Statistics *

Claudio Agostinelli
claudio@unive.it
Dipartimento di Scienze Ambientali, Informatica e Statistica
Università Ca' Foscari di Venezia
San Giobbe, Cannaregio 873, Venezia
Tel. 041 2347446, Fax. 041 2347444
http://www.dst.unive.it/~claudio

February 5, 2013

Contents

*Notes from the Probability and Statistics 2's class in the Computer Science program. Most of the materials is taken from Pawitan [2001].

1 Probability review

In this lecture we review very briefly some elementary concepts on probability theory we will use during the class.

1.1 Probability Space

Definition 1 (Random Experiment)**.** An experiment (or trial) is said to be a **random experiment**, for a certain individual, in a particular situation, if the individual is not able with certainty to indicate the result (the outcome), independently of the fact the experiment is already done or not.

Remarks:

- The result of a trial is an **outcome** (i.e. one of a specific set of possibilities).
- An **event** is a collection of outcomes that have prescribed characteristics.
- A random experiment could be:
  - unique
  - repeatable

Probability Space

Definition 2. A **Probability Space** is a triple $(\Omega, \mathcal{A}, \mathrm{P})$ where

Ω is a sample space;

$\mathcal{A}$ is a σ–algebra of subsets of Ω;

P is a probability measure.

Ω: Sample space

Definition 3. The sample space Ω (sometime called basic space) is the set of all possible outcomes ω of a random experiment.

Remark:

- There may be several possible way of specifying a sample space for the same random experiment; this dependes on the events we are interesting. In any case they lead to the same results.

Example 4. A random experiment may consist in drawing a dice with 6 faces. The possible outcomes are the numbers of the faces, that is, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$ and a natural way of defining the sample space Ω is to take the union of all possible outcomes, that is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

$\mathcal{A}$: σ–algebra

We are not going to use this concept during the preliminary on statistics, we will come back to it during the probability class.

Definition 5 (σ–algebra)**.** A class $\mathcal{A}$ of subsets of Ω is a σ–algebra (sometile called a σ–field) if

1. $\Omega \in \mathcal{A}$;
2. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$;
3. If $A_1, \cdots, A_i, \cdots \in \mathcal{A}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Remark:

- Every set in the class (and only these sets) is call an **event**.

Example 6. For $\Omega = \{1, 2, 3, 4, 5, 6\}$

- $\mathcal{A}_1 = \{\emptyset, \Omega\}$;
- $\mathcal{A}_2 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$;
- $\mathcal{A}_3 = \mathcal{P}(\Omega)$[1]

are examples of σ–algebra on Ω. Notice that, while the set $A = \{\text{an even number}\} = \{2, 4, 6\}$ is an event on the space $\{\Omega, \mathcal{A}_2\}$ and $\{\Omega, \mathcal{A}_3\}$ it is not in $\{\Omega, \mathcal{A}_1\}$.

P: Probability measure

Definition 7. Given a measurable space $(\Omega, \mathcal{A})$, a **Probability measure** P is an application $\mathrm{P} : \mathcal{A} \to \mathbb{R}^+$ such that

1. (not negative) If $A \in \mathcal{A}$ then $\mathrm{P}(A) \geq 0$;
2. (normalization) $\mathrm{P}(\Omega) = 1$;
3. (σ–additive) If $\{A_i\}_{i=1}^{\infty}$ is a sequence of events (elements of $\mathcal{A}$) pairwise incompatible (that is $A_i \cap A_j = \emptyset, i \neq j$), then

$$\mathrm{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

This is the assiomatic definition of Kolmogorov.

Example 8. Given the space $(\Omega, \mathcal{P}(\Omega))$ a probability measure P is

$$\mathrm{P}(\{i\}) = \frac{1}{6} \qquad i = 1, 2, \cdots, 6$$

[1] we use the symbol $\mathcal{P}$ to indicate the power set, that is the class of all possible subset of a set.

Example 9. A coin is tossed until we obtain a first head. Then the number of tails could be $0, 1, 2, \cdots$. This random experiment could be describe by a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ as follows

- $\Omega = \{0, 1, 2, \cdots\}$;
- $\mathcal{A} = \mathcal{P}(\Omega)$;
- $\mathrm{P}(\{n\}) = \frac{1}{2^{n+1}}$ that is the probability of having n tails before the first head.

We are interesting on the probability of the event $\{n \text{ is even}\}$.

Since $\mathcal{A} = \mathcal{P}(\Omega)$ then $\{n \text{ sia pari}\}$ is an event. Further

$$\{n \text{ is even}\} = \{0\} \cup \{2\} \cup \{4\} \cup \{6\} \cup \cdots$$

then

$$\mathrm{P}(\{n \text{ is even}\}) = \mathrm{P}(\{0\} \cup \{2\} \cup \{4\} \cup \{6\} \cup \cdots)$$

For the aim it is simpler to consider the probabiliy of the complement event, that is

$$\mathrm{P}(\{n \text{ is odd}\}) = \mathrm{P}(\{1\} \cup \{3\} \cup \{5\} \cup \cdots)$$

In fact,

$$\begin{aligned} \mathrm{P}(\{n \text{ is odd}\}) &= \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \frac{1}{2^8} + \cdots \\ &= \frac{1}{1-q} - 1 = \frac{4}{3} - 1 = \frac{1}{3} \end{aligned}$$

with $q = 1/4$, since it is a geometric progression with $q = 1/4$ where the first term is missed (1). Since $\{n \text{ sia pari}\} = \Omega / \{n \text{ sia dispari}\}$ we have

$$\mathrm{P}(\{n \text{ is even}\}) = 1 - \frac{1}{3} = \frac{2}{3} \, .$$

1.2 Conditional Probability

Conditional Probability

Definition 10. Let $(\Omega, \mathcal{A}), \mathrm{P})$ be a probability space. Let H be an event (an element of $\mathcal{A}$), with $\mathrm{P}(H) \neq 0$. The **Conditional Probability** derived by P under the condition "H holds" is the probability function P_H on the space $(\Omega, \mathcal{A})$

$$\mathrm{P}_H(A) = \frac{\mathrm{P}(A \cap H)}{\mathrm{P}(H)}$$

for each event $A \in \mathcal{A}$.

The probabiliy $\mathrm{P}_H(A)$ is the **Conditional Probability** of A, derived by P, given H. We denote this probability as

$$\mathrm{P}(A|H) \ .$$

Remarks:

- $\mathrm{P}(A|H)$ is a probability function concentrated on H. It coincides with P in the particular case of $\mathrm{P}(H) = 1$ that is H is an almost sure event.
- $\mathrm{P}_H(A) = \mathrm{P}(A|H)$ is a probability function.

After H happened, each event $A \in \mathcal{A}$ such that $A \cap H = \emptyset$ we have $\mathrm{P}_H(A) = \mathrm{P}(A|H) = 0$.

Hence, two equivalent probability space for the same problem are

$$(\Omega, \mathcal{A}, \mathrm{P}_H) \qquad (\Omega_H, \mathcal{A}_H, \mathrm{P}_H)$$

where

$$\Omega_H = \{A \subseteq H : A \in \Omega\} = H$$

since the outcomes (elementary events) are only the ones implying H and

$$\mathcal{A}_H = \{A \cap H : A \in \mathcal{A}\}$$

Definition 11 (Conditional Probability Space)**.** The probability space $(\Omega_H, \mathcal{A}, \mathrm{P}_H)$ is called the **Conditional Probability Space**.

Example 12. A dice is run twice. W know, only, that the sum of the two outcomes is 6. What is the probability of having 3 in the first run?

Let H the event "the sum of the two outcomes is 6" and A the event "face 3 in the first run" hence we have

$$H = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$$
$$A = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\}$$

from which $\mathrm{P}(H) = \frac{5}{36}$ and $\mathrm{P}(A) = \frac{6}{36}$. Further $A \cap H = \{(3,3)\}$ and hence $\mathrm{P}(A \cap H) = \frac{1}{36}$.

$$\mathrm{P}(A|H) = \frac{\frac{1}{36}}{\frac{5}{36}} = \frac{1}{5}$$

1.3 Stochastic Independence

Stochastic Independence

Definition 13. Given a probability space $(\Omega, \mathcal{A}, P)$, two events A, B are **stochastic independent** if and only if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B) \ .$$

Remark

- If two events A, B are stochastic independent then

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

and the same holds for $\mathrm{P}(B|A) = \mathrm{P}(B)$.

Example 14. Let us consider the result of tossing a dice. As usual we have $\Omega = \{1, 2, 3, 4, 5, 6\}$. If $\mathrm{P}(\{\omega\}) = \frac{1}{6}$ for $\omega \in \Omega$, and let $A = \{2, 4, 6\}$ and $B = \{3, 6\}$, then

$$A \cap B = \{6\},\ \mathrm{P}(A) = \frac{1}{2},\ \mathrm{P}(B) = \frac{1}{3} \text{ e } \mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B) = \frac{1}{6} \ .$$

Hence A and B are stochastic independent.

Definition 15. Given n events $A_1, A_2, \cdots, A_n$ from the probability space $(\Omega, \mathcal{A}, P)$, they are **stochastic independent** if and only if

$$\mathrm{P}(A_{i_1} \cap A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathrm{P}(A_{i_1}) \cdot \mathrm{P}(A_{i_2}) \cdots \mathrm{P}(A_{i_n}) \ .$$

for each $k = 2, 3, \cdots, n$ and for each subset $(i_1 < i_2 < \cdots < i_k)$ of the integers $1, 2, \cdots, n$.

1.4 Random Variables

1.4.1 Univariate random variables

Definition 16 (Random Variable)**.** Given a measurable space $(\Omega, \mathcal{A})$, a **real random variable** $X(\omega) : \Omega \to \mathbb{R}$ is a real function such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A} \ \text{ for each real value } x. \tag{1}$$

Remarks:

- the probability function does not play a role in the definition of a random variable;
- when $\mathcal{A}$ is the power set of Ω then the condition (1) holds.
- the random variable $X(\omega)$ is defined on the space $(\mathbb{R}, \mathcal{B})$ for an appropriate σ–algebra $\mathcal{B}$. Often this σ–algebra is the Borel σ–algebra ($\mathcal{B}(\mathbb{R})$). We are not going on in these details here.

Probability function and Random Variable

Let us consider the interval $(a, b]$ and the set $A \in \mathcal{A}$ such that

$$A = \{\omega \in \Omega : a < X(\omega) \leq b\} \in \mathcal{A} \ .$$

They are equivalent in the sense they implying each others, since when A holds, that is $\omega \in \mathcal{A}$, then $X \in (a, b]$ and viceversa. Since the event A has probability $\mathrm{P}(A)$, we can let, for each couple $a < b$,

$$\mathrm{P}_X((a, b]) = \mathrm{P}(X \in (a, b]) = \mathrm{P}(\{\omega \in \Omega : a < X \leq b\}) \ .$$

The function P_X is a probability function defined on the space $(\mathbb{R}, \mathcal{B}(\mathbb{R})$ and it is call the **Law** of the random variable X and given it we can determined the probability $\mathrm{P}_X(B) = \mathrm{P}(X \in B)$ for each $B \in \mathcal{B}(\mathbb{R})$.

Distribution Function

Definition 17. Let X a real random variable. The **Probability Distribution Function** of X is the function $y = F(x)$, defined for every real number x, such that

$$F(x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathrm{P}(X \leq x) \qquad x \in \mathbb{R} \ .$$

Discrete Random Variable

Definition 18. A r.v. X on the space $(\Omega, \mathcal{A})$ is a **discrete random variable** if the set $R_X = \cup_{\omega \in \Omega}\{X(\omega)\}$ is countable or finite. R_X is called the **support** of the r.v. X.

If X is a discrete random variable with support $R_X = \{x_1, x_2, \cdots\}$, then its Law (often call the discrete probability density function) is a real function $p(x)$ such that

$$p(x) = \begin{cases} \mathrm{P}(X = x_i) > 0 & x = x_i \in R_X \\ 0 & x \notin R_X \end{cases}$$

Connection between probability function and distribution function

For a discrete random variable X we have

$$\mathrm{p}(X = x) = F(x) - F(x^-)$$

Binomial distribution

This distribution describe the number of success in a sequence of fixed and finite number of independent trials.

Let us consider an experiment where the possible results are $\Omega = \{A, A^c\}$ (A="success", A^c ="unsuccess") and that this experiment will replicate for $N \geq 1$ times. At each trial we link a r.v. X_i, $i = 1, 2, \cdots, N$ so that $X_i = 1$ when A holds and $X_i = 0$ when A^c holds. Further, since the problem we have $\mathrm{P}_{\mathrm{X_i}}(X_i = 1) = \mathrm{P}(A) = p$, $0 \leq p \leq 1$. What is the probability of the total number of succecces in the N trials. That is, what is the Law of the r.v.

$$X = X_1 + X_2 + X_3 + \cdots + X_N \ ?$$

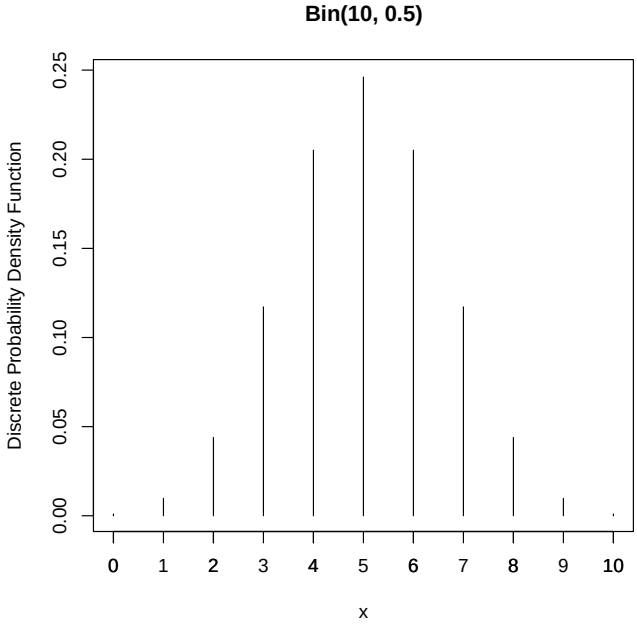
Binomial Distribution

Definition 19 (Binomial Distribution)**.** The r.v. X is a **Binomial Distribution** Law with parameters $N \geq 1$ (number of trials) and $0 \leq p \leq 1$ (probability of success in each trial), if

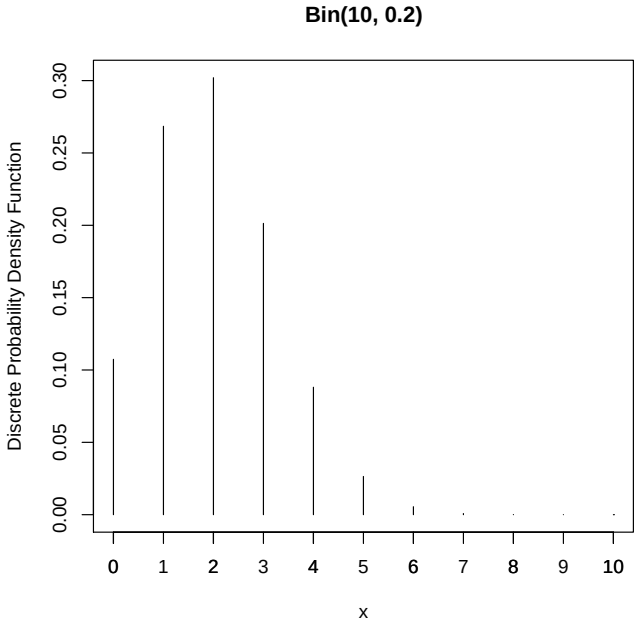
$$p(x) = \mathrm{P}(X = x) = \begin{cases} \binom{N}{x} p^x (1-p)^{N-x} & x = 0, 1, \cdots, N \\ 0 & \text{otherwise .} \end{cases}$$

we denote it $X \sim \mathrm{Bin}(N, p)$.

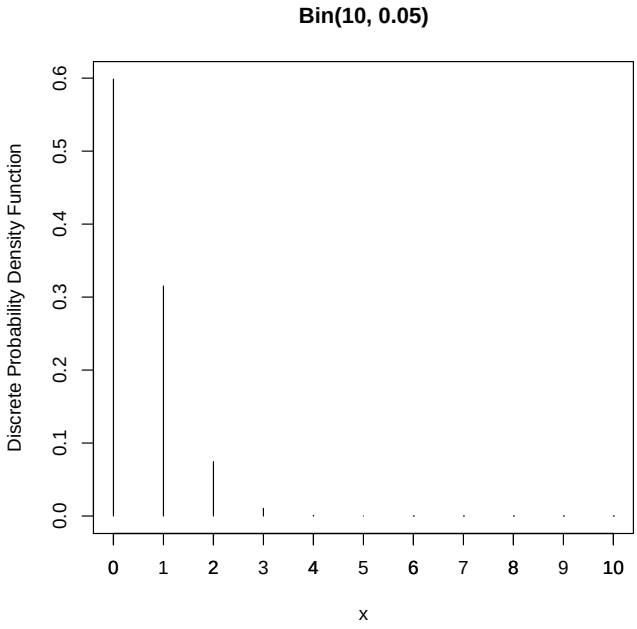
The special case $N = 1$ is called a **Bernoulli** random variable.



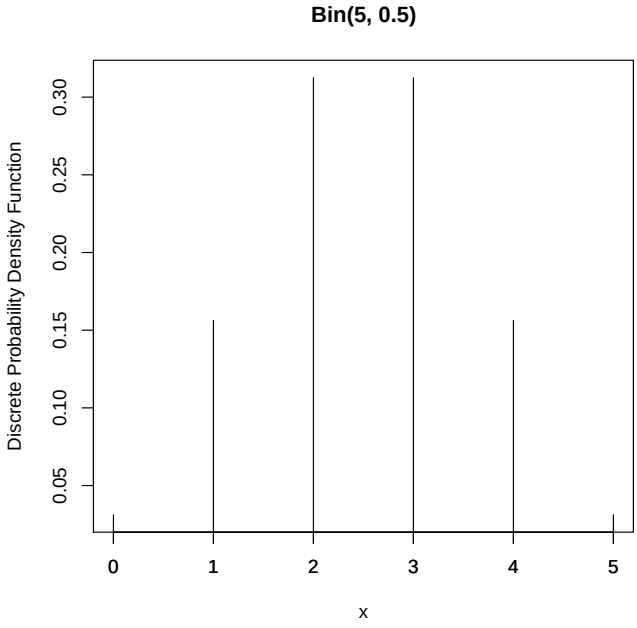
Probability Density Function for a Binomial random variable $\text{Bin}(10, 0.5)$.



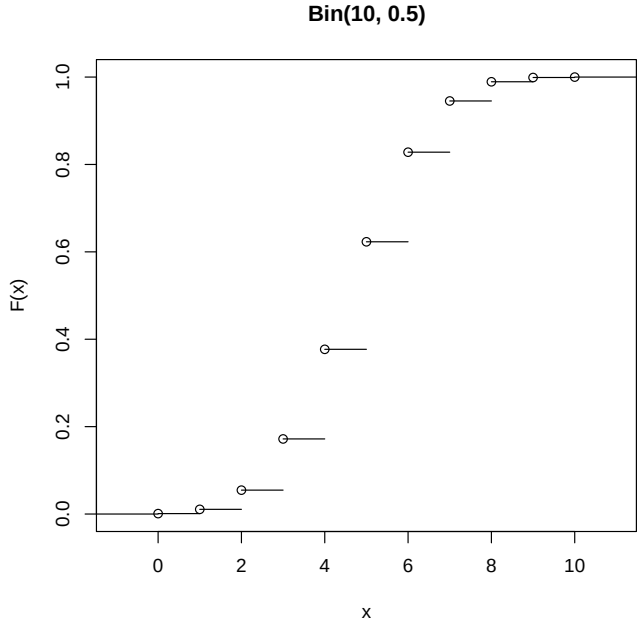
Probability Density Function for a Binomial random variable $\text{Bin}(10, 0.2)$.



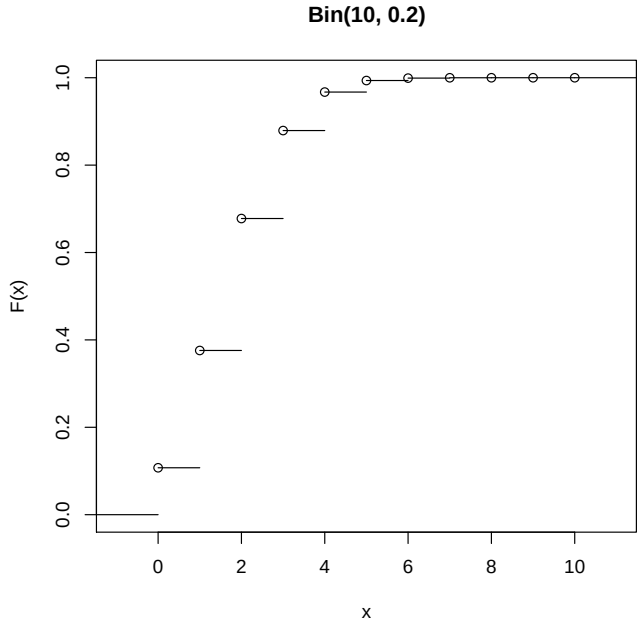
Probability Density Function for a Binomial random variable $\text{Bin}(10, 0.05)$.



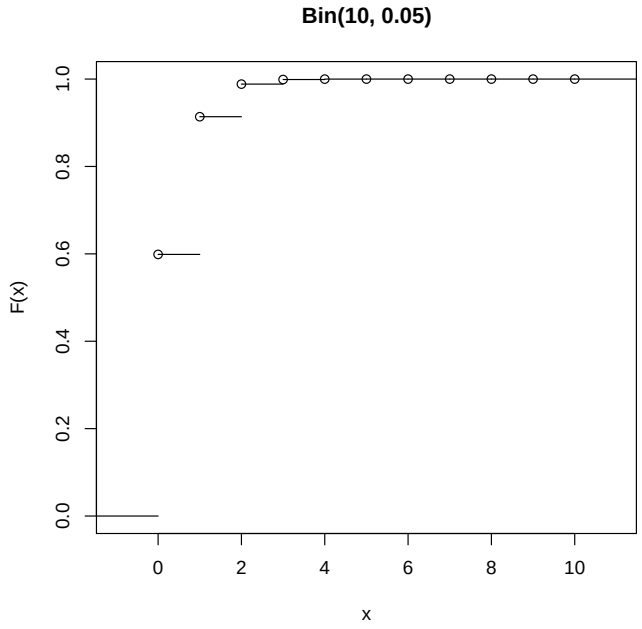
Probability Density Function for a Binomial random variable $\text{Bin}(5, 0.5)$.



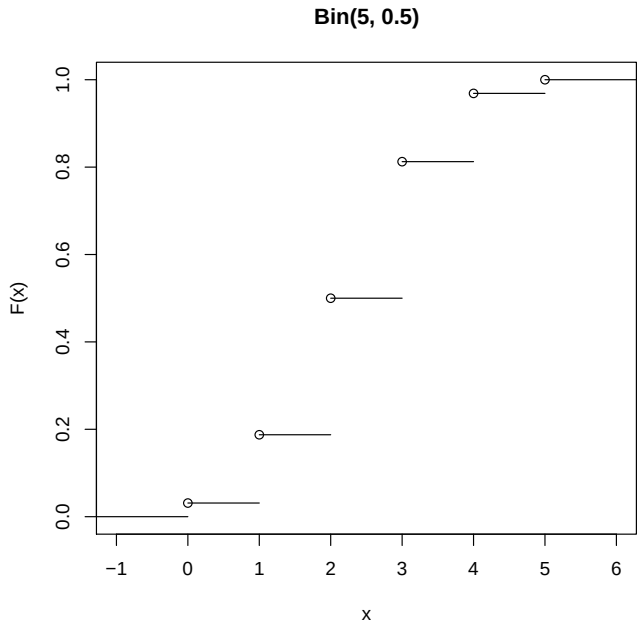
Distribution Function for a Binomial random variable $\text{Bin}(10, 0.5)$.



Distribution Function for a Binomial random variable $\text{Bin}(10, 0.2)$.



Distribution Function for a Binomial random variable $\text{Bin}(10, 0.05)$.



Distribution Function for a Binomial random variable $\text{Bin}(5, 0.5)$.

Absolutely Continuous Random Variable

Definition 20 (Probability Density Function)**.** A r.v. X has a **Probability Density Function** or briefly Density if the probability of events $X \in (a, b]$ could be described as

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

where $f(x)$ is the **density** of the r.v. X with the following properties

- $f(x) \geq 0$ for each $x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x)\ dx = 1$

Density and Distribution Function

For a r.v. X with density $f(x)$ and distribution function $F(x)$ we have

$$\begin{aligned} \mathrm{P}(X \in (a, b]) &= \int_a^b f(x)\ dx \\ &= \int_{-\infty}^b f(x)\ dx - \int_{-\infty}^a f(x)\ dx \\ &= \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) \\ &= F(b) - F(a) \end{aligned}$$

Density and Distribution Function

If the distribution function $F(x)$ is derivable then

$$f(x) = F'(x) = \frac{\partial}{\partial x} F(x)$$

and

$$F(x) = \int_{-\infty}^x f(t)\ dt$$

Example 21. Let $\Omega = [0, 1]$ and $\mathcal{A} = \mathcal{B}([0, 1])$. Let the probability function P on $\mathcal{A}$ defined by the distribution function $F(x)$

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

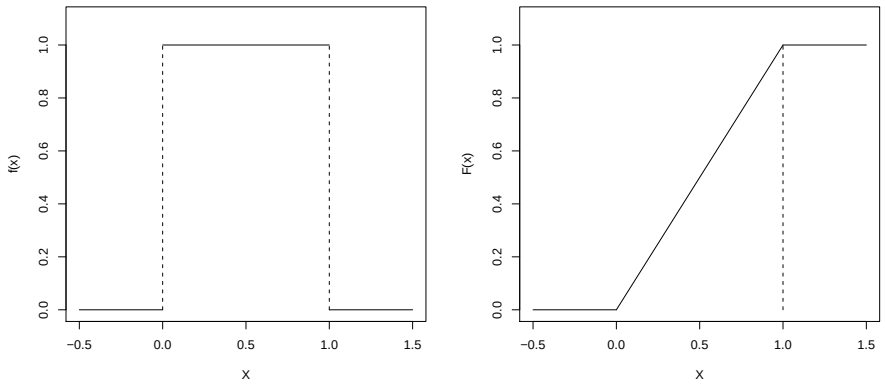
Let us consider the identity random variable $X(\omega) = \omega$ and hence

$$\begin{aligned} \mathrm{P}(a < X \leq b) &= \mathrm{P}(\{\omega \in [0,1] : a < X(\omega) \leq b\}) \\ &= \mathrm{P}(\{\omega \in [0,1] : a < \omega \leq b\}) \\ &= \mathrm{P}((a,b]) \end{aligned}$$

$$= \begin{cases} b - a = \int_a^b 1 \, dx & 0 \leq a < b \leq 1 \\ b & a < 0 < b \\ 0 & a < b < 0 \end{cases}$$

The density of X is

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



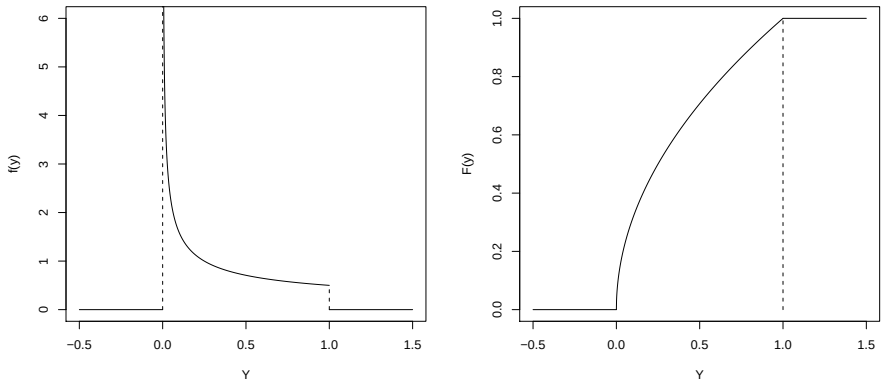
Left: Density and Right: Distribution, of the r.v. X.

Example 22. On the same space we can define the r.v. $Y(\omega) = \omega^2$, in this case we have

$$\begin{aligned} \mathrm{P}(a < Y \leq b) &= \mathrm{P}(\{\omega \in [0,1] : a < Y(\omega) \leq b\}) \\ &= \mathrm{P}(\{\omega \in [0,1] : a < \omega^2 \leq b\}) \\ &= \mathrm{P}(\{\omega \in [0,1] : \sqrt{a} < \omega \leq \sqrt{b}\}) \\ &= \mathrm{P}((\sqrt{a}, \sqrt{b}]) \\ &= \begin{cases} \sqrt{b} - \sqrt{a} = \int_a^b \frac{1}{2\sqrt{x}} \, dx & 0 \leq a < b \leq 1 \\ \sqrt{b} & a < 0 < b \\ 0 & a < b < 0 \end{cases} \end{aligned}$$

And the density of Y is

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



Left: Density and Right: Distribution, of the r.v. Y.

Normal (or Gaussian) Distribution

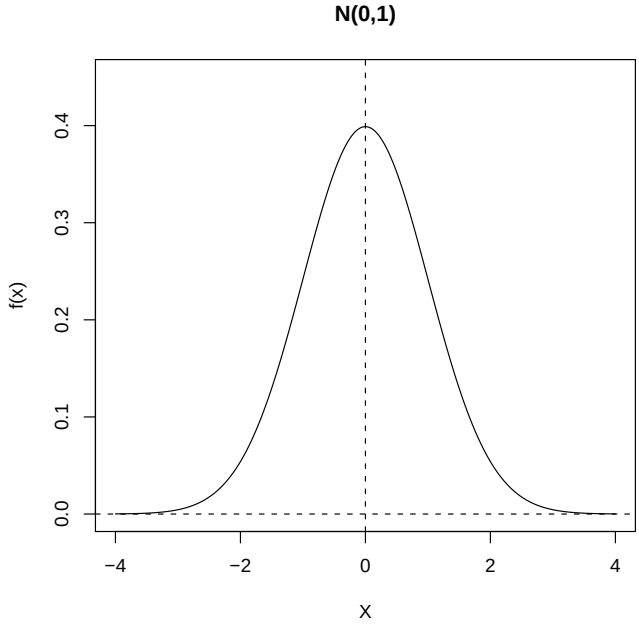
Definition 23 (Normal Distribution)**.** A r.v. X has a **Normal Distribution** with parameters $-\infty < \mu < +\infty$ (location and mean) and $0 < \sigma < +\infty$ (scale and variance) if its density is

$$m(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \qquad -\infty < x < +\infty$$

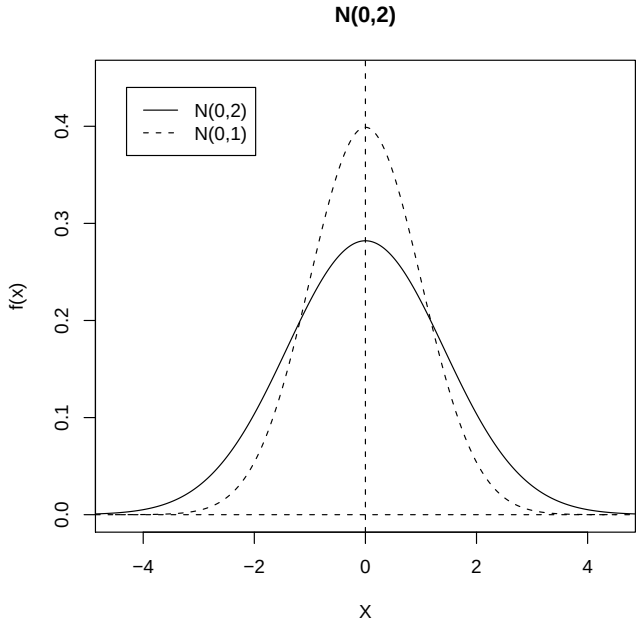
and we denote it by $X \sim \mathrm{N}(\mu, \sigma^2)$

Remark:

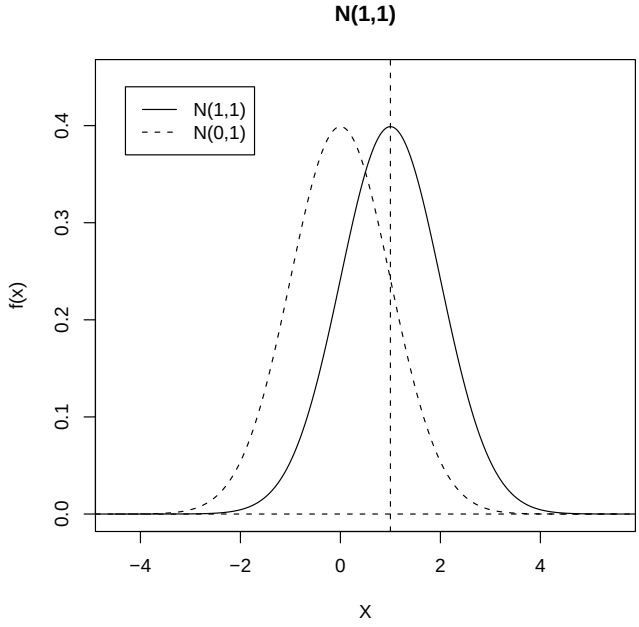
- A r.v. $X \sim \mathrm{N}(0, 1)$ is called a **Standard Normal**;
- ϕ and Φ are often used for the density and the distribution function for the Standard Normal.



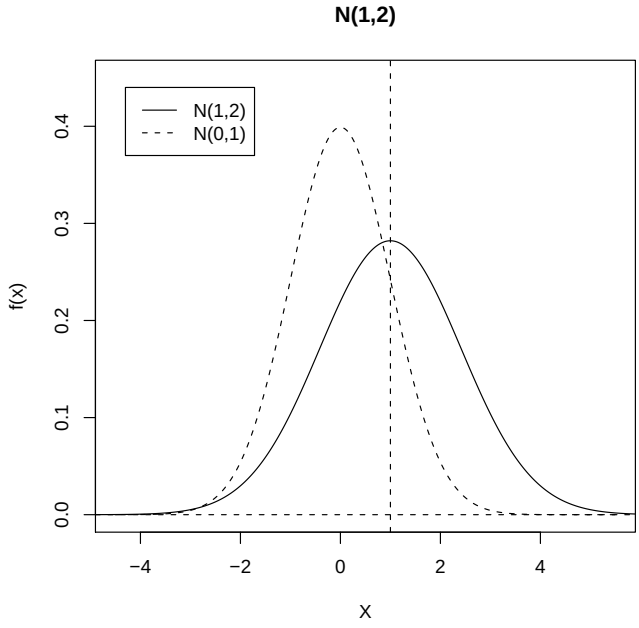
Density ϕ of the Standard Normal $\mathrm{N}(0,1)$.



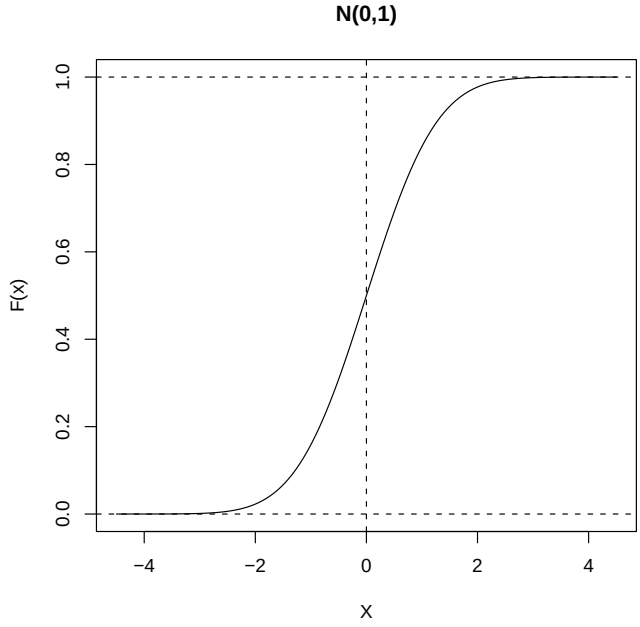
Comparison between the densities of the $\mathrm{N}(0,1)$ and $\mathrm{N}(0,2)$.



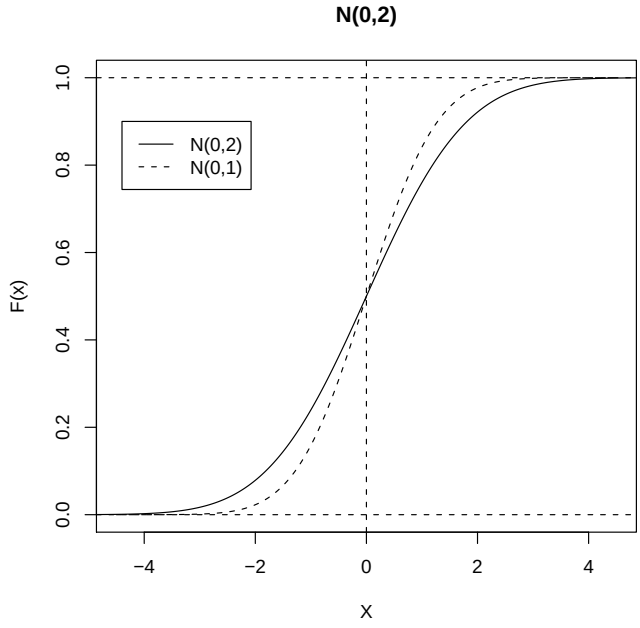
Comparison between the densities of the $\mathrm{N}(0,1)$ and $\mathrm{N}(1,1)$.



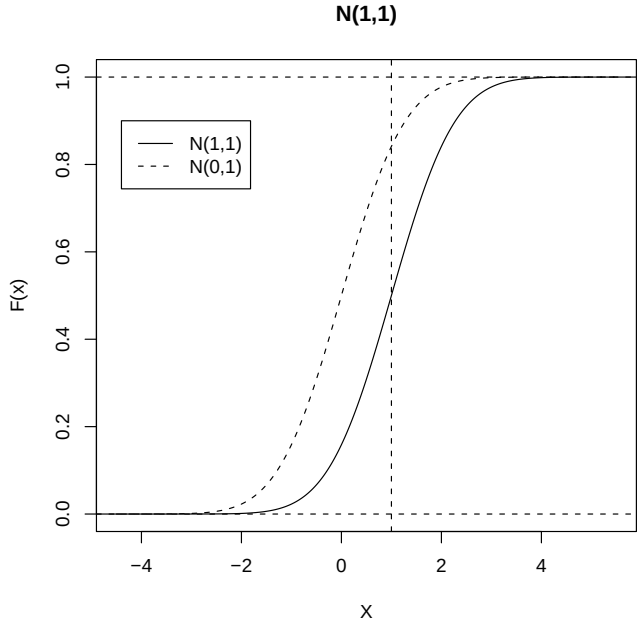
Comparison between the densities of the $\mathrm{N}(0,1)$ and $\mathrm{N}(1,2)$.



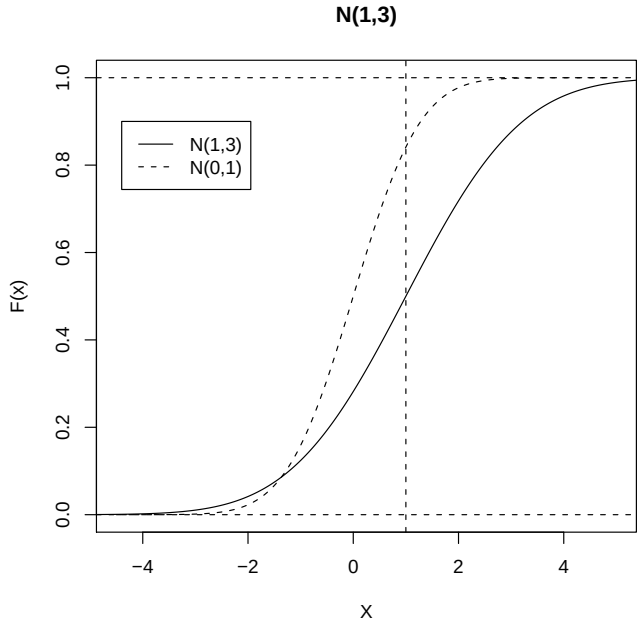
Distribution Φ of the Standard Normal $\mathrm{N}(0,1)$.



Comparison between the distributions of the $\mathrm{N}(0,1)$ and $\mathrm{N}(0,2)$.



Comparison between the distributions of the $\mathrm{N}(0,1)$ and $\mathrm{N}(1,1)$.



Comparison between the distributions of the $\mathrm{N}(0,1)$ and $\mathrm{N}(1,2)$.

Quantile

Definition 24 (Quantile)**.** A *quantile* of order $0 \leq \alpha \leq 1$ of a r.v. X is the value $Q_\alpha(X) = x_\alpha$ such that

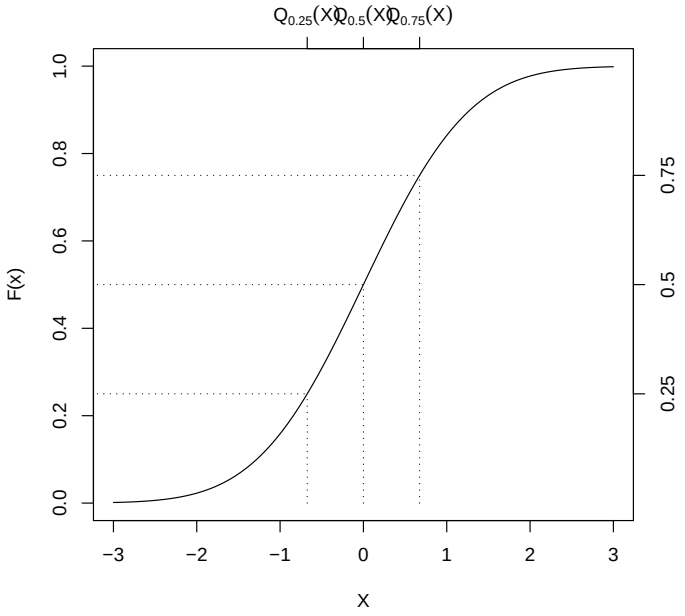
$$\mathrm{P}(X \leq x_\alpha) \geq \alpha \text{ and } \mathrm{P}(X \geq x_\alpha) \geq 1 - \alpha$$

or in different form

$$F(x_\alpha) \geq \alpha \text{ and } 1 - F(x_\alpha) \geq 1 - \alpha$$

Remark:

- **Quartile**:
  - 1^o quartile: $q_1 = \mathrm{Q}_{0.25}(X) = x_{0.25}$,
  - median: $\mathrm{Me}(X) = q_2 = \mathrm{Q}_{0.5}(X) = x_{0.5}$,
  - 3^o quartile: $q_3 = \mathrm{Q}_{0.75}(X) = x_{0.75}$
  - 4^o quartile: $q_4 = \mathrm{Q}_1(X) = x_1$.



Quartiles for a Standard Normal random variable.

Expected value for a discrete random variable

Definition 25. Let X be a discrete random variable with Law $p_X(x)$. We define the **Expected Value** of X the (finite) quantity

$$\mathrm{E}(X) = \sum_{x \in R_X} x \; p_X(x)$$

if

$$\mathrm{E}(|X|) = \sum_{x \in R_X} |x| \; p_X(x) < \infty$$

Expected value for a continuous random variable

Definition 26. Let X be a continuous random variable with density $f_X(x)$ and distribution function $F_X(x)$. We define the **Expected Value** of X the (finite) quantity

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x \; f_X(x) \; dx = \int_{-\infty}^{+\infty} x \; dF_X(x)$$

if

$$\mathrm{E}(|X|) = \int_{-\infty}^{+\infty} |x| \; f_X(x) \; dx < \infty$$

Expected value for a function of a random variable

Let Y and X such that there exists a function g so that $Y = g(X)$ then

$$\mathrm{E}(Y) = \mathrm{E}(g(X)) = \begin{cases} \sum_{y \in R_Y} y p_Y(y) & = \sum_{x \in R_X} g(x) p_X(x) \\ & \text{if } Y \text{ is a discrete r.v.} \\ \int_{R_Y} y \; dF_Y(y) & = \int_{R_X} g(x) \; dF_X(x) \\ & \text{if } Y \text{ is a r.v. with density} \end{cases}$$

Moments

Definition 27. Given a r.v. X we will call

$$\mu_r = \mathrm{E}(X^r)$$

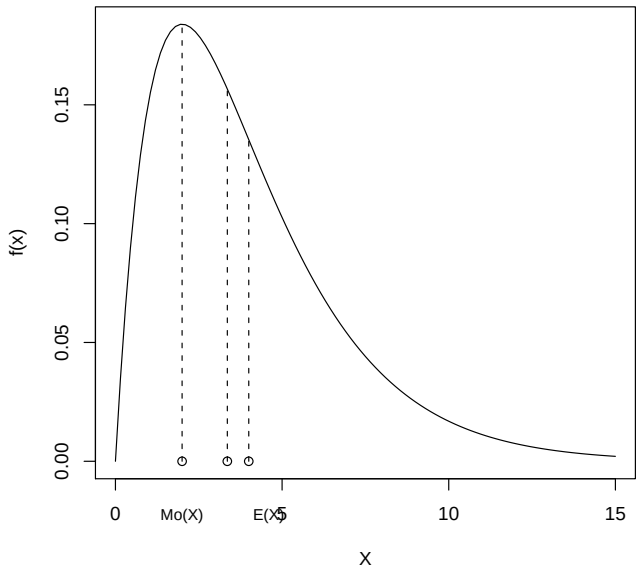
the non–centered moments of order r (positive integer) while

$$\bar{\mu}_r = \mathrm{E}((X - \mu_1)^r)$$

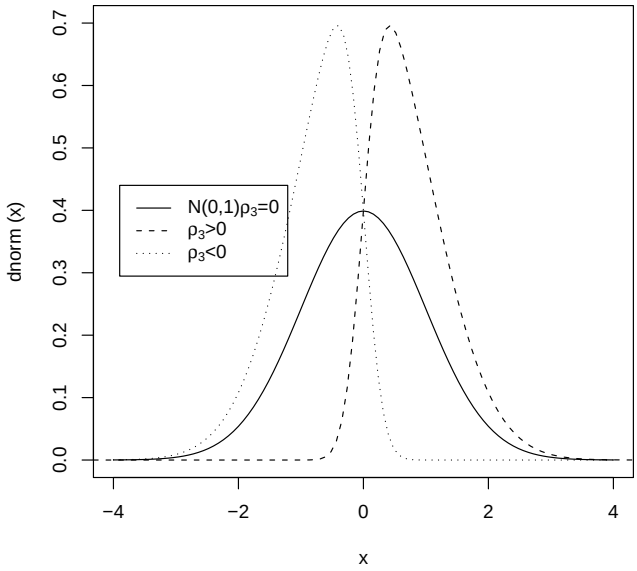
is the centered (from the mean) moments of order r.

Most common summary values based on the moments

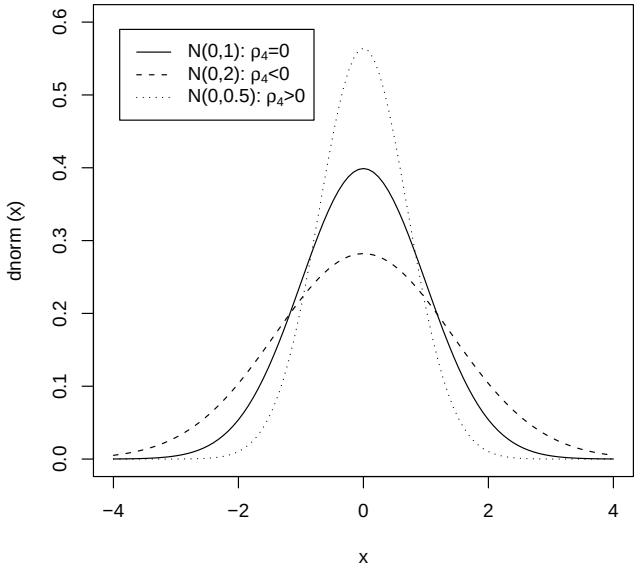
- Mean: $\mu = \mu_1 = \mathrm{E}(X)$
- Variance: $V(X) = \sigma^2 = \bar{\mu}_2 = \mathrm{E}((X - \mu_1)^2)$
- Standard Deviation: $\sigma = \sqrt{\sigma^2}$
- Variation coefficient: $CV = \frac{\sigma}{\mu}$
- Asymmetry index: $\rho_3 = \frac{\bar{\mu}_3}{\sigma^3}$
- Kurtosis index: $\rho_4 = \frac{\bar{\mu}_4}{\sigma^4} - 3$



Mean, Median, Mode.



Asymmetry index



Kurtosis index

1.4.2 Multivariate Random Variables

Definition 28. Let $(\Omega, \mathcal{A}, P)$ a probability space. Let $X_i(\omega)$ $(i = 1, \cdots, n)$, n random variables in this space such that

$$\mathbf{X}(\omega) = (X_1(\omega), \cdots, X_n(\omega)) : \Omega \to \mathbb{R}^n$$

then $\mathbf{X}$ is called a multivariate random variable of dimension n.

Bivariate random variable

We restrict, for simplicity, our attention to random variables for $n = 2$.

Let $Z(\omega) = (X(\omega), Y(\omega))$ a bivariate random variable. It is not enough in general to consider $F_X(x)$ and $F_Y(y)$ for defining the (join) distribution function $F_Z(z)$ but

$$F_Z(z) = F_{X,Y}(x, y) = \mathrm{P}(\{X \leq x\} \cap \{Y \leq y\}) \qquad (x, y) \in R_{X,Y}$$

Bivariate discrete random variable

Definition 29. For two given discrete random variables X and Y, the bivariate random variable $Z = (X, Y)$ (which is discrete) has the (join) Law

$$p_Z(z) = \begin{cases} p_{X,Y}(x, y) = & \mathrm{P}(\omega : \{X(\omega) = x\} \cap \{Y(\omega) = y\}) \\ & (x, y) \in R_{X,Y} \\ = & 0 \qquad \text{otherwise} \end{cases}$$

that is

$$\begin{aligned} \mathrm{P}(x_1 < X \leq x_2, y_1 < Y \leq y_2) &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) \\ &- F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1) \end{aligned}$$

$$F_{X,Y}(x, y) = \sum_{\{(u,v):u \leq x, v \leq y\}} p_{X,Y}(u, v)$$

Bivariate random variable with density

Definition 30. The bivariate random variable $Z = (X, Y)$ has density if there exists a bivariate function $f_{X,Y}(x, y)$ such that

- $f_{X,Y}(x,y) \geq 0$, $\forall (x,y) \in \mathbb{R}^2$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dx\ dy = 1$$

- $$\mathrm{P}(a < x \leq b, c < y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\ dx\ dy$$

This function is called the (join) density $f_Z(z) = f_{X,Y}(x,y)$.

From which we have

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\ du\ dv$$

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,v)\ dv \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(u,y)\ du$$

and $f_X(x)$ and $f_Y(y)$ are the (marginal) density of X and Y respectively. We have also

$$F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x,y) \qquad F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x,y)$$

Conditional Distribution

Definition 31 (Conditional Probability for the r.v. $X|Y=y$)**.** Let (X,Y) a bivariate discrete random variable with Law

$$p_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y)$$

hence according with the definition of the conditional probability

$$p_{X|Y}(X = x|Y = y) \quad = \mathrm{P}(\{X = x\}|\{Y = y\}) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \qquad y \in R_Y (p_Y(y) > 0)\ .$$

for each fixed value $y \in R_Y$. The function $p_{X|Y}(X = x|Y = y)$ is called the conditional probability of the random variable $X|Y = y$.

Stochastic Independence

Definition 32. Two discrete random variables X and Y are **stochastic independent** if and only if

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y) \qquad \forall (x,y) \in R_{X,Y} = R_X \times R_Y$$

this implies that

$$p_{X|Y}(x|y) = p_X(x), p_{Y|X}(y|x) = p_Y(y) \qquad \forall (x,y) \in R_{X,Y}$$

Stochastic Independence

The same definition hold for a couple of random variable with density as follows

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \forall y : f_Y(y) > 0$$

and two r.v. X, Y are **stochastic independent** if and only if

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

which implies

$$f_{X|Y}(x|y) = f_X(x) \qquad \forall (x,y) \in R_{X,Y}$$

2 Elements of Likelihood Inference

2.1 Likelihood function

From deduction to induction

Example 33. Let us consider the following situation: we have one big box containings many O–rings of three different dimensions (say, $9, 10, 11mm$) with the same proportions. We need 10 O–ring of $10mm$ and we pick 10 (randomly) from the box.

Using the probability theory we might describe (with enough accuracy) our phenomenon through a Binomial model with the following form:

$$X \sim \text{Bin}(n = 10, \theta = \frac{1}{3})$$

where $\{1\}$ ="success" = "we pick an O-ring of $10mm$ and $\{0\}$ ="unsucess" = "we pick an O–ring of $9mm$ or $11mm$.

So that, the probability of having 0 success out of 10 is

$$P(X=0;\theta=\frac{1}{3})=\binom{10}{0}\theta^0(1-\theta)^{10-0}=\left(\frac{2}{3}\right)^{10}=0.017\ .$$

In the same manner for any other value. This is a deductive result yields by the probability theory.

Now, let suppose the proportions of the three O–rings inside the box is unknown. After draw out of the box 10 O–rings we find that 8 are $10mm$ large. The question is "What is the true proportion θ of the $10mm$ large O–rings with respect to the total number of O–rings inside the box?"

We can not follows the same lines, but we can still say that if we knew the value of the proportion θ then

$$P(X=8;\theta)=\binom{10}{8}\theta^8(1-\theta)^{10-8}$$

and so

$$\begin{aligned}
P(X=8;\theta=0) &= \binom{10}{8}0^8 1^{10-8}=0\\
P(X=8;\theta=0.6) &= \binom{10}{8}0.6^8 0.4^{10-8}=0.121\\
P(X=8;\theta=0.7) &= \binom{10}{8}0.7^8 0.3^{10-8}=0.233
\end{aligned}$$

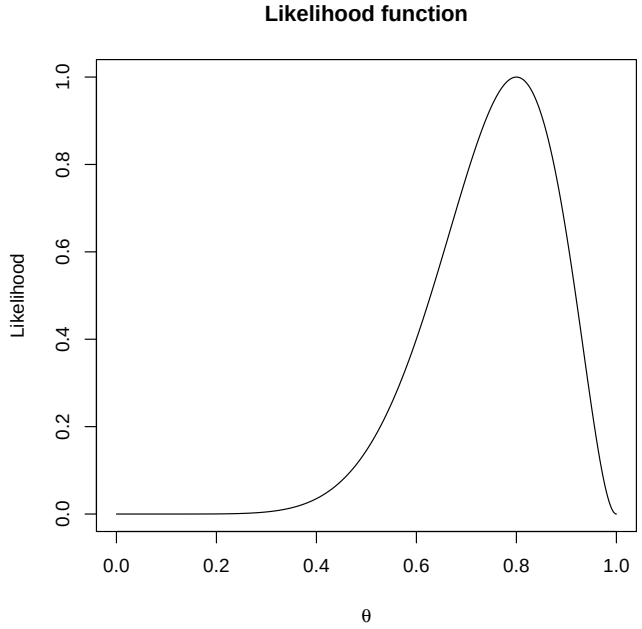
We have thus found a deductive way of comparing different θ's: compared the probability of the observed data under different values of θ.

That is, in a simple and deductive way we have found a *numerical quantity to express the order of preferences on* θ.

This quantity

$$L(\theta)=\mathrm{P}(X=8;\theta)$$

as a function of θ is called the *Likelihood* function.



Likelihood function of the success probability θ in a binomial experiment with $n = 10$ and $x = 8$. The function is normalized to have unit maximum. The Likelihood function plot shows θ is unlikely to be less than 0.5 or to be greater than 0.95, but is more likely to be in between.

Likelihood function

Definition 34. Assuming a statistical model parametrized by a fixed and unknown θ, the **Likelihood function** $L(\theta)$ is the probability of the observed data x considered as a function of θ.

or using Fisher [1922, pag. 310] definition:

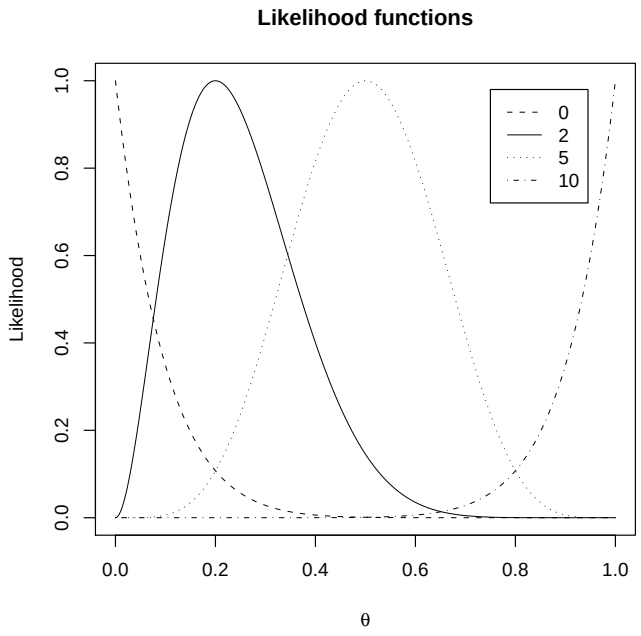
Definition 35. The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

Likelihood in Discrete models

There is no ambiguity about the probability of the observed data in the discrete models, since it is a well–defined nonzero quantity.

Example 36. For the binomial distribution we have

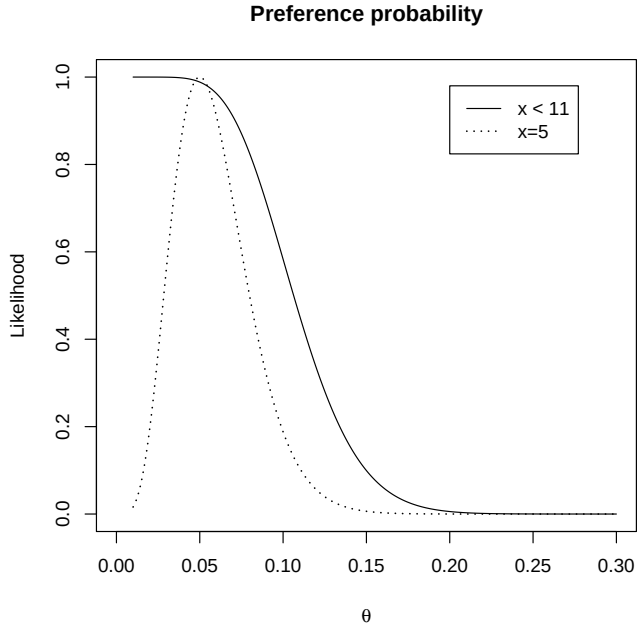
$$\begin{aligned} L(\theta) &= \mathrm{P}(X = x; \theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \end{aligned}$$



Likelihood functions of the success probability θ in four binomial experiments with $n = 10$ and outcomes x, $0, 2, 5, 10$. The functions are normalized to have unit maximum.

Example 37 (New product). Suppose 100 persons are asked if they prefer a new product with respcet to the old one. It is known only that $x \leq 10$ like the new product. The exact number is unknown. Then the information about θ is given by the likelihood function

$$\begin{aligned} L(\theta) &= P(X \leq 10; \theta) \\ &= \sum_{x=0}^{10} \binom{100}{x} \theta^x (1-\theta)^{100-x} \ . \end{aligned}$$



Likelihood functions from two binomial experiments: $n = 100$ and $x \leq 10$, and $n = 100$ and $x = 5$.

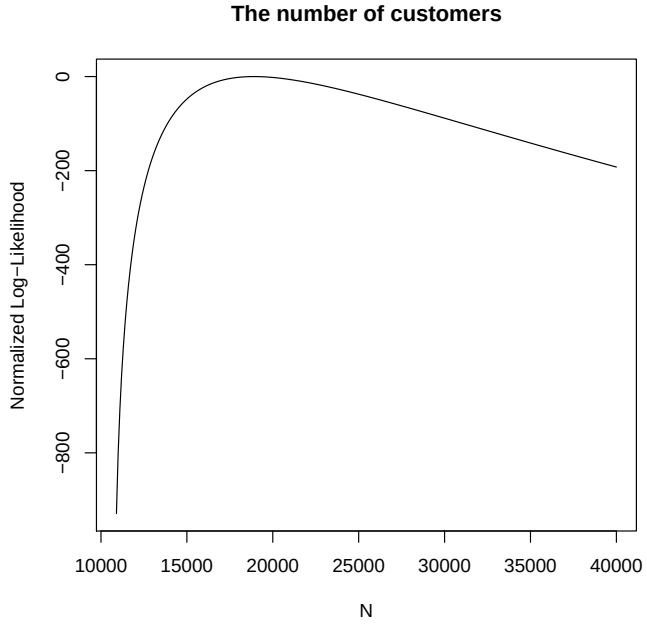
Example 38. A useful technique for counting a population is to mark a subset of the population, then take a random sample from the mixture of marked and unmarked individuals. This capture–recapture technique is used, for example, to count the number of wild animals. In census applications a post–enumeration survey is conducted and one considers the previously counted individuals as "marked" and the new ones as "unmarked"; the proportion of new individuals in the survey would provide an estimate of the undercount during the census.

Example 39 (Fidelity cards). A Megastore use the fidelity cards. Until now, it distributed $N_1 = 10341$ cards which are used (at least one time a month). To estimate the number of customers (N) in a month, it samples $n = 1000$ customers, and finds $n_2 = 453$ without fidelity card and $n_1 = 547$ with fidelity card.

Assuming the customers were sampled at random, the likelihood of N

can be computed based on the hypergeometric probability

$$\begin{aligned} L(N) &= P(n_1 = 547) \\ &= \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}}{\binom{N}{n}} \\ &= \frac{\binom{10341}{547}\binom{N-10341}{453}}{\binom{N}{1000}} \end{aligned}$$



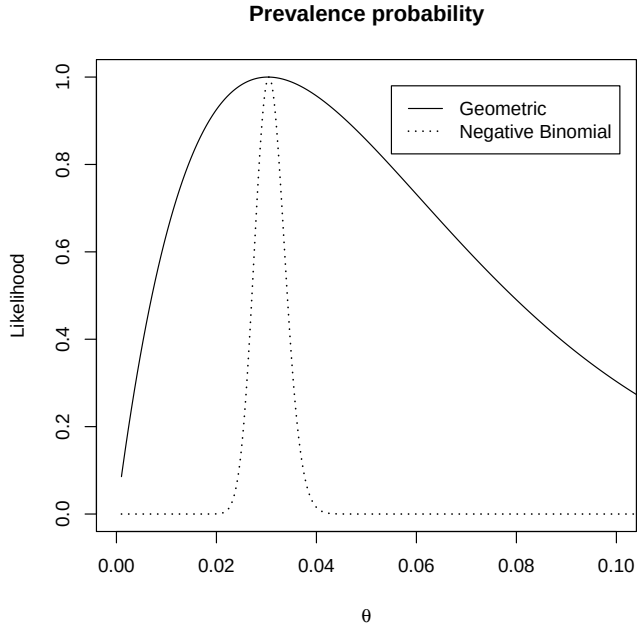
The likelihood of the number of customers.

Example 40 (CATI). A **C**omputer **A**ssisted **T**elephone **I**nterview center has found a person with the requested characteristics after 32 calls. Assuming the subjects are independent, the likelihood of the prevalence probability θ is given by the geometric distribution

$$L(\theta) = (1-\theta)^{32}\theta \ .$$

The center has decided to stop after they found 100 persons with the requested characteristcs, at which point they called 3178 persons. The likelihood of θ is given by a negative binomial distribution

$$L(\theta) = \binom{3178-1}{100-1}\theta^{100}(1-\theta)^{3178-100} \ .$$



The likelihood of the prevalence probability of the person with requested characteristics.

Likelihood in Continuous models

A slight technical issue arises when dealing with continuous outcomes, since theoretically the probability of any point value x is zero. We can resolve this problem by admitting that in real life there is only finite precision: observing x is short for observing $(x - \varepsilon/2, x + \varepsilon/2)$, where ε is the precision limit. If ε is small enough then

$$\begin{aligned} L(\theta) &= \mathrm{P}(X \in (x - \varepsilon/2, x + \varepsilon/2); \theta) \\ &= \int_{x-\varepsilon/2}^{x+\varepsilon/2} m(x;\theta)\ dx \approx \varepsilon m(x;\theta) \end{aligned}$$

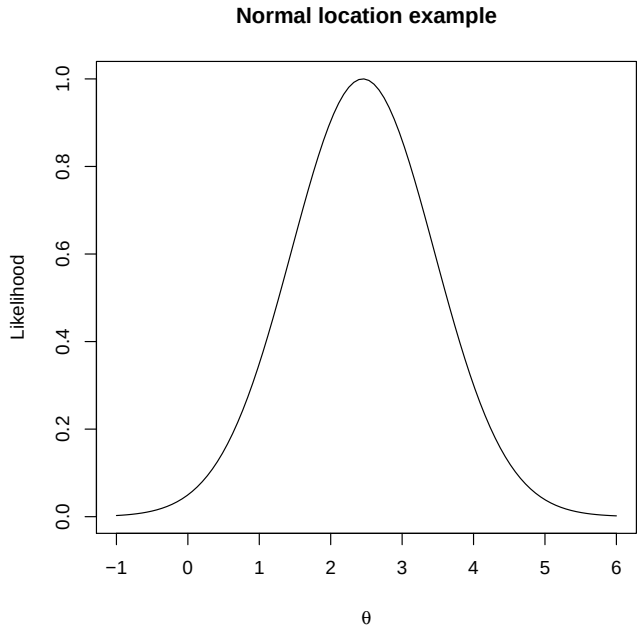
where $m(x;\theta)$ is the probability density function of the model.

For the pourpose of comparing θ within the model $m(x;\theta)$ the likelihood is only meaningful up to an arbitrary constant, so we can ignore ε. Hence, when the outcomes are measuered with good precision we can based the likelihood only on the probability density function $m(x;\theta)$.

Example 41. Suppose x is a sample from a Normal location model $N(\theta, 1)$; the likelihood of θ is

$$L(\theta) = \phi(x;\theta,1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right]$$

where $\phi(x;\theta,1)$ is the density function of the normal location model with scale equal to 1.

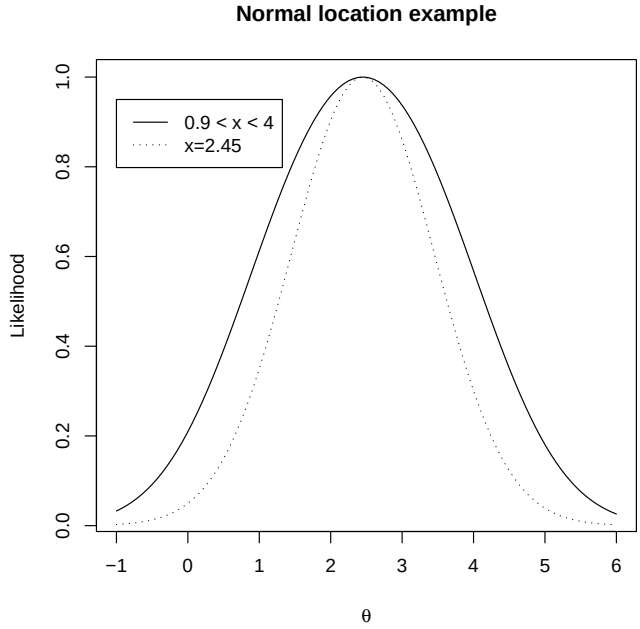


The Likelihood function of the normal location model based on $x = 2.45$.

Example 42. Suppose we have an experiment based on Normal location model $N(\theta, 1)$ and it is known only that $0.9 < x < 4$; the likelihood of θ is

$$L(\theta) = P(0.9 < X < 4; \theta) = \Phi(4 - \theta) - \Phi(0.9 - \theta) \ ,$$

where $\Phi(x)$ is the distribution function of the standard normal.



The Likelihood function of the normal location model based on $0.9 < x < 4$ and $x = 2.45$.

Example 43. Suppose $x_1, \cdots, x_n$ are an identically and independently distributed (iid) sample from $N(\theta, 1)$, an only the maximum $x_{(n)}$ is reported and n is known, while the others are missing. The distribution function of $X_{(n)}$ is

$$\begin{aligned} F_{X_{(n)}}(t) &= \mathrm{P}(X_{(n)} \leq t) \\ &= \mathrm{P}(X_i \leq t, \text{for each i, } i = 1, \cdots, n) \\ &= \mathrm{P}(X_1 \leq t) \cdots \mathrm{P}(X_n \leq t) \\ &= \Phi(t - \theta)^n \end{aligned}$$

taking derivative we have that the density function of $X_{(n)}$ is

$$m(x_{(n)}; \theta) = n \left[\Phi(x_{(n)} - \theta)\right]^{n-1} \phi(x_{(n)} - \theta)$$

so that the likelihood of θ is $L(\theta) = m(x_{(n)}; \theta)$.

Exercise 1. *Suppose $x_1, \cdots, x_n$ are an identically and independently distributed (iid) sample from $Exp(\theta)$, an only the minimum $x_{(1)}$ is reported and n is knwon. Derived the likelihood function for θ.*

2.2 Combining Likelihoods

Combining Likelihoods

Given two independent observations x_1 and x_2 from the probabilistic models $m_1(x_1;\theta)$ and $m_2(x_2;\theta)$ that share a common parameter θ, then the likelihood from the combined data is

$$\begin{aligned} L(\theta) &= m_1(x_1;\theta)\ m_2(x_2;\theta) \\ &= L_1(\theta)\ L_2(\theta) \end{aligned}$$

where $L_1(\theta)$ and $L_2(\theta)$ are the likelihoods from the individual observations.

Applying the previous results, let $x_1, \cdots, x_n$, an iid sample of size n from a model $m(x;\theta)$ we have a likelihood function as

$$L(\theta) = \prod_{i=1}^{n} m(x_i;\theta)$$

Often a log scale leads a simpler calculation so that we can define a **Log–Likelihood** function as

$$\begin{aligned} \ell(\theta) &= \log_e L(\theta) = \log_e \prod_{i=1}^{n} m(x_i;\theta) \\ &= \sum_{i=1}^{n} \ell(x_i;\theta) = \sum_{i=1}^{n} \log_e m(x_i;\theta)\ . \end{aligned}$$

In log scale a combining log–likelihood is simple the sum of the single Log-Likelihoods (We will drop the subindex e from the log from now on).

Example 44. Let $x_1, \cdots, x_n$ be an iid sample from a Normal location model $N(\mu, \sigma^2)$ with known scale σ^2. The contribution of each single observation x_i to the Likelihood is

$$L_i(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right]\ ,$$

so that the combining Likelihood is

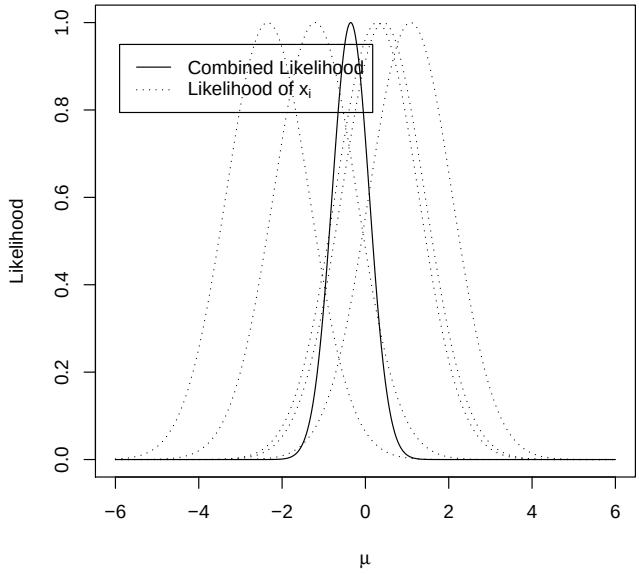
$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right] \end{aligned}$$

Example 45. Using the previous setting we have the following log–likelihood function

$$\ell_i(\theta) = -\frac{1}{2}\log\left[2\pi\sigma^2\right] - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} \ ,$$

so that the combining log–likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n}\log m(x_i;\theta) \\ &= -\frac{n}{2}\log\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$



The Likelihood function of the normal location model $N(\mu, 1)$ based on each single observations $-1.207, 0.277, 1.084, -2.346, 0.429$ and for the whole sample.

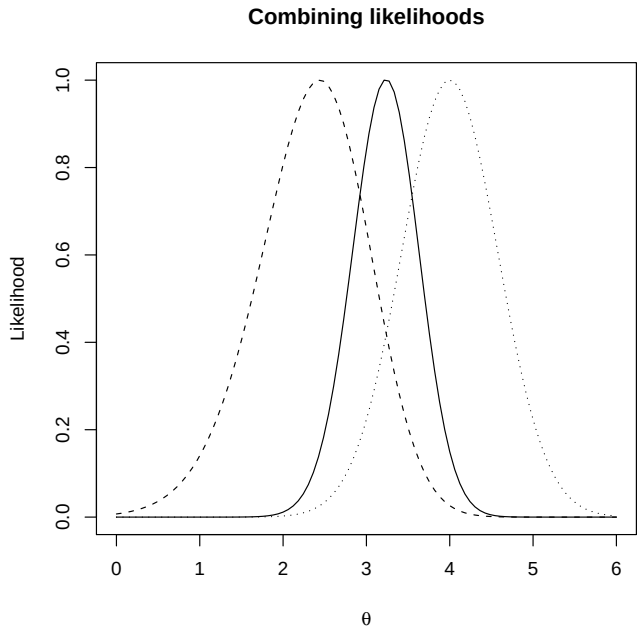
Example 46. Suppose we have two independet samples taken from $N(\theta, 1)$. From the first sample it is reported that the sample size is $n_1 = 5$, and the maximum $x_{(5)} = 3.5$. The second sample has size $n_2 = 3$, and only the sample mean $\bar{y} = 4$ is reported. We have

$$L_1(\theta) = 5\left[\Phi(x_{(5)}-\theta)\right]^4\phi(x_{(5)}-\theta) \ ,$$

and, since $\bar{Y} \sim N(\theta, 1/3)$,

$$L_2(\theta) = \frac{1}{\sqrt{2\pi/3}}\exp\left[-\frac{3}{2}(\bar{y}-\theta)^2\right] \ .$$

The combined Likelihood is $L(\theta) = L_1(\theta)L_2(\theta)$.



Likelihood based on the maximum $x_{(5)} = 3.5$ of the first sample (dashed line), on the sample mean $\bar{y} = 4$ of the second sample (dotted line), and on the combined data (solid line).

2.3 Likelihood ratio

Likelihood ratio

How should we compare the likelihood of different values of a parameter, say $L(\theta_1)$ versus $L(\theta_2)$? Suppose y is a one–to–one transformation of the observed data x; if x is continuous,

$$m_y(y;\theta) = m_x(x(y);\theta) \left| \frac{\partial x}{\partial y} \right| ,$$

so that the likelihood based on the new data y is

$$L(\theta; y) = L(\theta; x) \left| \frac{\partial x}{\partial y} \right| .$$

Obviously x and y should carry the same information of θ, so to compare θ_1 and θ_2 only the likelihood ratio is relevant since it is **invariant** with respect

to the transformation, infact

$$\frac{L(\theta_2; y)}{L(\theta_1; y)} = \frac{L(\theta_2; x)\left|\frac{\partial x}{\partial y}\right|}{L(\theta_1; x)\left|\frac{\partial x}{\partial y}\right|} = \frac{L(\theta_2; x)}{L(\theta_1; x)}$$

2.4 Point estimation by Maximum Likelihood

Maximum Likelihood Estimate

The obvious role of the maximum likelihood estimate (MLE) is to provide a point estimate for a parameter of interest. When log–likelihood function is well approximated by a quadratic function, then we can use this fact to get and study the properties of the MLE.

First we define the **score function** $S(\theta)$ as the first derivative of the log–likelihood

$$S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta)$$

Hence the MLE $\hat{\theta}$ is the solution of the **score equation**

$$S(\theta) = 0 \ .$$

Observed Fisher Information ($I(\theta)$)

At the maximum, the second derivative of the log–likelihood is negative, so we define the curvature at $\hat{\theta}$ as $I(\hat{\theta})$, where

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} \log L(\theta) \ .$$

A large curvature $I(\hat{\theta})$ is associated with a tight or strong peak, intutitively indicating less uncertainty about θ.

Example 47. Let $x_1, \cdots, x_n$ be an iid sample from a normal location model $N(\mu, \sigma^2)$, where σ^2 is known. Ignoring irrelevant constant terms

$$\ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \ ,$$

so

$$S(\mu) = \frac{\partial}{\partial\mu} \ell(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \ .$$

Solving $S(\mu) = 0$ we have $\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i/n$ as the MLE of μ.

The second derivative of the log–likelihood gives the observed Fisher information

$$\begin{aligned} I(\hat{\mu}) &= -\frac{\partial^2}{\partial \mu^2}\ell(\mu)\Big|_{\mu=\hat{\mu}} \\ &= -\frac{\partial}{\partial \mu}S(\mu)\Big|_{\mu=\hat{\mu}} \\ &= -\frac{\partial}{\partial \mu}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\Big|_{\mu=\hat{\mu}} \\ &= \frac{n}{\sigma^2} \end{aligned}$$

From a well known results on the sum of indipendent normal distribution we have

$$\begin{aligned} var(\hat{\mu}) &= var(\bar{X}) \\ &= var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Hence

$$var(\hat{\mu}) = \frac{\sigma^2}{n} = I^{-1}(\hat{\mu})$$

This is an important example, for it is a common theme in statistics that many properties which are exactly true in the normal case are approximately true in regular problems.

Example 48. Based on x from the $Bin(n,\theta)$ the log–likelihood function is (proportional)

$$\ell(\theta) = x\log\theta + (n-x)\log(1-\theta)$$

The score function is

$$\begin{aligned} S(\theta) &= \frac{\partial}{\partial\theta}\ell(\theta) \\ &= \frac{x}{\theta} - \frac{n-x}{1-\theta}\ , \end{aligned}$$

giving the MLE $\hat{\theta} = x/n$.

Exercise 2. *Repeat the same example when we have a sample $x_1, x_2, \cdots, x_m$ of size m.*

The observed Fisher information is

$$\begin{aligned} I(\theta) &= -\frac{\partial^2}{\partial\theta^2}\ell(\theta) \\ &= \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \ , \end{aligned}$$

so at the MLE we have

$$I(\hat{\theta}) = \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

and since the $Bin(n;\theta)$ model is a regular case we have

$$var(\hat{\theta}) \approx \frac{\hat{\theta}(1-\hat{\theta})}{n}$$

Example 49. For some models or in realistic problems often we do not have a closed form solution to the score equation. Suppose an iid sample $x_1, \cdots, x_n$ of size n is taken from a $Ga(\lambda, \omega)$ with density

$$m(x;\lambda,\omega) = \frac{\lambda^\omega x^{\omega-1}\exp(-\lambda\ x)}{\Gamma(\omega)} \ .$$

The log-likelihood function is

$$\ell(\lambda,\omega) = n\ \omega\log\lambda + (\omega-1)\sum_{i=1}^{n}\log x_i - \lambda\sum_{i=1}^{n} x_i - n\ \log\Gamma(\omega)$$

Since we have two parameters we need to take partial derivative from both parameters.

$$S(\lambda,\omega) = \left\{ \begin{array}{lcl} \frac{\partial}{\partial\lambda}\ell(\lambda,\omega) &=& n\frac{\omega}{\lambda} - \sum_{i=1}^{n} x_i = 0 \\ \frac{\partial}{\partial\omega}\ell(\lambda,\omega) &=& n\ \log\lambda + \sum_{i=1}^{n}\log x_i - n\psi(\omega) = 0 \end{array} \right.$$

where $\psi(\omega)$ is the function *digamma*. From the first equation we have $\lambda = n\omega/\sum_{i=1}^{n} x_i = \omega/\bar{x}$ and then substitute in the second equation we have

$$g(\omega) = \sum_{i=1}^{n}\log x_i + n(\log\omega - \log\bar{x}) - n\psi(\omega) = 0$$

The last equation is not linear, a possible way to solve it, is to use the Newton–Raphson method, that is using a first–order Taylor's expansion we have

$$0 = g(\omega_1) \approx g(\omega_0) + (\omega_1 - \omega_0)\ g'(\omega_0)$$

where $g'(\omega) = n/\omega - n\psi'(\omega)$ e $\psi'(\omega)$ is the *trigamma* function.

and hence

$$\omega_1 \approx \omega_0 - \frac{g(\omega_0)}{g'(\omega_0)}$$

As initial values of the algorithm we might use the method of moments estimates as follows.

$$\left\{ \begin{array}{lcl} E(X) = \frac{\omega}{\lambda} & = & \bar{x} \\ E(X^2) = \frac{\omega(1+\omega)}{\lambda^2} & = & m_{(2)}\ . \end{array} \right.$$

where $m_{(2)}$ is the second order empirical moment from the data.

From the first equation we have $\lambda = \omega/\bar{x}$ so that from the second equation

$$m_{(2)} = \bar{x}^2 \left(\frac{1}{\omega} + 1 \right)$$

and finally $\hat{\omega} = \bar{x}^2/(m_{(2)} - \bar{x}^2)$, that is

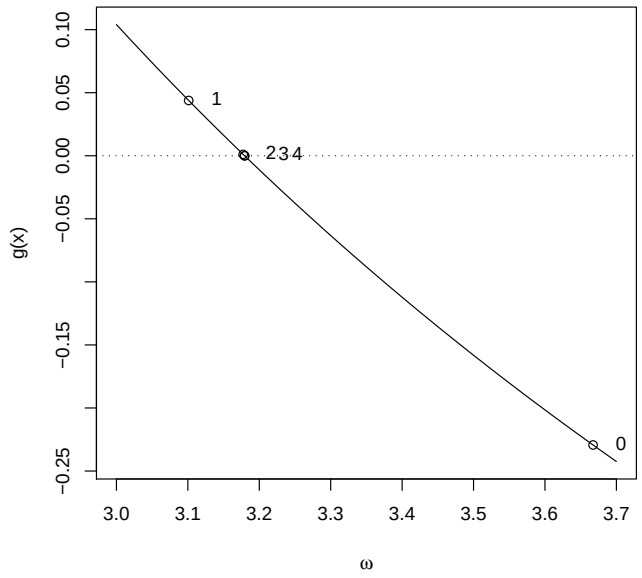
$$\begin{array}{rcl} \hat{\omega} & = & \dfrac{\bar{x}^2}{m_{(2)} - \bar{x}^2} \\ \hat{\lambda} & = & \dfrac{\hat{\omega}}{\bar{x}} = \dfrac{\bar{x}}{m_{(2)} - \bar{x}^2}\ , \end{array}$$

Example 50. Let $x = (0.319, 1.007, 1.033, 0.493, 1.121, 0.558, 1.432, 0.538, 0.601, 0.166)$ an iid sample of size $n = 10$ from a $Ga(\lambda, \omega)$ model. Following the method of moments we have

$$\begin{array}{rcl} \hat{\lambda}_M & = & 5.046 \\ \hat{\omega}_M & = & 3.667 \end{array}$$

and using them in the Newton-Raphson method we have the following iterations

	Iter.	ω	λ	$g(\omega)$
1	0	3.667	5.046	−0.229
2	1	3.101	4.267	0.044
3	2	3.177	4.372	0.001
4	3	3.179	4.375	0.000
5	4	3.179	4.375	0.000



The $g(\omega)$ function and iteration of the Newton–Raphson algorithm starting from the method of moments value.

Approximation of the Log–Likelihood

Using a second–order Taylor expansion of the log–likelihood around $\hat{\theta}$

$$\ell(\theta) \approx \ell(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

and since $S(\hat{\theta}) = 0$ we get

$$\log \frac{L(\theta)}{L(\hat{\theta})} = \ell(\theta) - \ell(\hat{\theta}) \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2 \tag{2}$$

providing a quadratic approxtimation of the normalized log–likelihood around $\hat{\theta}$.

When the normal location with known scale is used the quadratic ap-

proxiamtion is exact, that is

$$\log \frac{L(\theta)}{L(\hat{\theta})} = \ell(\theta) - \ell(\hat{\theta}) = -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$

and since, in this situation, the Maximum Likelihood estimator $\hat{\theta}$ is normally distributed, a good quadratic approximation of the log–likelihood corresponds to a normal approximation of $\hat{\theta}$. So that, a reasonably regular likelihood means $\hat{\theta}$ approximatelly normal.

We can check the regularity of the log-likelihood by plotting the true log–likelihood and the approximation together in a range of the normalized log–likelihood between -4 and 0.

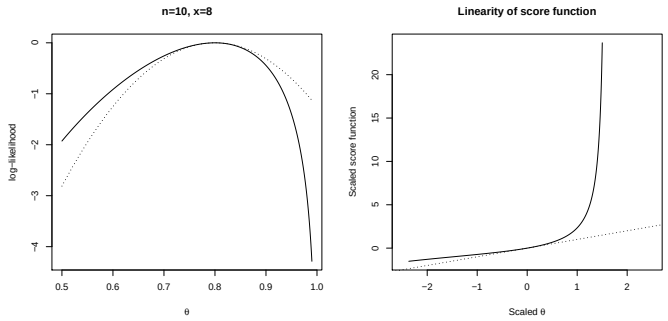
Alternatively, we can study the approximation using the score function. We can take derivative of the quadratic approximation (2) to get

$$\begin{aligned} \frac{\partial}{\partial \theta}\left[\ell(\theta) - \ell(\hat{\theta})\right] &\approx -\frac{\partial}{\partial \theta}\left[\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2\right] \\ S(\theta) &\approx -I(\hat{\theta})(\theta - \hat{\theta}) \end{aligned}$$

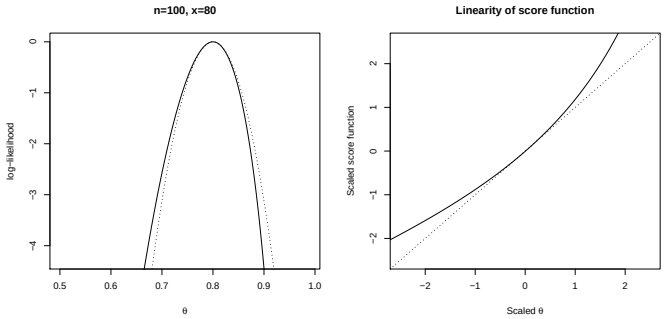
or

$$-I^{-1/2}(\hat{\theta}) S(\theta) \approx I^{1/2}(\hat{\theta})(\theta - \hat{\theta})$$

Example 51. Let $x_1, \cdots, x_n$ a sample of size 10 from a Bernoulli $Bin(1, \theta)$ where we have $\sum_{i=1}^{n} x_i = 8$ successes. Here $\hat{\theta} = 0.8$ and $I(\hat{\theta}) = n/(\hat{\theta}(1 - \hat{\theta})) = 62.5$. If $n = 100$ and $\sum_{i=1}^{n} x_i = 80$ successes then still $\hat{\theta} = 0.8$ but $I(\hat{\theta}) = n/(\hat{\theta}(1 - \hat{\theta})) = 625$. The quadratic approximation is poor in the first case while for the bigger sample size is satisfactory as shown by the next figures.



Quadratic approximation of the log–likelihood function. Left; the true log–likelihood (solid) and the approximaiton (dotted) for the Bernoulli parameter θ (sample size $n = 10$). Right: Linearity check of the score function, showing a poor approximation.



Quadratic approximation of the log–likelihood function. Left; the true log–likelihood (solid) and the approximaiton (dotted) for the Bernoulli parameter θ (sample size $n = 100$). Right: Linearity check of the score function, showing a satisfactory approximation.

In the cases where the quadratic approximaiton is good we might report only the MLE $\hat{\theta}$ and the curvature $I(\hat{\theta})$ because they well summarize the log-likelihood, ate least in a neighborhood of $\hat{\theta}$. In these cases we can use $I(\hat{\theta})$ also for approximating the variance of the MLE as $var(\hat{\theta}) \approx I^{-1}(\hat{\theta})$.

If the likelihood function is not regular, then the curvature of the log–likelihood at the MLE or the variance is not meaningful.

2.5 Interval estimation by Likelihood–based intervals

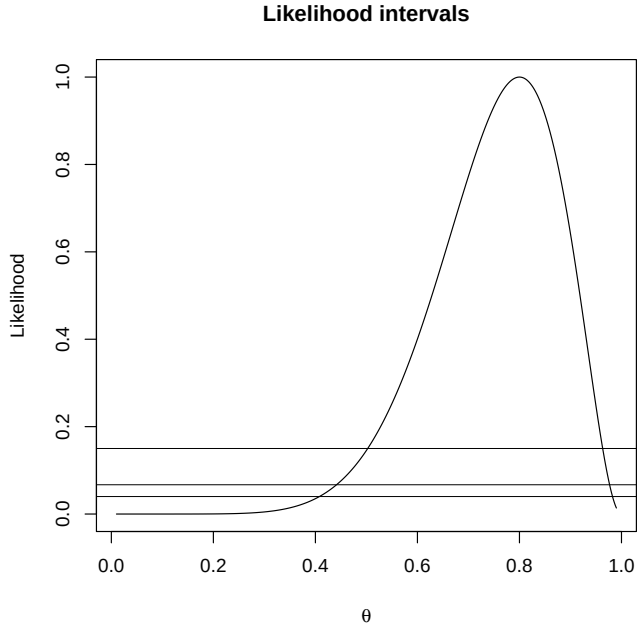
Pure Likelihood–based intervals

Definition 52. A likelihood interval is defined as a set of parameter values with high enough likelihood

$$\left\{ \theta : \frac{L(\theta)}{L(\hat{\theta})} > c \right\} ,$$

for some cutoff point c, where $L(\theta)/L(\hat{\theta})$ is the normalized likelihood.

Fisher suggested that parameter values with less than 1/15 or 6.7% normalized likelihood "are obviously open to grave suspicion". This prescription only works for scalar parameters; in general there is a calibration issue we have to deal with.



Likelihood intervals at 4%, 6.7% and 15% cutoff for the binomial parameter θ (sample $x = 8$ from a $Bin(10, \theta)$) are $(0.408, 0.982)$, $(0.441, 0.977)$ and $(0.503, 0.963)$.

Probability–based intervals

Probabiliy–based intervals are intervals derived, as observations, from statements on the form

$$\mathrm{P}(Z_1 \leq \theta \leq Z_2) = 1 - \alpha$$

where Z_1 and Z_2 are random variables and α is a constant in $[0, 1]$, in general small and close to 5%.

The probability–based intervals with respect to the pure likelihood–based intervals have the advantage to be externally validated, in the sense that: a 5% likelihood does not have a strict meaning, in contrast, a 5% probability is always meaningful as a long-term frequency.

Very often likelihood intervals are related to a corresponding probability intervals in exact way as in the case of the location parameter in the normal model or via a good approximation as in the cases of regular log–likelihood.

Example 53 (Normal location). The normalized log–likelihood is

$$\ell(\mu) - \ell(\hat{\mu}) = -\frac{n}{2\sigma^2}(\bar{X} - \mu)^2 \ .$$

but it is a well known result that

$$\bar{X} \sim N(\mu, \sigma^2/n) \tag{3}$$

and hence

$$\frac{n}{\sigma^2}(\bar{X} - \mu)^2 \sim \chi^2_1 \ ,$$

and finally

$$W = -2(\ell(\mu) - \ell(\hat{\mu})) \sim \chi^2_1 \ .$$

which is called Wilk's likelihood ratio statistic.

Notice that formula (3) is the key for constracting optimal probability–based interval.

From the previous result we have

$$\begin{aligned} \mathrm{P}\left(\frac{L(\mu)}{L(\hat{\mu})} > c\right) &= \mathrm{P}\left(W < -2\log c\right) \\ &= \mathrm{P}\left(\chi^2_1 < -2\log c\right) \end{aligned}$$

so, if for some $0 < \alpha < 1$ we choose a cutoff

$$\chi^2_{1,(1-\alpha)} = -2\log c$$

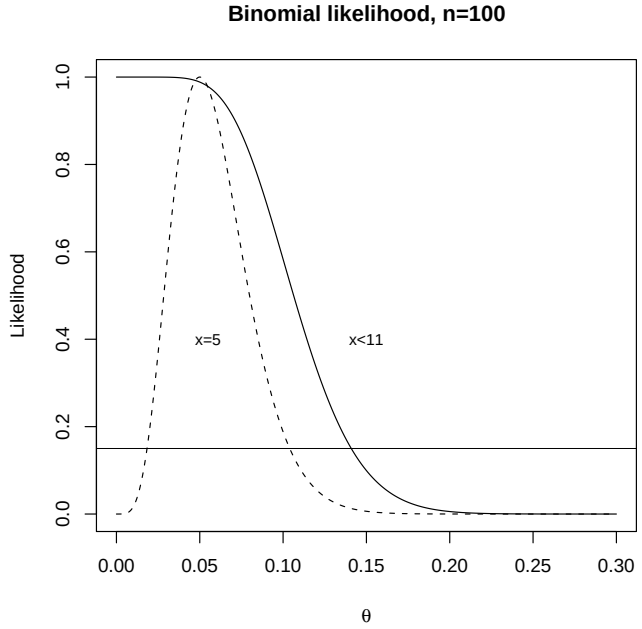
that is

$$c = \exp\left[-\frac{1}{2}\chi^2_{1,(1-\alpha)}\right]$$

where $\chi^2_{1,(1-\alpha)}$ is the $100(1-\alpha)$ percentile of χ^2_1 hence

$$\mathrm{P}\left(\theta : W(\theta) < -2\log c\right) = 1 - \alpha$$

Example 54 (Cont. example 37). When we observed $x \leq 10$ the likelihood is not regular and we do not have any theoretical justification, so that only pure likelihood–based intervals are possible, for instance with a 15% cutoff we have the interval $(0, 0.141)$. If we observed $x = 5$ the information on θ would have been more precise and the likelihood reasonably regular. So we can report an approximated 95% confidence interval $0.019 < \theta < 0.104$.



Likelihood functions from two binomial experiments: $n = 100$ and $x < 11$, and $n = 100$ and $x = 5$. The latter is reasonably regular.

2.6 Standard error and Wald statistic

Standard error and Wald statistic

In regular cases where a quadratic approximation of the log–likelihood works well and $I(\hat{\theta})$ is meaningful, we have

$$-2(\ell(\theta) - \ell(\hat{\theta})) \approx I(\hat{\theta})(\theta - \hat{\theta})^2 \tag{4}$$

so the likelihood interval $\{\theta : L(\theta)/L(\hat{\theta}) > c\}$ is approximately

$$\hat{\theta} \pm \sqrt{-2\log c}\; I(\hat{\theta})^{-1/2} \; .$$

In analogy with the normal model (where everything is exact) $I(\hat{\theta})^{-1/2}$ provides the **standard error** of the MLE $\hat{\theta}$.

The right hand side of (4) and the last remark leads to the **Wald statistic**

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

Further we can use **Wald confidence intervals** based on the quantity

$$\hat{\theta} \pm z_{1-\alpha}\mathrm{se}(\hat{\theta})$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal, while the standard error $\text{se}(\hat{\theta})$ is evaluated as $I(\hat{\theta})^{-1/2}$.

Wald intervals might be called **MLE–based intervals**

Remarks:

- Wald intervals are always symmetric, but likelihood intervals can be asymmetric;
- Computationally the Wald interval is much easier to compute than the likelihood–based interval;
- If the log–likelihood is regular the two intervals will be similar;
- If they are not similar a likelihood–based CI is preferable.

2.7 Rao statistic

Rao (or Score) statistic [2]

Since as we already know

$$S(\theta) \approx -I(\hat{\theta})(\theta - \hat{\theta})$$

and hence

$$-2(\ell(\theta) - \ell(\hat{\theta})) \approx \frac{S(\theta)^2}{I(\hat{\theta})} \approx \frac{S(\theta)^2}{I(\theta)}$$

where the last approximation holds since $I(\theta)$, under regular log–likelihood is a conituous function of θ.

The last two expressions are approximations of the so called **Rao statistic** or Score statistic. As Wilks and Wald statistics could be used for constructing approximated confidence intervals.

Exercise 3. *Construct the three confidence intervals based on Wilks, Wald and Rao statistics for a binomial model with* $n = 100$ *and* $x = 80$.

2.8 Invariance principle

Parametric family

Definition 55. A family of distribution $\mathcal{M}$ is a parametric family if there exists a one–to–one function between the elements in $\mathcal{M}$ and a subset of $\mathbb{R}^p$ space for some fixed value p, that is,

$$\mathcal{M} = \{M(x;\theta); \theta \in \Theta\}$$

[2] see the discussion in section 9.7 pag. 246 of Pawitan [2001]

where $M(x;\theta)$ is the distribution function indexed by θ and $\Theta \subseteq \mathbb{R}^p$ is the parametric space.

Remark:

- when a discrete density or a density is available for all elements of the family then the family can be defined based on the corresponding density, that is,
$$\mathcal{M} = \{m(x;\theta); \theta \in \Theta\}$$

Example 56 (Normal location and scale parametric family).

$$\mathcal{M} = \{\mathrm{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+/0\}$$

where $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (\mathbb{R}^+/0) \subset \mathbb{R}^2$ and $p = 2$.

Parametrization and Re–parametrization

There are many possibilities, in general, of find a function between the distribution functions in a parametric family and a subset in $\mathbb{R}^p$, for instance

Example 57 (Exponential location parametric family).

$$\mathcal{M} = \{\lambda \exp[-\lambda x]; \lambda \in \mathbb{R}^+/0\}$$

and

$$\mathcal{M} = \{\frac{1}{\mu} \exp[-\frac{x}{\mu}]; \mu \in \mathbb{R}^+/0\}$$

where $E(X) = \mu = 1/\lambda$.

Invariance principle

Example 58. In the first binomial example with $n = 10$ and $x = 8$, the likelihood ratio of $\theta_1 = 0.8$ versus $\theta_2 = 0.3$ is

$$\frac{L(\theta_1 = 0.8)}{L(\theta_2 = 0.3)} = \frac{\theta_1^8(1-\theta_1)^2}{\theta_2^8(1-\theta_2)^2} = 208.744$$

i.e. $\theta_1 = 0.8$ is about 200 times more likely than $\theta_2 = 0.3$.

Suppose we are now interested in expressing θ in the logit scale as

$$\psi = \log\left(\frac{\theta}{(1-\theta)}\right)$$

what is our likelihood of $\psi_1 = \log(0.8/0.2) = 1.386$ with respect to $\psi_2 = \log(0.3/0.7) = -0.847$? Since our information is not changed we would like that even the likelihood remains unchanged.

In fact, since $\theta = e^{\psi}/(1+e^{\psi})$

$$\begin{aligned}\frac{L^*(\psi_1 = 1.386)}{L^*(\psi_2 = -0.847)} &= \frac{\left(e^{\psi_1}/(1+e^{\psi_1})\right)^8 \left(1 - e^{\psi_1}/(1+e^{\psi_1})\right)^2}{\left(e^{\psi_2}/(1+e^{\psi_2})\right)^8 \left(1 - e^{\psi_2}/(1+e^{\psi_2})\right)^2} \\ &= \frac{\theta_1^8(1-\theta_1)^2}{\theta_2^8(1-\theta_2)^2} \\ &= \frac{L(\theta_1 = 0.8)}{L(\theta_2 = 0.3)} \ .\end{aligned}$$

That is our information is **Invariant** to the choice of the parametrization.

This property holds for every re–parametrization (for one–to–one re–parametrization but also in other cases) of the likelihood.

Let $\mathcal{M}$ a parametric family based on the parameter θ, and consider the re–parametrization $\psi = g(\theta)$ which is not one–to–one function. In this case

$$L^*(\psi) = \max_{\theta: g(\theta) = \psi} L(\theta) \tag{5}$$

Example 59. Let us consider the binomial model with $n = 10$, $x = 8$ and $\psi = g(\theta) = (\theta - 0.5)^2$. The likelihood $L^*(\psi = 0.04)$ is then evaluated as

$$\begin{aligned}L^*(\psi = 0.04) &= \max_{\theta: (\theta-0.5)^2 = 0.04} L(\theta) \\ &= \max(L(\theta = 0.3), L(\theta = 0.7)) \\ &= \max(0, 0.005) \\ &= 0.005\end{aligned}$$

Invariance property of the MLE

an important property of the invariance of likelihood ratio is the so called invariance property of the MLE.

Theorem 60. *If $\hat{\theta}$ is the MLE of θ in a parametric family $\mathcal{M}$ and $g(\theta)$ is a function of θ, then $g(\hat{\theta})$ is the MLE of $g(\hat{\theta})$.*

Remark:

- The function $g(\theta)$ does not have to be one–to–one, but the definition of the lilkelihood of $g(\theta)$ must follow (5);
- Optimal properties of $\hat{\theta}$ may not be transfer to (and viceversa) $g(\hat{\theta})$ via invariance property;
- The quadratic approximation may be improved by a re–parametrization.

Example 61. In the binomial model with $n = 10$ and $x = 8$ we have for the following re–parametrization $\psi_1 = g_1(\theta) = \theta/(1-\theta)$, $\psi_2 = g_2(\theta) = \log(g_1(\theta))$ and $\psi_3 = g_3(\theta) = \theta^2$ the following MLEs:

- $\hat{\theta} = 8/10 = 0.8$;
- $\hat{\psi}_1 = \hat{g_1(\theta)} = g_1(\hat{\theta}) = \hat{\theta}/(1-\hat{\theta}) = 0.8/0.2 = 4$;
- $\hat{\psi}_2 = \hat{g_2(\theta)} = g_2(\hat{\theta}) = \log g_1(\hat{\theta}) = \log 4 = 1.386$;
- $\hat{\psi}_3 = \hat{g_3(\theta)} = \hat{\theta}^2 = 0.8^2 = 0.64$.

Observed Fisher Information under re–parametrization

We saw how the MLE changes using the invariance property of the likelihood. What happens to the Fisher Information? The following theorem answer to the question

Theorem 62. *Let the paramatric family $\mathcal{M}$ parametrized with a scalar θ and consider the re–parametrization $\psi = g(\theta)$ then*

$$I^*(\psi) = I^*(g(\theta)) = I(\theta) \left| \frac{\partial}{\partial \theta} g(\theta) \right|^{-2}$$

and hence, the standard error of $\hat{\psi}$ is

$$\text{se}(\hat{\psi}) = \text{se}(\hat{\theta}) \left| \frac{\partial}{\partial \theta} g(\theta) \right|_{\theta=\hat{\theta}} .$$

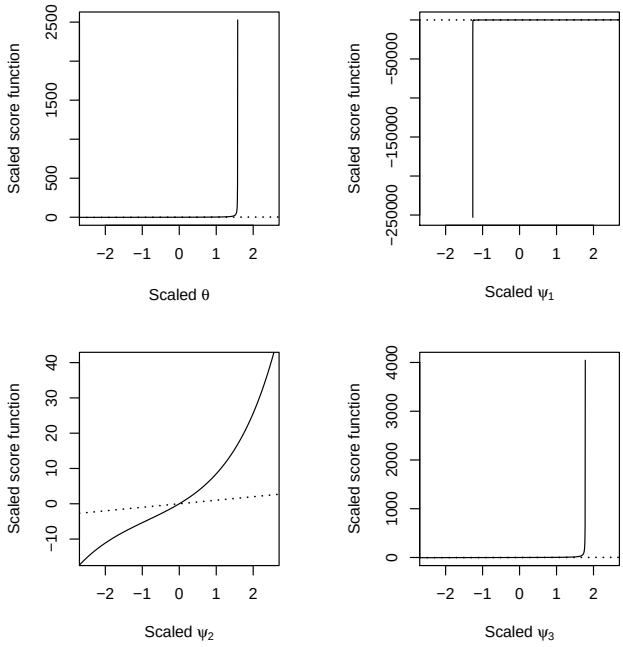
Example 63.

- $I(\hat{\theta}) = n/(\hat{\theta}(1-\hat{\theta})) = 62.5$;
- $I^*(\hat{\psi}_1) = I(\hat{\theta})(1-\hat{\theta})^4 = 0.1$;
- $I^*(\hat{\psi}_2) = I^*(\hat{\theta})(\hat{\psi}_1)^2 = I(\hat{\theta})(1-\hat{\theta})^4 \left(\hat{\theta}/(1-\hat{\theta})\right)^2 = 1.6$;
- $I^*(\hat{\psi}_3) = I^*(\hat{\theta})/(4\hat{\theta}^2) = 24.414$.

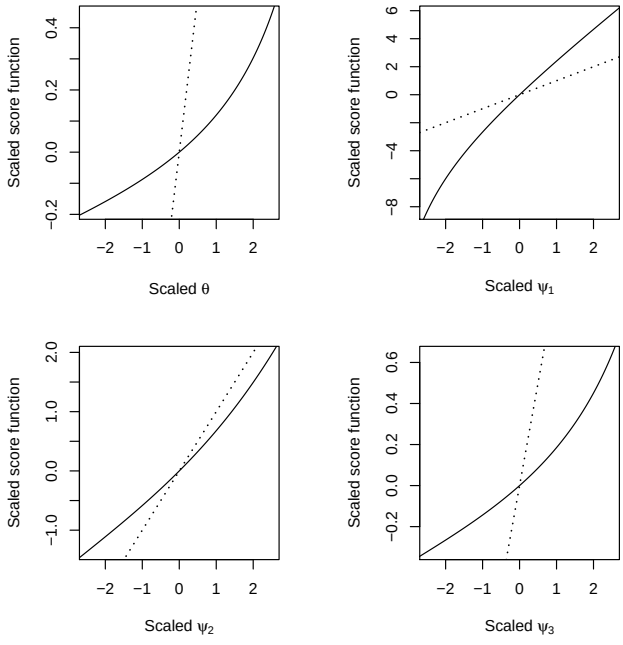
and hence (if the quadratic approximation is good)

- $\text{se}(\hat{\theta}) = I(\hat{\theta})^{-1/2} = 0.126$;
- $\text{se}(\hat{\psi}_1) = I^*(\hat{\psi}_1)^{-1/2} = 3.162$;
- $\text{se}(\hat{\psi}_2) = I^*(\hat{\psi}_2)^{-1/2} = 0.791$;

- $\text{se}(\hat{\psi}_3) = I^*(\hat{\psi}_3)^{-1/2} = 0.202$;



Linearity check for different parametrizations, $n = 10$ and $x = 8$.



Linearity check for different parametrizations, $n = 100$ and $x = 80$.

Improving the quadratic approximation

As we saw from the previous figures, the quadratic approximation can be different depending on the parametrization we are working with. Hence we can have a better Confidence Interval based on Wald statistic by using a parametrization with good quadratic approximation.

Example 64. Form the previous figures we saw that the parametrization in ψ_2 has a better quadratic approximation then that in θ. So we first construct a 95% CI based on ψ_2 and then we inverted to the parametrization in θ.

$$\begin{aligned}(a(\hat{\psi}_2), b(\hat{\psi}_2)) &= \hat{\psi}_2 \pm z_{0.975}\text{se}(\hat{\psi}_2) \\ &= 1.386 \pm 1.96 \times 0.791 = (-0.163, 2.936)\end{aligned}$$

then a 95% Wald CI for θ is

$$\left(\frac{\exp[a(\hat{\psi}_2)]}{1+\exp[a(\hat{\psi}_2)]}, \frac{\exp[b(\hat{\psi}_2)]}{1+\exp[b(\hat{\psi}_2)]}\right) = (0.459, 0.95)$$

while the interval directly obtain using the θ parametrization is

$$\begin{aligned}(a(\hat{\theta}), b(\hat{\theta})) &= \hat{\theta} \pm z_{0.975}\text{se}(\hat{\theta}) \\ &= 0.8 \pm 1.96 \times 0.126 = (0.552, 1.048)\end{aligned}$$

For this case we can construct the exact 95% CI based on the exact distribution of $\hat{\theta}$ which is

$$(0.555, 0.975)$$

Recall that the likelihood–based interval is (0.503, 0.963)

Likelihood–based CI and Wald CI

We do not know how to transform a parameter to get a more regular likelihood and hence a better quadratic approximation. This difficulty is automatically overcome by the likelihood–based intervals since

- The Wald interval is correct only if

$$\frac{\hat{\theta}-\theta}{\text{se}(\hat{\theta})} \sim N(0,1)$$

- In contrast, because of the **invariance** property of the likelihood ratio, the likelihood–based interval is correct as long as *there exists a one–to–one transformaiton $g(\cdot)$, which we do not need to know*, so that

$$\frac{g(\hat{\theta})-g(\theta)}{\text{se}(g(\hat{\theta}))} \sim N(0,1)$$

3 Basic models and simple applications

In this section we use the idea presented in the previous lecture in different models.

3.1 Binomial and Bernoulli models

Binomial and Bernoulli models

The Bernoulli model ($\text{Bin}(1,\theta) = \theta(1-\theta)$) is useful for experiments with dichotomous outcomes. Each experimental unit is thought of as a trial; often assumed

- to be independent each others,
- each with the same probability θ for a successful outcome.

Suppose we observe $(x_1, \cdots, x_n)$ which are realization of n Bernoulli experiment with $\text{P}(X_i = 1) = \theta$ and $\text{P}(X_i = 0) = 1-\theta$. The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

Remark:

Since $X = X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n,\theta)$ we have that observing $x = \sum_{i=1}^n x_i$ or $(x_1, \cdots, x_n)$ carries (for the likelihood) the same information in fact

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

which is proportional (and hence equivalent) to the previous likelihood based on the n Bernoulli trials.

Discrete data are usually presented in grouped form. For example, suppose $x_1, \cdots, x_N$ is an iid sample from $\text{Bin}(n,\theta)$ with n known. We first summarize the data as in the following table

k	0	1	$\cdots$	n
n_k	n_0	n_1	$\cdots$	n_n

where n_k is the number of x_i's equal to k, so $\sum_{k=0}^n n_k = N$. We can now think of the data $(n_0, n_1, \cdots, n_n)$ as having a multinomial distribution with probabilities $(p_0, p_1, \cdots, p_n)$ given by the binomial probabilities

$$p_k = \binom{n}{k}\theta^k(1-\theta)^{n-k} \ .$$

The log–likelihood is given by

$$\begin{aligned}
\ell(\theta) &= \sum_{k=0}^{n} n_k \log p_k \\
&= \sum_{k=0}^{n} n_k \log \left[\binom{n}{k} \theta^k (1-\theta)^{n-k} \right] \\
&= \sum_{k=0}^{n} n_k \log \binom{n}{k} + \sum_{k=0}^{n} k n_k \log \theta + \sum_{k=0}^{n} (n-k) n_k \log(1-\theta)
\end{aligned}$$

$$\begin{aligned}
S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) &= \frac{\sum_{k=0}^{n} k n_k}{\theta} - \frac{\sum_{k=0}^{n} (n-k) n_k}{1-\theta} \\
&= \frac{(1-\theta) \sum_{k=0}^{n} k n_k - \theta \sum_{k=0}^{n} (n-k) n_k}{\theta(1-\theta)} \\
&= \frac{\sum_{k=0}^{n} k n_k - \theta n \sum_{k=0}^{n} n_k}{\theta(1-\theta)}
\end{aligned}$$

and solving the score equation $S(\theta) = 0$ we have

$$\hat{\theta} = \frac{\sum_{k=0}^{n} k n_k}{n \sum_{k=0}^{n} n_k} = \frac{\sum_{k=0}^{n} k n_k}{nN}$$

To ensure that the stationary point is a maximum and in order to evaluate the standard error of the MLE $\hat{\theta}$ we calculate the second derivative of the log–likelihood at $\hat{\theta}$

$$\begin{aligned}
\frac{\partial^2}{\partial \theta^2} \ell(\theta) &= \frac{\partial}{\partial \theta} S(\theta) \\
&= \frac{-n \sum_{k=0}^{n} n_k \theta(1-\theta) - \left(\sum_{k=0}^{n} k n_k - \theta n \sum_{k=0}^{n} n_k \right) (1-2\theta)}{\theta^2 (1-\theta)^2}
\end{aligned}$$

which is at $\hat{\theta}$ equal to

$$\begin{aligned}
\left. \frac{\partial^2}{\partial \theta^2} \ell(\theta) \right|_{\theta=\hat{\theta}} &= \frac{-nN\hat{\theta}(1-\hat{\theta}) - (N\hat{\theta} - N\hat{\theta})(1-2\hat{\theta})}{\hat{\theta}^2 (1-\hat{\theta})^2} \\
&= -\frac{nN\hat{\theta}(1-\hat{\theta})}{\hat{\theta}^2 (1-\hat{\theta})^2} = -\frac{nN}{\hat{\theta}(1-\hat{\theta})}
\end{aligned}$$

which is always negative, hence $\hat{\theta}$ is the MLE.

The observed Fisher Information is

$$I(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta)\Big|_{\theta=\hat{\theta}} = \frac{nN}{\hat{\theta}(1-\hat{\theta})}$$

so that an estimate of the standard error of $\hat{\theta}$ is

$$\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{nN}}$$

Example 65. An industry producing steel plates has two factories. In each factory there are 10 continuous flow production plants. While in the first factory (A) the plants come from the same branch and are quite similar, in the second factory (B) the plants come from different branch and are quite different in terms of obsolescence. The producer has recorded for the last 5 years the number of plants in production every day for the two factories, hereafter is the table summarizing the data

	0	1	2	3	4	5	6	7	8	9	10
A	0	1	2	10	60	145	239	309	239	121	22
B	0	0	0	5	45	131	280	348	241	72	13

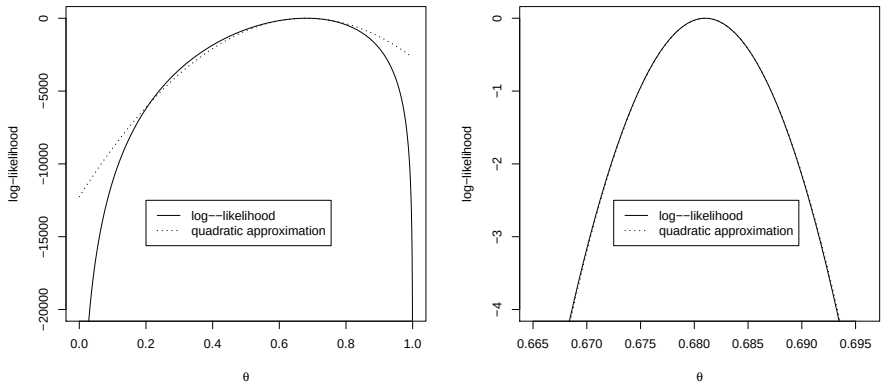
Using a binomial model to describe the problem we have for the first factory (A) the following estimate of the proportion θ,

$$\hat{\theta} = \frac{\sum_{k=0}^{n} k n_k}{nN} = \frac{7818}{10 \times 1148} = 0.681$$

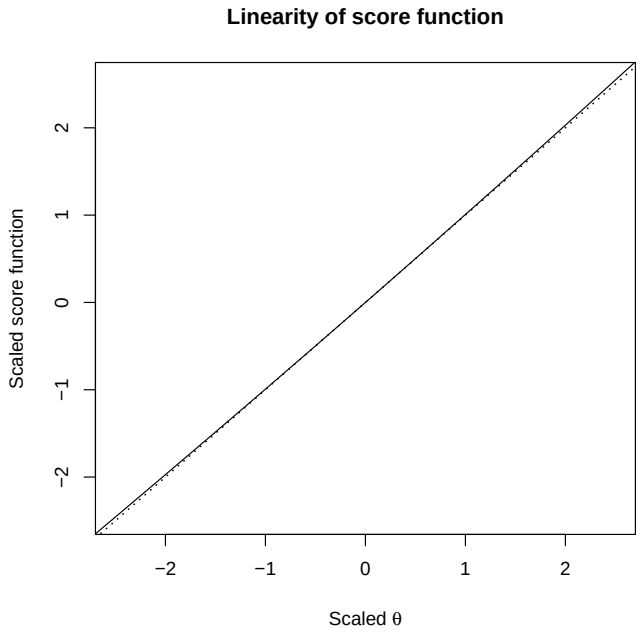
with standard error

$$\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{nN}} = \sqrt{\frac{0.681(1-0.681)}{101148}} = 0.004\ .$$

To check the quadratic approximation we have the two figures in the following slides.



Log–likelihood and quadratic approximation.



Linearity check of the score function.

3.2 Binomial model with under– or over–dispersion

For modelling purposes the binomial model has a weakness in that it specifies a rigid relationship between the mean and the variance in fact, given θ we have

$$\begin{aligned} E(X) &= n\theta \\ var(X) &= n\theta(1-\theta) \ . \end{aligned}$$

We can check if the assumed model (binomial family) is a good one for our data by using the χ^2 Pearson's statistic. First of all we can estimate p_k by

$$\hat{p}_k = \binom{n}{k} \hat{\theta}^k (1-\hat{\theta})^{n-k}$$

and then we can compute the expected frequencies

$$e_k = N\hat{p}_k$$

which leads for the factory A to the following table

	0	1	2	3	4	5	6	7	8	9	10
$\hat{p}_k$	0.00	0.00	0.00	0.01	0.05	0.12	0.22	0.26	0.21	0.10	0.02
e_k	0.01	0.27	2.57	14.62	54.63	139.96	248.99	303.76	243.19	115.37	24.63
n_k	0.00	1.00	2.00	10.00	60.00	145.00	239.00	309.00	239.00	121.00	22.00
r_k	−0.11	1.42	−0.35	−1.21	0.73	0.43	−0.63	0.30	−0.27	0.52	−0.53

We then obtain the goodness–of–fit Pearson's statistic

$$\chi^2 = \sum_{k=0}^{n} \frac{(n_k - e_k)^2}{e_k} = 5.436$$

which is higtly insignificant at 9 degrees of freedom (The p–value is 0.795).

Hence our assumption for factory A seems reasonable.

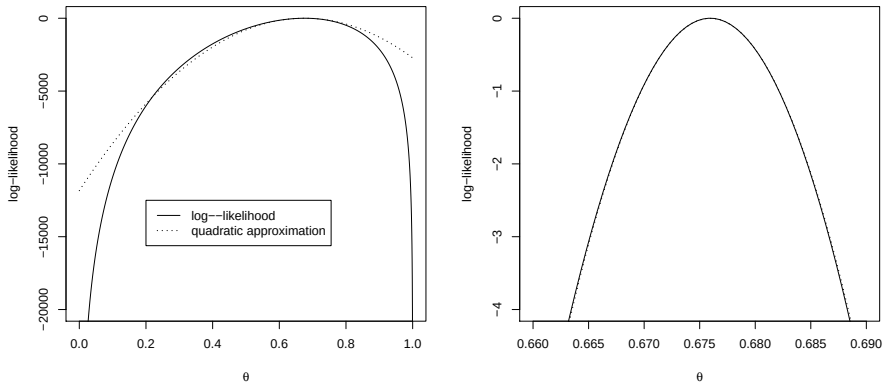
We can redo the same analysis for the data of factory B and we obtain

$$\hat{\theta} = \frac{\sum_{k=0}^{n} k n_k}{nN} = \frac{7818}{10 \times 1135} = 0.676$$

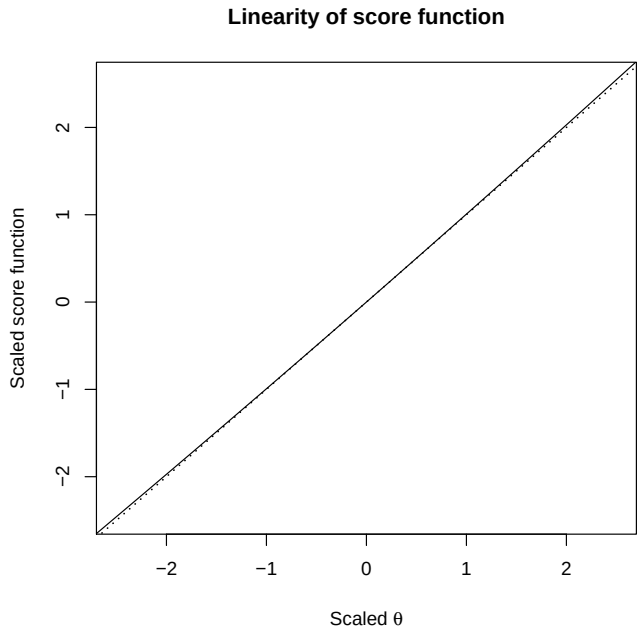
with standard error

$$\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{nN}} = \sqrt{\frac{0.676(1-0.676)}{101135}} = 0.004 \ .$$

To check the quadratic approximation we have the two figures in the following slides.



Log–likelihood and quadratic approximation.



Linearity check of the score function.

We can check if the assumed model (binomial family) is a good one for the data factory B

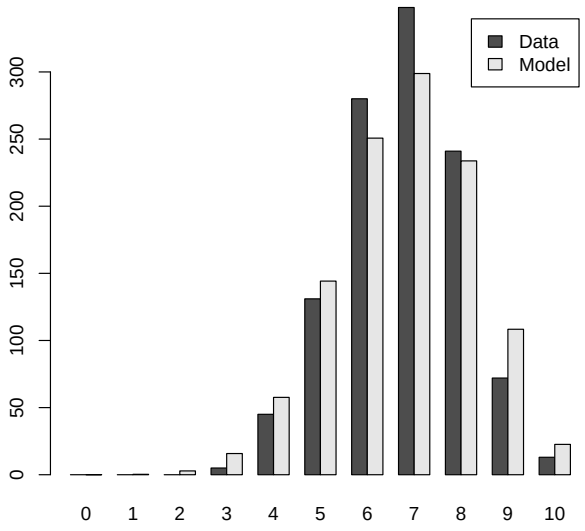
We then obtain the goodness–of–fit Pearson's statistic

$$\chi^2 = \sum_{k=0}^{n} \frac{(n_k - e_k)^2}{e_k} = 42.514$$

which is highly significant at 9 degrees of freedom (The p–value is 0).

Hence our assumption for factory B does not seem reasonable.

	0	1	2	3	4	5	6	7	8	9	10
$\hat{p}_k$	0.00	0.00	0.00	0.01	0.05	0.13	0.22	0.26	0.21	0.10	0.02
e_k	0.01	0.30	2.84	15.78	57.62	144.22	250.70	298.82	233.74	108.35	22.60
n_k	0.00	0.00	0.00	5.00	45.00	131.00	280.00	348.00	241.00	72.00	13.00
r_k	−0.12	−0.55	−1.68	−2.71	−1.66	−1.10	1.85	2.84	0.47	−3.49	−2.02



Frequencies observed from data of factory B and the estimated frequencies $\hat{p}_k$ from the binomial model with $\hat{\theta} = 0.676$.

Underdispersion

We now describe theoretically how a standard binomial model might fail.

Suppose $X_1, \cdots, X_n$ are independent Bernoulli trials with probabilities $p_1, \cdots, p_n$. Let $X = \sum_{i=1}^{n} X_i$; then

$$E(X) = \sum_{i=1}^{n} p_i = n\theta$$

where $\theta = \sum_{i=1}^{n} p_i/n$.

But,

$$\begin{aligned} var(X) &= \sum_{i=1}^{n} var(X_i) \\ &= \sum_{i=1}^{n} p_i(1-p_i) \\ &= \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i^2 \\ &= \sum_{i=1}^{n} p_i - (\sum_{i=1}^{n} p_i)^2/n - \left[\sum_{i=1}^{n} p_i^2 - (\sum_{i=1}^{n} p_i)^2/n\right] \\ &= n\theta - n\theta^2 - n\sigma_p^2 \\ &= n\theta(1-\theta) - n\sigma_p^2 \end{aligned}$$

where we have defined the variance among the p_i's as

$$\sigma_p^2 = \frac{1}{n}\left[\sum_{i=1}^{n} p_i^2 - \frac{(\sum_{i=1}^{n} p_i)^2}{n}\right]$$

So, allowing individual Bernoulli probabilities to vary produces less–than–standard binomial variance.

Example 66 (Cont.). With the data reported from the factory B we are not able to estimate single proportions for each of the 10 plant inside the factory. So, our example stop here. In other cases, when the data carries the necessary informations, we may want to use a more flexible model to take into account the different relation between mean and variance, such as the **Exponential Dispersion model** among others.

Overdispersion

A similar problem is presented when the data are more dispersed then the model. Suppose, X_i's for $i = 1, \cdots, m$, are independent $\text{Bin}(n, p_i)$, and let $X = X_I$ be a random choice from one of these $X_i's$, i.e. the random index $I = i$ has probability $1/m$. This process produces a mixture of binomials

with marginal probability

$$\begin{aligned} \mathrm{P}(X = x) &= E(\mathrm{P}(X_I = x|I)) \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathrm{P}(X_i = x) \\ &= \frac{1}{m}\sum_{i=1}^{m} \binom{n}{x} p_i^x (1-p_i)^{n-x} \end{aligned}$$

which does not simplify further.

The first moment is

$$\begin{aligned} E(X) &= E(E(X_I|I)) \\ &= \frac{1}{m}\sum_{i=1}^{m} E(X_i) \\ &= \frac{n}{m}\sum_{i=1}^{m} p_i = n\theta \end{aligned}$$

where we set $\theta = \sum_{i=1}^{m} p_i/m$.

The variance is

$$\begin{aligned} var(X) &= E(var(X_I|I)) + var(E(X_I|I)) \\ &= \frac{1}{m}\sum_{i=1}^{m} var(X_i) + \frac{1}{m}E(X_i)^2 - \left(\sum_{i=1}^{m} E(X_i)/m\right)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m} np_i(1-p_i) + \frac{1}{m}\sum_{i=1}^{m}(np_i)^2 - (n\theta)^2 \\ &= n\theta(1-\theta) + n(n-1)\sigma_p^2 \end{aligned}$$

where σ_p^2 is the variance among the p_i's as defined previously. So, here we have greater–than–standard binomial variation.

3.3 Negative Binomial model

Negative Binomial model

In the so–called negative or inverse binomial experiment we continue a Bernoulli trial with parameter θ until we obtain x successes, where x is fixed in advance. Let n be the number of trials needed; the likelihood function is

$$\begin{aligned} L(\theta) &= \mathrm{P}(N = n; \theta) \\ &= \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x} \ . \end{aligned}$$

Again here we find the same likelihood function as the one from the binomial experiment, even though the sampling property is quite different.

Example 67. In the production of steel plates, we would like to have 1000 "good" plates. The producer has recorded the number of plates the production plant has to make in order to get the expected number of "good" plates in the last years.

	1	2	3	4	5	6	7	8	9	10
1	125	122	107	116	110	109	124	125	94	108

We are interesting in the probability of success.

The log–likelihood in this case is (let $m = 10$ the number of experiments and $x = 1000$)

$$\ell(\theta) = \sum_{i=1}^{m} x \log\theta + \sum_{i=1}^{m} (n_i - x)\log(1-\theta)$$

from which the score equation is

$$\frac{mx}{\theta} - \frac{\sum_{i=1}^{m}(n_i - x)}{1-\theta}$$

which leads to

$$\hat{\theta} = \frac{mx}{\sum_{i=1}^{m} n_i} = \frac{mx}{mx + \sum_{i=1}^{m} d_i} = 0.898$$

where d_i is the number of failures before to have x successes, i.e. $d_i = n_i - x$.

$$\begin{aligned} I(\hat{\theta}) &= \frac{mx}{\hat{\theta}^2} + \frac{\sum_{i=1}^{m}(n_i - x)}{(1-\hat{\theta})^2} \\ &= \frac{mx(1-\hat{\theta})^2 + \hat{\theta}^2(\sum_{i=1}^{m} n_i - mx)}{\hat{\theta}^2(1-\hat{\theta})^2} \\ &= \frac{\hat{\theta}\sum_{i=1}^{m} n_i - \hat{\theta}^2 \sum_{i=1}^{m} n_i}{\hat{\theta}^2(1-\hat{\theta})^2} \\ &= \frac{\sum_{i=1}^{m} n_i}{\hat{\theta}(1-\hat{\theta})} = 121269.258 \end{aligned}$$

where we have used the fact $mx = \hat{\theta}\sum_{i=1}^{m} n_i$. By the quadratic approximation we have

$$se(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{\sum_{i=1}^{m} n_i}} = 0.003$$

and hence a 95% Wald CI for θ is $(0.892, 0.903)$.

Geometric model

The **Geometric model** is a special case of the Negative binomial model when $x = 1$, i.e.,

$$p(N = n) = \theta(1 - \theta)^(n - 1) \qquad n = 1, \cdots$$

3.4 Poisson model

Poisson model

A discrete random variable X has a Poisson distribution with parameter λ if

$$\mathrm{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \; ; \qquad \lambda > 0 \; , x \geq 0 \; .$$

we indicate it by $X \sim Pois(\lambda)$.

This model is extremely versatile for modelling any data involving counts: the number of accidents on a highway each year, the number of deaths due to a certain illnesss per week, the number of insurance claims in a region each year, ecc.

Remark:

- A Taylor expansion leads to $e^{\lambda} = \sum_{x=0}^{\infty} \lambda^x / x!$.

Poisson approximation of the Binomial

A Poisson model with mean λ is an approximation of the binomial model with large n and a small success probability $\pi = \lambda/n$

$$\begin{aligned}
\mathrm{P}(X = x) &= \binom{n}{x}\pi^x(1 - \pi)^{n-x} \\
&= \frac{n!}{(n-x)!x!}\left(\frac{\lambda}{n}\right)^x\left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{n(n-1)\cdots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&\to \frac{\lambda^x}{x!}e^{-\lambda} \; .
\end{aligned}$$

On observing an iid sample $x_1, \cdots, x_n$ from $Pois(\lambda)$, the likelihood of λ is

$$L(\lambda) = \prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} \propto e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}$$

so that the log–likelihood is

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \ \log \lambda$$

which leads to

$$S(\lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

and the MLE of λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The observed Fisher Information could easily evaluated as

$$\begin{aligned} I(\hat{\lambda}) &= -\frac{\partial}{\partial\lambda}\ell(\lambda)\Big|_{\lambda=\hat{\lambda}} \\ &= \frac{\sum_{i=1}^{n} x_i}{\lambda^2}\Big|_{\lambda=\hat{\lambda}} \\ &= \frac{n\hat{\lambda}}{\hat{\lambda}^2} \\ &= \frac{n}{\lambda} \end{aligned}$$

and hence a standard error for $\hat{\lambda}$ is

$$se(\hat{\lambda}) = \sqrt{\frac{\bar{x}}{n}} \ .$$

Poisson model for grouped data

In the case of grouped data the likelihood based on observing n_k for $k = 0, 1, \cdots$, is

$$L(\lambda) = \prod_{k=0}^{+\infty} p_k^{n_k}$$

where $p_k = \mathrm{P}(X = k; \lambda) = \lambda^k e^{-\lambda}/k!$, and the log–likelihood is

$$\ell(\lambda) = -\lambda \sum_k n_k + \sum_k k n_k \log \lambda \ .$$

The solution of the Score equation leads to the MLE

$$\hat{\lambda} = \frac{\sum_k k n_k}{\sum_k n_k}.$$

The Poisson assumption can be checked if we have grouped data. The idea of a Poisson plot is to plot k against a function of n_k such that it is expected to be a straight line for Poisson data. Since (p_k defined as before):

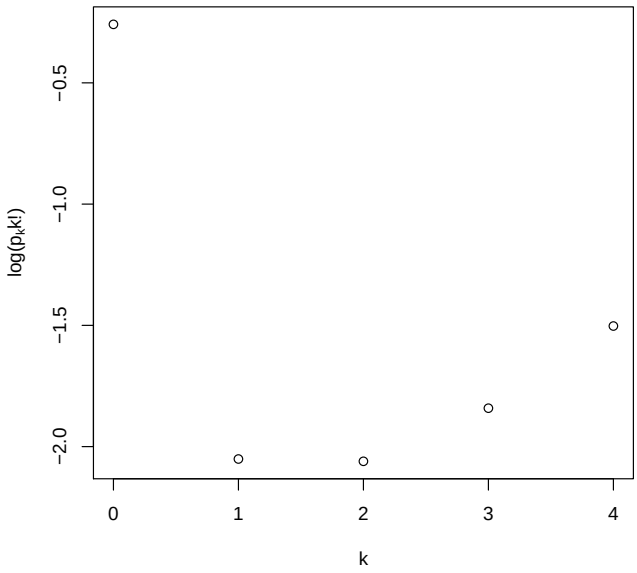
$$\log p_k + \log k! = \lambda + k \log \lambda \ ,$$

indicating that we should plot $\log p_k k!$ versus k, where p_k is estimated by $n_k / \sum_k n_k$.

Example 68 (Truck crash counts). The table below reported the number of truck crashes build using a proprietary, driver-level dataset from JB Hunt, one the largest truckload firms in the U.S. Even though the results are not expected to be fully representative of the population of for-hire truckload drivers. The dataset is rich because it contains human resources, operations, and safety data for 11,540 unscheduled over-the-road dry-van tractor-trailer drivers of JB Hunt, a major U.S. for-hire truckload company, over 26 calendar-months. Some drivers are observed for a single month while others are observed for the entire 26 months. On average, each driver is observed for 9.2 months. We use Table 3. pag. 7 from Rodríguez et al. [200?].

k	0	1	2	3	4+
n_k	8909	1484	735	305	107

Frequency of crashes by driver.



Poisson plot check for the Track crash dataset.

In the previous figure we show the Poisson plot for the Track crash dataset, where we ignore the fact that the last category included 4 and more than 4 crashes.

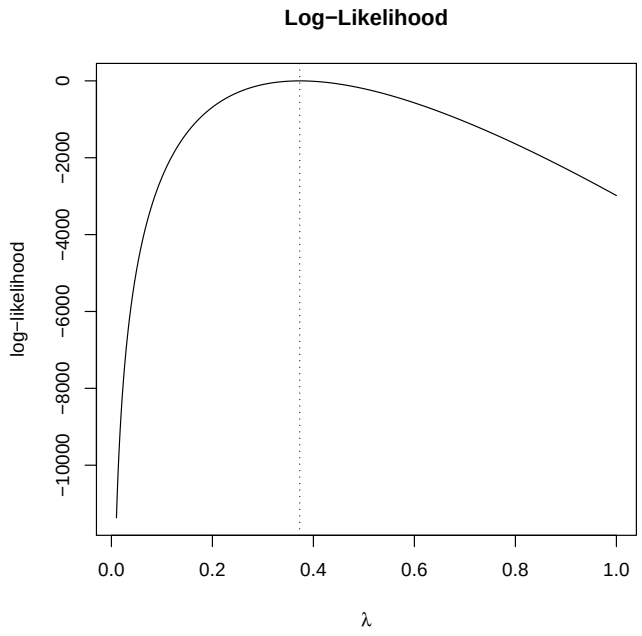
The plot suggests that a Poisson model is not very appropriate, in particular, the frequency of zero crash is much higher then what suggested by a Poisson model.

Hereafter, we still stay with the Poisson model, later we will consider more reliable models.

Likelihood for the Truck crash dataset

$$\begin{aligned} L(\lambda) &= \left(\prod_{k=0}^{3} p_k^{n_k}\right) \mathrm{P}(X > 3; \lambda)^{n_{4+}} \\ &= \prod_{k=0}^{3} \left(\lambda^k e^{-\lambda}/k!\right)^{n_k} (1 - F(3; \lambda))^{n_{4+}} \end{aligned}$$

This likelihood is not very tractable and we will maximized the log–likelihood numerically. We got the MLE $\hat{\lambda} = 0.373$. Notice, that assuming that the last category is represented by $k = 4$ leads to $\hat{\lambda}_{approx} = \sum_{k=0}^{4} k n_k / \sum_{k=0}^{4} n_k = 0.372$, which is a good approximation of the MLE and could be used as an initial value for the numerical algorithm.

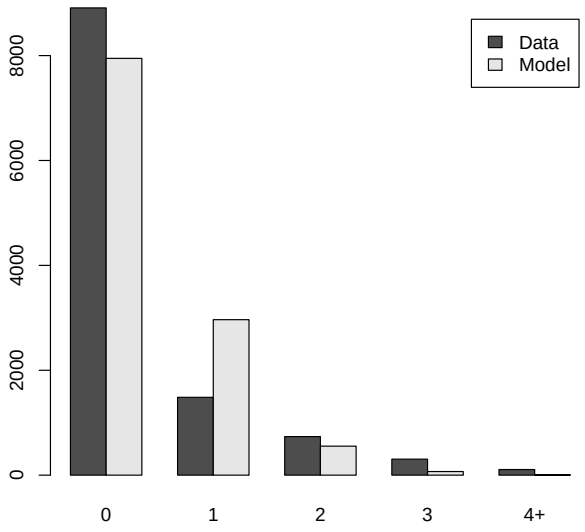


Log–likelihood of the Poisson model for the Track crash dataset.

We can check the adequacy of the model using the Pearson χ^2 statistic

	0	1	2	3	4+
$\hat{p}_k$	0.69	0.26	0.05	0.01	0.00
e_k	7947.07	2964.39	552.88	68.74	6.92
n_k	8909.00	1484.00	735.00	305.00	107.00
r_k	10.79	−27.19	7.75	28.49	38.04

and $\chi^2 = 3174.96$ which strongly suggests that the model is not adequate.



Frequencies observed from the data and estimated $\hat{p}_k$ from the Poisson model with $\hat{\lambda} = 0.373$.

Poisson with overdispersion

Like the binomial model, the Poisson model imposes a strict relationship between the mean and variance that may not be appropriate for the data, in particular we have

$$E(X) = \lambda = var(X) \ .$$

Overdisperion can occur if, conditionally on θ, an outcome X_θ is Poisson with mean θ, and θ is a random variable with mean μ and variance σ^2. For example, individuals vary in their propensity to have accidents, so even if the number of accidents per individual is Poisson, the marginal distribution will show some overdispersion. In this setup, the marginal distribution of X (the mixture of X_θs) has

$$\begin{aligned} E(X) &= E(E(X_\theta|\theta)) = \mu \\ var(X) &= E(var(X_\theta|\theta)) + var(E(X_\theta|\theta)) \\ &= \mu + var(\theta) \\ &= \mu + \sigma^2 \ , \end{aligned}$$

showing an extra variability compared with the Poisson model.

Remark:

- If the propensity of each individual is the same than the r.v. θ is degenerate and we do not have any extra variability.
- we have observations only from X and we do not to which X_θ they belong to.

If θ has a gamma distribution we will get a closed form formula for the marginal probabilities. Specifically, let X_θ be Poisson with mean θ, where θ has density

$$m(\theta; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \theta^{\alpha-1} \lambda^\alpha e^{-\lambda\theta} \ .$$

A random sample of X from the mixture of X_θ for all θ has mean

$$E(X) = \mu = \frac{\alpha}{\lambda}$$

and variance

$$\begin{aligned} var(X) &= \mu + \sigma^2 \\ &= \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2} \ . \end{aligned}$$

For likelihood modelling we need to compute the marginal probability, for $x = 0, 1, \cdots$, but first notice that the join distribution of X and θ is

$$\begin{aligned} \mathrm{P}(X = x, \theta = t) &= \mathrm{P}(X = x|\theta = t)\mathrm{P}(\theta = t) \\ &= \frac{t^x e^{-t}}{x!} \frac{1}{\Gamma(\alpha)} t^{\alpha-1} \lambda^\alpha e^{-\lambda t} \end{aligned}$$

where λ and α are parameters and the marginal distribution of X is

$$\begin{aligned} \mathrm{P}(X = x) &= \int_0^{+\infty} \mathrm{P}(X = x, \theta = t)\, dt \\ &= E_\theta \left(\mathrm{P}(X = x|\theta) \right) \end{aligned}$$

$$\begin{aligned} \mathrm{P}(X = x) &= E_\theta \left(\mathrm{P}(X = x|\theta) \right) \\ &= E_\theta \left(e^{-\theta} \frac{\theta^x}{x!} \right) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha) x!} \int_0^{+\infty} t^x e^{-t} t^{\alpha - 1} e^{-\lambda t}\, dt \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha) x!} \int_0^{+\infty} t^{x + \alpha - 1} e^{-(\lambda + 1)t}\, dt \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha) x!} \frac{\Gamma(x + \alpha)}{(\lambda + 1)^{x + \alpha}} \end{aligned}$$

The last result holds since the definition of the Gamma function is (see Abramowitz and Stegun [1972], Chapter 6: Gamma and Related Functions.)

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$$

and we have to find the following integral (with obvius notation)

$$g(x) = \int_0^{+\infty} t^{x-1} e^{-\lambda t}\, dt$$

that could be solved using the change of variable formula, in our case, let $z = \lambda t$ so that $t = z/\lambda$ and the Jacobian is $t' = 1/\lambda$

$$\begin{aligned} g(x) &= \int_0^{+\infty} \frac{z^{x-1}}{\lambda^{x-1}} e^{-z} \frac{1}{\lambda}\, dz \\ &= \frac{1}{\lambda^x} \int_0^{+\infty} z^{x-1} e^{-z}\, dz = \frac{\Gamma(x)}{\lambda^x} \end{aligned}$$

Poisson-Gamma model as Negative Binomial model

When α is integer we have $\Gamma(x+\alpha) = (x+\alpha-1)!$. In this case

$$\begin{aligned} \mathrm{P}(X=x) &= \frac{\lambda^\alpha}{(\lambda+1)^{x+\alpha}} \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)x!} \\ &= \binom{x+\alpha-1}{\alpha-1} \left(\frac{\lambda}{\lambda+1}\right)^\alpha \left(1-\frac{\lambda}{\lambda+1}\right)^x \end{aligned}$$

since clearly

$$\frac{\Gamma(x+\alpha)}{\Gamma(\alpha)x!} = \binom{x+\alpha-1}{\alpha-1} \quad \text{and} \quad \frac{\lambda^\alpha}{(\lambda+1)^{x+\alpha}} = \left(\frac{\lambda}{\lambda+1}\right)^\alpha \left(1-\frac{\lambda}{\lambda+1}\right)^x$$

which is a Negative Binomial model : the outcome x is the number of failures recorded when we get exactly α successes, and the probability of success is

$$\pi = \frac{\lambda}{\lambda+1} \ .$$

Generalized Negative Binomial model

Note, however, that as a parameter of the gamma distribution α does not have to be an integer. Given the probability formula, and on observing data $x_1, \cdots, x_n$, we can construct the likelihood of (α, λ) or (α, π). When α is not integer we call the previous model the **Generalized Negative Binomial model**.

An estimate of the (mean) accident rate μ is

$$\hat{\mu} = \hat{\alpha}\frac{1-\hat{\pi}}{\hat{\pi}}$$

Using this model for the Track crash dataset we have

$$\ell(\alpha, \pi) = \sum_{k=0}^{3} n_k \log \mathrm{P}(X=k) + n_{4+} \log(1 - \mathrm{P}(X>3))$$

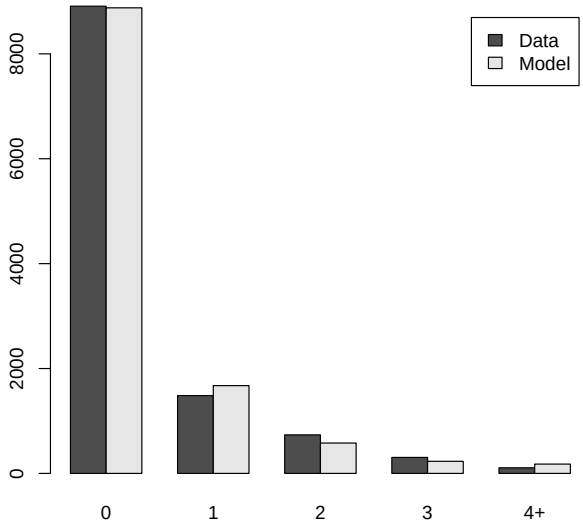
where $\mathrm{P}(X=k)$ is the probability of the Generalized Negative Binomial model.

By numerical optimization we have

$$\hat{\alpha} = 0.374 \qquad \hat{\pi} = 0.496 \qquad \hat{\mu} = 0.38$$

and $\chi^2 = 115.255$ which is still a strong suggestion against the model (The threshold is $\chi^2_{4,0.95} = 9.488$).

	0	1	2	3	4+
$\hat{p}_k$	0.77	0.15	0.05	0.02	0.02
e_k	8877.68	1673.72	579.51	231.12	177.97
n_k	8909.00	1484.00	735.00	305.00	107.00
r_k	0.33	−4.64	6.46	4.86	−5.32



Frequencies observed from the data and estimated $\hat{p}_k$ from the Generalized Negative Binomial model with $(\hat{\alpha}, \hat{\pi}) = (0.374, 0.496)$.

Zero Inflated Poisson model

As the problem of Truck crashes dataset suggests perhaps is not a problem of overdispersion that generate a poor fit of the model to the data. There are cases where, just one category, for the nature of the problem, behave differently from the rest of the categories. In our case, both model (Poisson and Generalized Negative Binomial) put a big effort in order to get a good estimation of the zero category and then they fails for the remains. This is due to the fact that perhaps, the drivers with no accidents are much different from the others.

The **Zero Inflated Poisson model** (ZIP) is a generalization of the Poisson model where the count for the category "no crashes" is modelled as a mixture of an indicator function and a Poisson. In formula, let $0 \leq p \leq 1$,

we have

$$\mathrm{P}(X=k)=\begin{cases} p+(1-p)\frac{\lambda^0 e^{-\lambda}}{0!} & k=0 \\ (1-p)\frac{\lambda^k e^{-\lambda}}{k!} & k>0 \end{cases}$$

The log–likelihood for this model is easy defined as usual but often a numerical algorithm is needed in order to get the MLE. For the Truck crashes dataset we have

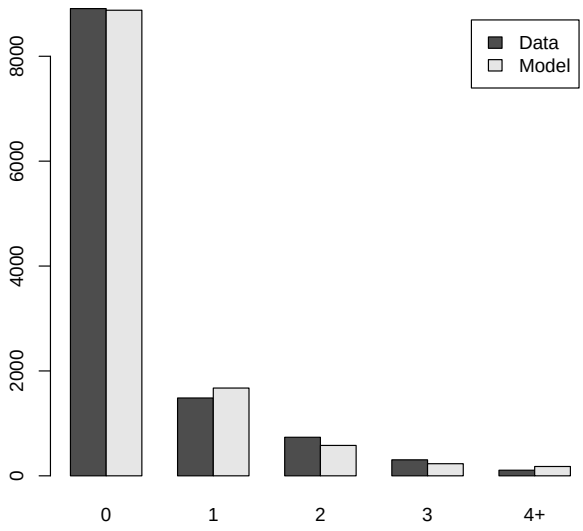
$$\begin{aligned} \ell(\lambda,p) &= n_0 \log(p+(1-p)\frac{\lambda^0 e^{-\lambda}}{0!}) \\ &+ \sum_{i=1}^{3} n_k \log((1-p)\frac{\lambda^k e^{-\lambda}}{k!}) \\ &+ n_{4+} \log\left((1-p)\left(1-\sum_{k=0}^{3}\frac{\lambda^k e^{-\lambda}}{k!}\right)\right) \end{aligned}$$

By numerical optimization we have

$$\hat{\lambda}=1.092 \qquad \hat{p}=0.657$$

	0	1	2	3	4+
e_k	8909.00	1450.79	792.20	288.39	99.63
n_k	8909.00	1484.00	735.00	305.00	107.00
r_k	0.00	0.87	−2.03	0.98	0.74

and $\chi^2=6.393$ which suggests an adequacy of the model (The threshold is $\chi^2_{3,0.95}=7.815$).



Frequencies observed from the data and estimated from the Zero Inflated Poisson model with $(\hat{\lambda}, \hat{p}) = (1.092, 0.657)$.

Zero Inflated Negative Binomial model

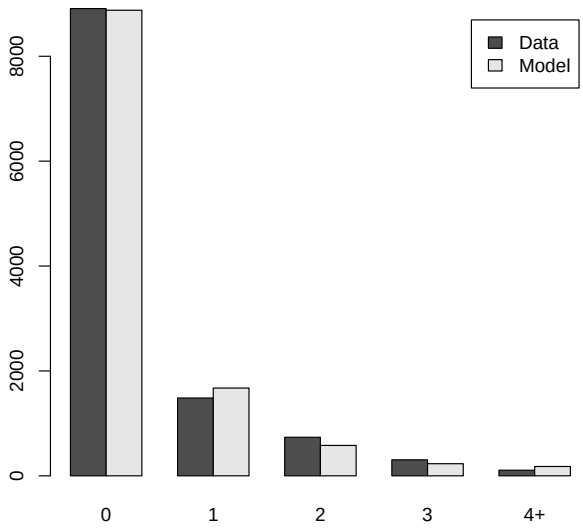
As we extended the Poisson model to the Zero Inflated Poisson model we can extend the Generalize Negative Binomial model to the **Zero Inflated Negative Binomial model** (ZINB) and the log–likelihood is very similar to the previuos one for ZIP.

By numerical optimization we have

$$\hat{\alpha} = 10.496 \qquad \hat{\pi} = 0.911 \qquad \hat{p} = 0.636 \qquad \hat{\mu} = 1.03$$

	0	1	2	3	4+
e_k	8909.01	1476.53	758.49	282.35	113.62
n_k	8909.00	1484.00	735.00	305.00	107.00
r_k	−0.00	0.19	−0.85	1.35	−0.62

and $\chi^2 = 2.967$ which suggests an adequacy of the model (The threshold is $\chi^2_{2,0.95} = 5.991$).



Frequencies observed from the data and estimated from the Zero Inflated Negative Binomial model with $(\hat{\alpha}, \hat{\pi}, \hat{p}) = (10.496, 0.911, 0.636)$.

Final Remark for the Truck crash dataset

While a simple Poisson model is quite inadequate, both the proposed Zero Inflated models are quite well. On the other side, using the lasts we summarize 5 numbers (the 5 counts)by

- an assumption (the form of the model);
- $2-3$ numbers (the parameters).

3.5 Comparing two proportions

Comparing two proportions

Let us consider a discrete bivariate r.v. (X, Y) with two items each. The observations could be summarized by a two entry table as follows

	X	Y	Total
Success	x	y	$t = x + y$
Failure	$m - x$	$n - y$	$N - t$
Total	m	n	$N = m + n$

To compare the two binomial proportions we can use several tools (among them: Pearson χ^2 statistic and Fisher's exact test). Here we will use the

likelihood analysis. Suppose the number of successes X is $Bin(m, \pi_x)$, and, independently, Y is $Bin(n, \pi_y)$. On observing x and y the joint likelihood of (π_x, π_y) is

$$L(\pi_x, \pi_y) = \pi_x^x (1-\pi_x)^x \pi_y^y (1-\pi_y)^y \ .$$

The comparison of two proportions can be expressed in various ways, for example using the difference $\pi_x - \pi_y$, the relative risk π_x/π_y or the log odds–ratio θ defined by

$$\theta = \log \frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}$$

When $\pi_x = \pi_y$ then $\theta = 0$. We have to choose another parameter in order to make the reparametrization one–to–one. A possible one is

$$\eta = \log \frac{\pi_y}{1-\pi_y} \ .$$

The reparametrization leads to

$$\begin{aligned} \pi_x &= \frac{e^{\theta+\eta}}{1+e^{\theta+\eta}} \\ \pi_y &= \frac{e^{\eta}}{1+e^{\eta}} \end{aligned}$$

Therefore, we get the joint likelihood

$$\begin{aligned} L(\theta, \eta) &= \left(\frac{\pi_x}{1-\pi_x}\right)^x (1-\pi_x)^m \left(\frac{\pi_y}{1-\pi_y}\right)^y (1-\pi_y)^n \\ &= \left(\frac{\pi_x/(1-\pi_x)}{\pi_y/(1-\pi_y)}\right)^x \left(\frac{\pi_y}{1-\pi_y}\right)^{x+y} (1-\pi_x)^m (1-\pi_y)^n \\ &= e^{\theta x} e^{\eta(x+y)} (1+e^{\theta+\eta})^{-m} (1+e^{\eta})^{-n} \ . \end{aligned}$$

Recall that $\hat{\pi_x} = x/m$ and $\hat{\pi_y} = y/n$ so that by the invariance property of the likelihood we have

$$\hat{\theta} = \log \frac{\hat{\pi}_x/(1-\hat{\pi}_x)}{\hat{\pi}_y/(1-\hat{\pi}_y)} = \log \frac{x/(m-x)}{y/(n-y)}$$

and in a similar way for $\hat{\eta}$

In order to derive the standard deviation of $\hat{\theta}$ we may use the **Delta method**. So, first we review this method.

Theorem 69 (Delta method)**.** *Let $\hat{\psi}$ be an estimate of ψ based on a sample of size n such that*

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N(0, \sigma^2) \ .$$

Then, for any function $h(\cdot)$ that is differentiable around ψ and $h'(\psi) \neq 0$ we have

$$\sqrt{n}(h(\hat{\psi}) - h(\psi)) \xrightarrow{d} N(0, \sigma^2 |h'(\psi)|^2).$$

We first derive the variance of $\hat{\eta}$ as follows

- Recall that $\hat{\pi}_y - \pi_y \xrightarrow{d} N(0, \pi_y(1-\pi_y)/n$;
- let $\eta = h(\pi_y) = \log \pi_y/(1-\pi_y)$ so that $h'(\pi_y) = (\pi_y(1-\pi_y))^{-1}$;
-

$$\begin{aligned} var(\hat{\eta}) &= \frac{\hat{\pi}_y(1-\hat{\pi}_y)}{n} \frac{1}{(\hat{\pi}_y(1-\hat{\pi}_y))^2} \\ &= \frac{1}{n\hat{\pi}_y(1-\hat{\pi}_y)} \\ &= \frac{1}{y} + \frac{1}{n-y} \end{aligned}$$

Since

$$\hat{\theta} = \log \frac{\hat{\pi}_x}{(1-\hat{\pi}_x)} - \log \frac{\hat{\pi}_y}{(1-\hat{\pi}_y)}$$

and the two terms on the right side are both approximately normal and independent each others then

$$se(\hat{\theta}) = \left(\frac{1}{x} + \frac{1}{m-x} + \frac{1}{y} + \frac{1}{n-y} \right)^{1/2}$$

3.6 Normal model

3.7 Exponential model

Multinomial Distribution

Definition 70. The extension of the binomial distribution to the case of more than two classes. For example, suppose that the probabilities of classes $1, 2, \cdots, m$ are $p_1, p_2, \cdots, p_m$ with $\sum_{j=1}^{m} p_j = 1$. Let X_j denote the number,

in a sample of size n, that are in class j, for $j = 1, 2, \cdots, m$. The random variable $X_1, X_2, \cdots, X_m$ have a multivariate distribution given by

$$\mathrm{P}(X_1 = n_1, X_2 = n_2, \cdots, X_m = n_m) = \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m} \ ,$$

$0 \leq n_1, n_2, \cdots, n_m \leq n$, with $\sum_{j=1}^{m} n_j = n$. The random variable X_j has expectation np_j and variance $np_j(1 - p_j)$ and the covariance of X_j and X_k $(j \neq k)$ is $-np_j p_k$.

References

M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1972.

A. Azzalini. *Statistical Inference Based on the Likelihood*. Chapman & Hall, 1996.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, 2002.

K.L. Chung and F. AitSahlia. *Elementary Probability Theory. With Stochastic Processes and an Introduction to Mathematical Finance*. Springer, 4th edition edition, 2003.

M.D. Cifarelli. *Introduzione al calcolo della probabilità*. McGraw Hill, 1998.

D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman & Hall, 2000.

R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368, 1922.

R.V. Hogg and A.T. Craig. *Introduction to Mathematical Statistics*. 5th edition edition, 1994.

G. Letta. *Probabilità elementare*. Zanichelli, 1993.

A.M. Mood, D.C. Boes, and F.A. Graybill. *Introduction to the Theory of Statistics*. McGraw-Hill Science/Engineering/Math, 3rd edition edition, 1974.

Y. Pawitan. *In All Likelihood*. Oxford Science Publications, 2001.

D.A. Rodríguez, M. Rocha, A.J. Khattak, and M.H. Belzer. The effects of truck driver wages and working conditions on highway safety: A case study. download from the www, 200?

L. Wasserman. *All of Statistics. A concise Course in Statistical Inference.* Springer, 2003.

4 GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The **"Document"**, below, refers to any such manual or work.

Any member of the public is a licensee, and is addressed as **"you"**. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A **"Modified Version"** of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A **"Secondary Section"** is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The **"Invariant Sections"** are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The **"Cover Texts"** are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A **"Transparent"** copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called **"Opaque"**.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include

proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The **"Title Page"** means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section **"Entitled XYZ"** means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as **"Acknowledgements"**, **"Dedications"**, **"Endorsements"**, or **"History"**.) To **"Preserve the Title"** of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in

covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright ©YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.