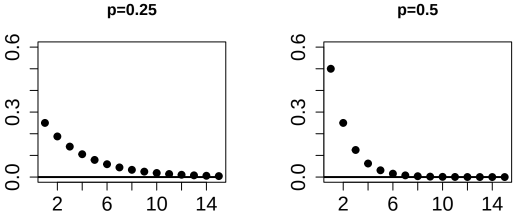
The geometric distribution is memoryless; this means that, given that the first success has not yet occurred, the conditional probability distribution of the number of additional trials does not depend on how many failures have been observed:

$$P(X > s + t | X > s) = P(X > t), \;\; s, t \in S$$

The **negative binomial distribution** generalizes the geometric one by considering the number of trials until the r-th success, with $r \geq 1$, in a sequence of independent Bernoulli experiments with the same success probability p.

The case with $r = 1$ define the geometric distribution.

The geometric probability functions $f(x)$ for $p = 0.25$ and $p = 0.5$ are described below



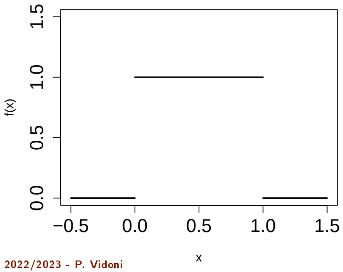
Continuous uniform distribution

The continuous uniform distribution describes equiprobability for continuous experiments, that is all intervals of the same length on the distribution's support are equally probable.

A r.v. X follows a **continuous uniform (rectangular) distribution** with parameter $a, b \in \mathbf{R}$, $a < b$, abbreviated as $X \sim U(a, b)$, if $\mathcal{S} = [a, b]$ and

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = (a+b)/2$, $V(X = (b-a)^2/12$ and, for $a = 0$ and $b = 1$, the density function is



Exponential distribution

The exponential distribution is often used to describe durations, failure times or waiting times, assuming a constant hazard rate λ.

A continuous r.v. X follows an **exponential distribution** with parameter $\lambda > 0$, abbreviated as $X \sim Esp(\lambda)$, if $\mathcal{S} = [0, +\infty)$ and

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$.

It can be viewed as a particular case of both the **gamma distribution** and the **Weibull distribution**.

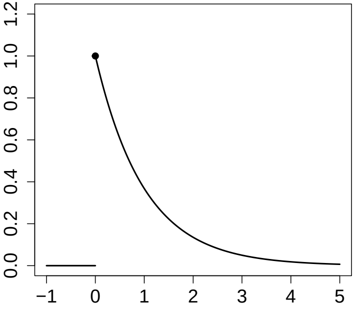
It also describes the (independent) times between two subsequent events in a **Poisson process**, which is a particular count process in which the interest events occur continuously and independently at a constant average rate λ.

The exponential distribution is the continuous analogue of the geometric distribution, having the property of being memoryless; namely,

$$P(X > s + t | X > s) = P(X > t), \ s, t \in S$$

The probability of failure in a given time interval is independent of the previous history.

The density function $f(x)$ for $\lambda = 1$ is



Normal distribution

The normal or Gaussian distribution has a central place in Probability and Statistics, largely as a result of the central limit theorem.

It has a “bell-shaped” density curve and it is often used as a model for a number of continuous measurement data (sometimes after a suitable transformation).

A continuous r.v. X follows a **normal (Gaussian) distribution** with parameters $\mu \in \mathbf{R}$ and $\sigma^2 > 0$, abbreviated as $X \sim N(\mu, \sigma^2)$, if $\mathcal{S} = \mathbf{R}$ and

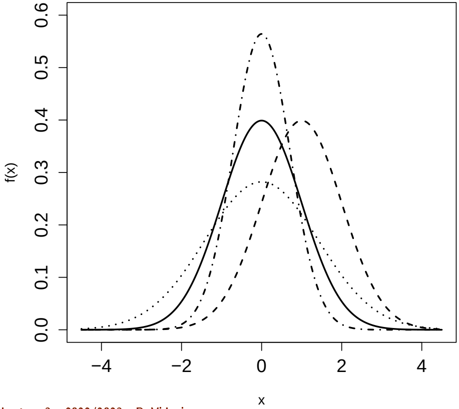
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \; x \in \mathbf{R}$$

Indeed, $E(X) = x_{0.5} = \mu$ and $V(X) = \sigma^2$; the normal distribution is closed with respect to linear transformations, namely, if $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.

A normal distributed r.v. Z having mean 0 and variance (standard deviation) 1 is referred to as the **standard normal r.v.**; note that $Z = (X - \mu)/\sigma$.

The distribution function of a normal distributed r.v. is not explicitly known, however numerical approximation are readily available, giving also the associated α-quantiles.

Density function $f(x)$ of $X \sim N(\mu, \sigma^2)$ with $\mu = 0$, $\sigma^2 = 1$ (—), $\mu = 1$, $\sigma^2 = 1$ (– –), $\mu = 0$, $\sigma^2 = 2$ $(\cdots)$ and $\mu = 0$, $\sigma^2 = 1/2$ (- · -)



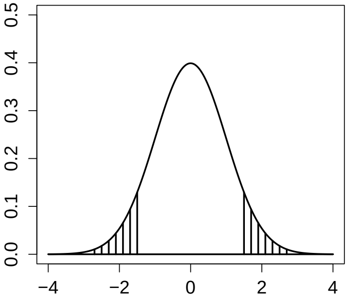
The standard normal density and distribution functions are usually indicated as $\phi(z)$ and $\Phi(z)$, respectively; indeed,

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \ F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

As a consequence of the symmetry of the normal density function, $\Phi(-z) = 1 - \Phi(z)$, $z \geq 0$.

Moreover, as described in the following figure, for $z \geq 0$

$$P(\mid Z \mid < z) = \Phi(z) - \Phi(-z), \quad P(\mid Z \mid > z) = 2\{1 - \Phi(z)\}$$



With regard to statistical applications, the notion of **critical value** of a standard normal distribution may be useful.

The α-critical value of Z, with $\alpha \in (0, 0.5)$, is the value z_α such that $P(Z > z_\alpha) = P(Z < -z_\alpha) = \alpha$.

z_α identifies the right α-level tail of $\phi(z)$, while $-z_\alpha$ defines the symmetric α-level tail on the left-hand side.

In particular,

α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
z_α	1.28	1.65	1.96	2.33	2.58	3.09	3.29

As a straightforward application of the above mentioned results,

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &\doteq 0.68, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\doteq 0.95, \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &\doteq 0.997. \end{aligned}$$

The central limit theorem

Consider i.i.d. r.v.'s $X_1, \ldots, X_n$, with mean μ and variance σ^2, then in the limit $n \to +\infty$

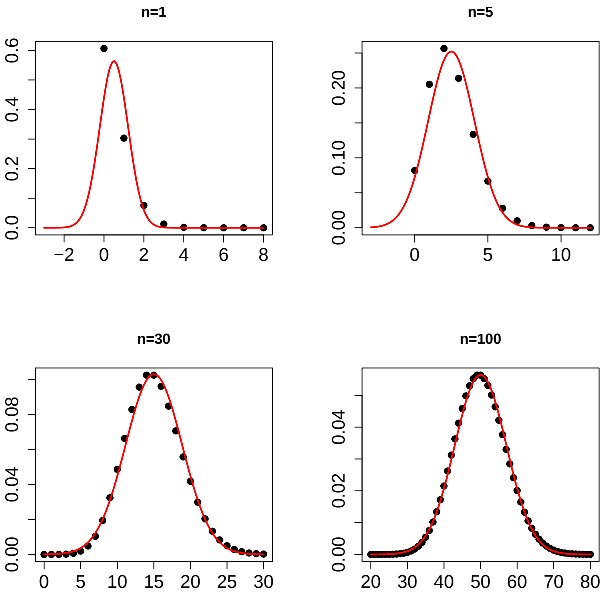
$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$$

Thus, for a large n, the following approximations hold:

$$\bar{X} \sim N(\mu, \sigma^2/n), \quad \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

As a simple application, let $X_1, \ldots, X_n$ be i.i.d. $Po(\lambda)$ distributed r.v.'s; the sample sum $\sum_{i=1}^{n} X_i$ follows a $Po(n\lambda)$ distribution.

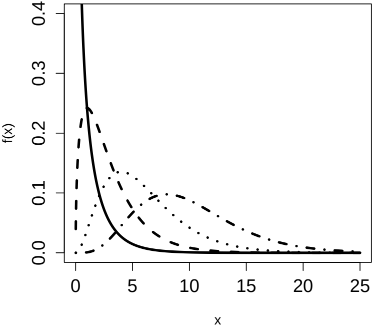
The following figure compares the true probability function of $\sum_{i=1}^{n} X_i$ with the density function of the approximating normal distribution $N(n\lambda, n\lambda)$, for $\lambda = 0.5$ and $n = 1, 5, 30, 100$.


n=1
0.0
0.2
0.4
0.6
−2
0
2
4
6
8
n=5
0.00
0.10
0.20
0
5
10
n=30
0.00
0.04
0.08
0
5
10
15
20
25
30
n=100
0.00
0.02
0.04
20
30
40
50
60
70
80

Chi-squared distribution

Let $Z_1, \dots, Z_k$ be i.i.d. standard normal distributed r.v.'s; the r.v. $Y = \sum_{i=1}^{k} Z_i^2$ follows a **chi-squared distribution** with $k \geq 1$ degrees of freedom, abbreviated as $\chi^2(k)$; it is a special case of the gamma r.v.

It is a continuous r.v. with $\mathcal{S} = [0, +\infty)$, $E(Y) = k$, $V(Y) = 2k$; the density function for $k = 1$ (—), $k = 3$ (– –), $k = 6$ $(\cdots)$, $k = 10$ (- · -) is given below

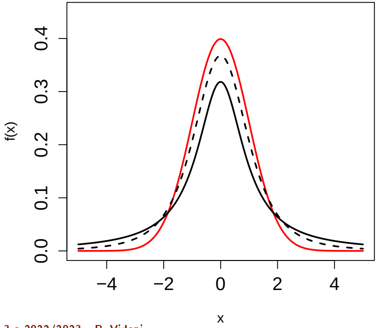


Student's *t* distribution

Let $Z \sim N(0,1)$ and $Y \sim \chi^2(k)$ be independent r.v.'s; the r.v. $T = Z/\sqrt{Y/k}$ follows a **Student's *t* distribution** with $k \geq 1$ degrees of freedom, abbreviated as $T \sim t(k)$.

It is a continuous r.v. with $\mathcal{S} = \mathbf{R}$, $E(T) = 0$, for $k > 1$ and $V(T) = k/(k-2)$, for $k > 2$; if $k \to +\infty$, it converges to a $N(0,1)$ r.v.

The density is symmetric and "bell-shaped", like the normal density, but it has heavier tails; some examples below for $k = 1$ (—), $k = 3$ (– –)



F distribution

Let $X \sim \chi^2(k)$ and $Y \sim \chi^2(m)$, with $k, m \geq 1$, be independent r.v.'s; the r.v. $F = (X/k)/(Y/m)$ follows an ***F*-distribution** with k and m degrees of freedom, abbreviated as $F \sim F(k, m)$.

It is also known as the **Fisher** or the **Snedecor distribution**; it is a continuous r.v. with $\mathcal{S} = [0, +\infty)$ and $E(F) = m/(m-2)$, for $m > 2$.

Density function for $k = 5$, $m = 5$ (—), $k = 5$, $m = 25$ (– –), $k = 25$, $m = 25$ (- · -)

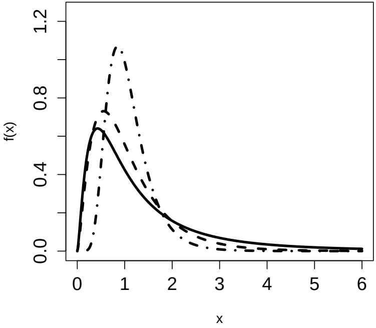


Table of contents

1 Summary and introduction

2 Probability distributions and random variables

3 Basic statistical models

4 Simulation of random samples

5 Model assumptions

Sampling from probability distributions

- Modern statistical software have routine to generate repeated random samples from a specified distribution.

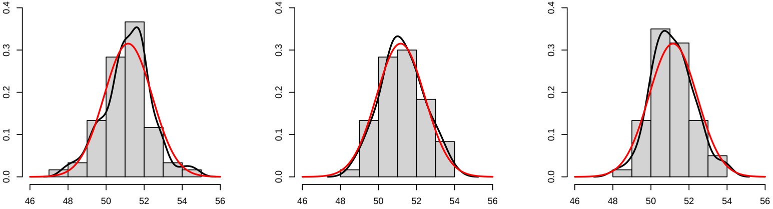
  Such a task is referred to as **simulation**, and it plays a prominent role in modern statistical practice.
- Summary statistics, possibly derived from a certain model, can be computed for each generated (simulated) sample, and their properties can be studied without intricate mathematics.
- Simulation is widely used to determine the properties of statistical procedures in cases where it has not been possible or convenient to derive analytical results.
- It has to be kept in mind, however, that computers generate **pseudo-random** numbers, typically based on deterministic algorithms, having specified a set of initial values.
- It is possible to specify the initial values by setting the **random seed**, thus forcing the generator to produce the same numbers.

Simulations from a normal distribution

Consider the data set y with 60 `mean annual temperatures` (°F) in New Haven, Connecticut, from 1912 to 1971.

Left panel: histogram and density estimate from the original data.

Central and right panels: histogram and density estimate based on simulated samples of dimension $n = 60$ from a normal distribution with $\mu = \bar{y} = 51.16$ and $\sigma^2 = s^2 = 1.60$.

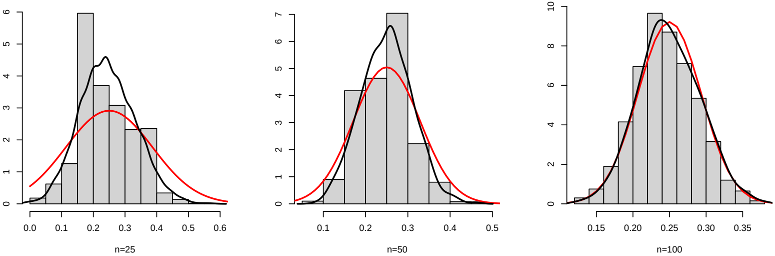


Simulation of the sample mean

The central limit theorem implies that, given an i.i.d. sample $X_1, \ldots, X_n$, for large enough n, the sampling distribution of $\bar{X} = \sum_{i=1}^{n} X_i/n$ will be closely approximated by a $N(\mu, \sigma^2/n)$ distribution.

1000 random samples of size $n = 25, 50, 100$ are simulated from a $Ber(p)$ distribution with $p = 0.25$.

The resulting plots show the distribution of the sample mean estimated by simulation using the histogram and the density estimate, together with the approximating normal density



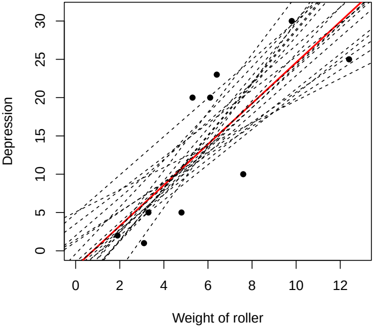
Simulation of regression data

It is possible to simulate a sample of n observations from the linear model

$$y_i = \alpha + \beta\, x_i + \varepsilon_i,\ i = 1, \ldots, n,$$

with $\varepsilon_i \sim N(0, \sigma^2)$, independent of one another, and fixed x_i values.

Consider the roller data set. Least squares lines for each of the 20 simulated sample, with $n = 10$, $\alpha = -2.087$, $\beta = 2.667$ and $\sigma^2 = 45.367$, together with the original data and the original fitted line



Sampling from a finite population

- It is possible to generate a sample from a finite population, which amounts to sampling from a finite set of numbers.
- Two variants are contemplated: sampling **without-replacement**, where no element is selected more than once, and sampling **with-replacement**, where repeated observations are allowed.
- This kind of sampling is useful for practical implementation of **randomization** techniques, that are very important in experimental design and have also a role in statistical inference.
- **Cluster sampling** is one of many different probability-based variants on simple random sampling: the clusters are independent, whereas the elements within the clusters are usually dependent.

Table of contents

Common model assumptions

- Common assumptions for statistical models, having a deterministic and a random component, are **independence** and **normality** of the elements of the random terms and **homogeneity of variance** (that is, standard deviations of all measurements are the same).
- When some assumptions do not hold, a statistical model may be invalid, failing to provide an adequate representation of the data.
- Some of the assumptions may be less important, and a certain method may be **robust** against them.
- An important part of applied statistics is about which assumptions are important and need to be **carefully checked**.
- **Non-parametric methods** have been developed to handle situations where normality or other assumptions are in question (without sensible alternatives); these methods assume little structure into a model and they are only sometimes useful.
- Particular attention is dedicated to the independence and the normality assumptions.

Randomness

- Typically, the data at hand are used as a window onto a much wider population, and they should be **collected in such a way that the randomness assumption is guaranteed**.
- For this reason, randomization in design experiments and random sampling in surveys are very important.
- Samples chosen haphazardly or in a careless fashion (e.g. a survey interview involving individuals found in a shopping center) and **self-selected samples** can totally invalidate a statistical analysis.
- Failure of the randomness assumption is a common reason for wrong statistical inferences, therefore it is crucial to **identify the nature of any possible lack of randomness**.
- As a matter of fact, random sampling assumption is made even when data selection mechanism does not guarantee randomness; in such case, it is crucial to consider carefully how this lack of randomness will affect the data.

Independence

- It is quite common to assume that the elements of a random sample are independent (and following the same distribution).
- However, suitable modifications of this **simple independent random sampling scheme** may be considered and therefore basic methods have to be modified or extended to handle such deviations from the basic experimental framework.
- It may happen that the lack of independence is due to the fact that sampled units are close in time or space or belong to the same cluster, such as in the case of subjects from the same street or from the same household.
- Whenever the randomness is guaranteed, if the data presents anyway temporal, spatial or cluster dependence, specific statistical models and methods have to be considered.

Checks for normality

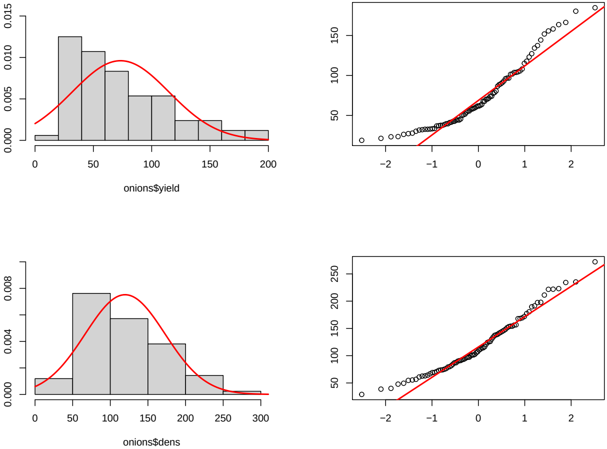
- Many data analysis methods are based on the assumption (at times implicit) that the data are normally distributed.
- Real data are never exactly normally distributed, being the normal assumption at best approximate (usually after a suitable data transformation).
- Large departures from normality is worrisome, whereas small departures are usually of no consequence.
- Things to check are **skewness**, **heavy** or **thin tails**, **outliers** and **undue discreteness** (as that caused by excessive rounding).
- With modest-sized samples, only gross departures will be detectable, and not even them for very small samples (with size 10 or less).
- Graphical tools for checking for normality: histograms are usually not effective and a better tool for assessing normality is the **normal probability plot** (**quantile-quantile plot**).
- Formal statistical tests for normality may also be considered.

Quantile-quantile plot

- For a normal probability plot, the data are sorted and then plotted against the ordered values to be expected if the data were from a normal distribution; namely, the observed quantiles are plotted against the theoretical normal quantiles.
- In case the data actually are from a normal distribution, with any mean and standard deviation, the plot should be approximately a straight line.
- It is actually useful to add a line that passes through two given quantiles (such as the 1st and 3rd quartiles) to help the eye to assess the linearity.
- The same idea can actually be employed for any interest distribution, other than the normal one, by plotting the sorted data against the ordered values that might be expected from the relevant distribution.

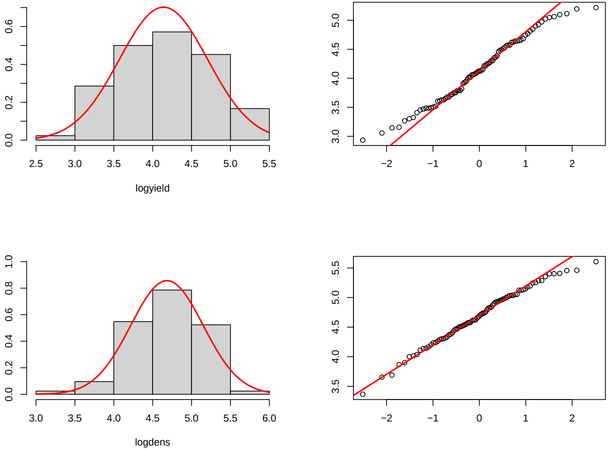
Example: onions

Data from an experiment on the production of white spanish onions in two South Australian locations: $n = 84$ observations for `dens`, areal density of plants (plants per m^2) and `yield`, onion yield (gr per plant)



The original observations for `dens` and `yield` show departures from normality.

The normal distribution assumption seems more plausible for the log-transformed data `logdens` and `logyield`



Why models matter: the Simpson's paradox

Statistical models are important because they can consider all the relevant information simultaneously, in a way that simple summaries do not allow for.

Let us consider data on the admission frequencies by `gender`, for the six largest departments at the University of California at Berkeley in 1973: frequencies classified by `admission status`, `gender` and `department`.

The focus concerns evidence, across the University as a whole, of sex-based discrimination.

Marginal admission rates for males and females

	admission	status	
gender	admitted	rejected	% admitted
male	1198	1493	44.5
female	557	1278	30.4

Apparently, female were discriminated, and this went under the name of *Berkeley gender bias case*.

A look at the results in each department show, however, that no single department was biased against women, as confirmed by the marginal admission rates for males and females for the six departments

	dept					
gender	A	B	C	D	E	F
male	62.1	63.0	36.9	33.1	27.7	5.9
female	82.4	68.0	34.1	34.9	23.9	7.0

As a fraction of those who applied, females were strongly favored in department A, and males somewhat favored in departments C and E.

The explanation of this paradox is in the different proportions of department applications for males and females, as described in the following table.

	dept					
gender	A	B	C	D	E	F
male	30.7	20.8	12.1	15.5	7.1	13.9
female	5.9	1.4	32.3	20.4	21.4	18.6

The overall bias arose because males favored departments where there were a relatively larger number of places, such as departments A and B.

This is just an instance of the **Simpson's paradox**, which refers to the fact that a relationship between two variables may change when the data are partitioned in subgroups, namely when another variable is taken into account.

Statistical models aim at considering all the relevant variables simultaneously and then they could be are the right tool to avoid such pitfalls, due to unsatisfactory and potentially misleading data summary.