SOR2002 : Statistical Inference

Jean Peyen
Queen's University Belfast

2023/2024

Caution: This document is a draft and may contain mistakes.

Chapter 1

Stages in a statistical investigation

The aim of a statistical investigation is to acquire data and to analyse them in order to observe the effect of a set of explanatory (or independent) variables on a set of response (or dependent) variables. A statistical investigation can be decomposed into the following stages :

1. **Problem definition and background understanding :** This involves defining a problem or a research question and understanding its context. It is essential to comprehend the background related to the problem and formulate it in statistical terms. It may involve discussing with specialist or performing a literature review. This step sets the direction for the entire investigation.

2. **Data collection :** This stage involves gathering the required data. The data collection method should be aligned with the objectives of the investigation and should be reliable and valid. It may involve ethical considerations.

3. **Preliminary data analysis and processing :** This includes summarising the data using descriptive statistics to gain initial insights. Data processing is also crucial at this stage, which involves dealing with problematic or missing entries. Data may also need to be anonymised or standardised. The goal is to clean the data and develop a preliminary understanding to make reasonable assumptions for further analysis.

4. **Definitive analysis :** This final stage involves formulating assumptions and building a statistical model that adheres to these assumptions. The model is then used to analyse the data and address the initial problem or research question. This stage often involves hypothesis testing, estimating model parameters, and interpreting results to draw conclusions that answer the original problem or question.

In this chapter we briefly outline methods of data collection and aspects of the definitive analysis. Next chapter will focus on the preliminary analysis.

1.1 Data collection

- **Experiment:** One or more variables is/are controlled to observe the effect on the dependent variables. In an ideal experiment we can conclude that any clear difference in response is due to the controlled variables and not to other factors.

- **Observational studies:** Subjects are observed without manipulating any variables. The investigator records data about the subjects, but does not intervene in any way.

- **Sample survey:** Data are extracted from responses from a sample of the population. The sampling scheme should involve some randomisation and be under the control of the investigator. Ideally, characteristics can be estimated accurately from a well-conducted survey. Surveys are essentially observational.

- **Retrospective study:** A response variable has been observed on different individuals and the history of these individuals is examined in order to try and assess which explanatory variables are important in determining the response.

- **Prospective study :** Individuals, chosen by the investigator, have various explanatory variables measured and are then observed through time to see what happens. The investigator chooses the individuals (perhaps by randomisation) and the explanatory variables.

1.2 Definitive analysis

1.2.1 Model formulation

Formulating a statistical model requires to make reasonable assumptions based on

- what is known with near certainty,
- what is safe to assume,
- what is uncertain.

The goal of a model is to be able to evaluate a response $\mathbf{y}$ as a function of explanatory variables $\mathbf{x}$. In practice models often need to incorporate systematic and random components. The random component allows to account for errors that cannot be predicted from the explanatory variables that are available. In general a statistical model is formulated in the following way

$$\underbrace{\mathbf{y}}_{\text{response}} = \underbrace{f(\mathbf{x})}_{\text{systematic part}} + \underbrace{\varepsilon}_{\text{random part}} .$$

The following aspects regarding the random component should be considered

- what assumptions of independence are appropriate,
- is the random part identically distributed across the data,
- if so, what probability distribution it follows.

The independence is contingent upon the data collection approach used. In a carefully designed experiment or sample survey for instance, it is possible to ensure it.

Assuming that the model is unbiased, the random part is centered. Additionally, a small standard deviation of the random part, compared to the response, contributes to more precise estimates.

1.2.2 Fitting

The main procedure is estimation although we may also perform appropriate significance tests. For models involving parameters the estimation stage consists of finding point estimates, their standard errors, and interval estimates of these parameters.

1.2.3 Validation

When a model has been fitted to a set of data, it has to be validated, i.e. the underlying assumptions need to be checked. Typical questions are:

- Is the systematic part of the model satisfactory? If not, how should it be altered, e.g. by transformation of variables, or inclusion of more variables?
- Can the model be simplified, e.g. by removal of variables?
- What can we say about the distribution of the random component? (nature of the distribution, variance...)

Several checking procedures involve looking at the residuals

$$\text{residual} = \text{observation} - \text{fitted value}$$

The residuals should be plotted on a straight line or if there is a large number of residuals, their histogram should be plotted. They should also be plotted against the fitted values and against any other variable of interest. If possible, it is also advisable to plot the residuals in the order in which the corresponding observations were collected to see if there is any trend with time. If the residual plots reveal an unexpected pattern, then the model requires modification.

Large residuals are of great importance and may arise for multiple reasons:

- there is a problem with the observations,
- the wrong model has been fitted,
- a inadequate form of analysis has been performed,
- the distribution of the random components is skewed.

It is also crucial to identify influential observations, which are those that, when removed, result in significant alterations to the fitted model.

1.2.4 Conclusion

Results of the definitive analysis are often presented in the forms of properly constructed tables and graphs. The results should be compared with those from any similar studies. When interpreting the results of the definitive analysis and then communicating our conclusions, we should be aware of the need that they be understood by non-statisticians as well as statisticians.

Chapter 2

Initial Data Analysis (IDA)

2.1 Data structure

The method of analysis that will be used depends on the structure of the data, including factors such as sample size, the grouping of experimental units, and the quantity and nature of the variables measured or observed. A small sample size can render model fitting unreliable, although such samples can be used to set conjectures. Conversely, managing and maintaining quality becomes challenging with very large samples.

Grouping of experimental units typically aims at comparing these groups across one or more variables. It is preferable to limit the number of groups and ensure each group has a sufficient sample size for reliable analysis.

The quantity of variables significantly impacts the analysis. Analyses involving one or two variables are generally more straightforward than those with many variables. In cases of numerous variables, it is crucial to determine their necessity and consider organising them into distinct sets. For complex scenarios with many variables, multivariate methods might be used.

The types of variables measured or observed also have a large influence on the method of analysis. We distinguish the following scales

1. **Nominal scale:** Experimental units are classified into classes, for which the order is not important (e.g. hair color).

2. **Ordinal scale:** As (1) but the classes have an ordering (e.g. patient condition : good, mild, serious, critical, dead).

3. **Interval scale:** It quantifies the differences between two experimental units, where the "zero point" on the interval scale is arbitrary (e.g. temperature 0 Celsius = 32 Fahrenheit).

4. **Ratio scale:** Like (3) but there is a non-arbitrary zero (e.g. mass, energy, count variable).

From type 1 to type 4, the variable types contain more and more information. (1) and (2) are categorical variables, only frequency calculations are meaningful. (3) and (4) are quantitative variables, in addition to frequency calculations, we may perform arithmetical operations (e.g. to compute mean, variance etc.).

A variable can be

- **Discrete:** Variables of type (1) and (2) are necessarily discrete. Variables of type (3) and (4) may be discrete or continuous. An example from type (4) taking discrete values is a count variable.

- **Continuous:** A continuous variable is one which is measured on an effectively continuous scale, taking all possible values in a subset of manifold with a non-empty interior. In this module, continuous values will usually belong to interval(s), however, statisticians occasionally need to deal with more complicated types of continuous data that might have no order or no well-defined mean (e.g. directional data).

2.2 Processing the data

Information is ultimately recorded on paper although the results of some experiments may be stored directly into a computer. Unless the dataset is small (both in number of subjects and variables), the data are analysed using a computer. Processing involves three stages

1. **Coding:** A standard method of coding the information must be decided upon and applied consistently. We need to distinguish the encoding of the output from the encoding of the computer input. The former needs to be readable by a human being. The later needs to be adapted to the computer architecture and the software environment, inconsistency may cause errors (errors of type conversion, overflow etc.)

    - **For quantitative variables:** Units, numerical form (decimal or scientific), precision (number of digits outputed, simple or double precision, short or long integers etc.)

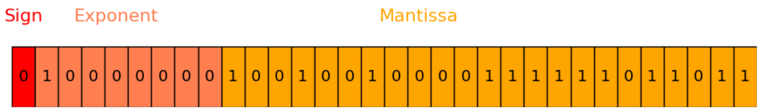


Figure 2.1: A floating point representation of π on 32 bits.

    - **Categorical variable:** The labels should be unequivocal. For ordinal variables they should reflect the ordering.

- **Missing information:** Choose a clear code that doesn't occur in proper entries (e.g. NaN, void).

2. **Input:** The data are entered into the computer and verified if possible.

3. **Screening and editing:** A first screening should be performed before the analysis this allow to

   - detect errors that should be corrected or omitted,
   - locate outliers (observations that are possible even without errors but that are not representative) and assess if they are influential,
   - uncover some basic distributional properties.

2.3 Modifying the data

Having screened the data we may wish to modify certain observations or variables. There are four main types of modification:

- **Adjust extreme observations:** One way is using trimmed mean, that is calculating the mean after discarding the lower and upper ends. We should be careful if the data is not symmetrically distributed.

- **Estimating missing observations:** Replace missing values of a variable by the mean or median of variables that are already present. If this variable is related to the others in some way, a regression model may be used.

- **Transforming variables:**
  - **To deal with the skewness:** Data is skewed when the mean and the median don't agree. It is right-skewed (resp. left-skewed) when then mean is larger (resp. smaller) than the median. Applying a concave function may correct right-skewed data while a convex function may correct left-skewness. Common transformations are log and powers smaller than one (concave), or the exponential and powers larger than one (convex). This is justified by the Jensen inequality, which states that if f is a convex function

    $$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

  - **To deal with the data range.** If the data is ranging in an interval (e.g. a probability ranges in $[0, 1]$), a function such as the logit can be applied to extend its range.

    $$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

- **Combining variables:** Multiple variables may be combined through arithmetic operations to reduce redundancy or to reduce the dimension of the data. It can also improve models performance in some cases.

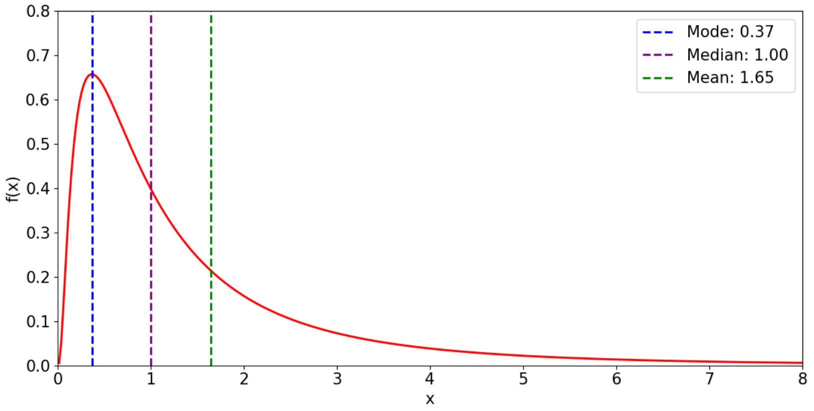


Figure 2.2: The standard log-normal distribution, obtained by exponentiating a Gaussian random variable.

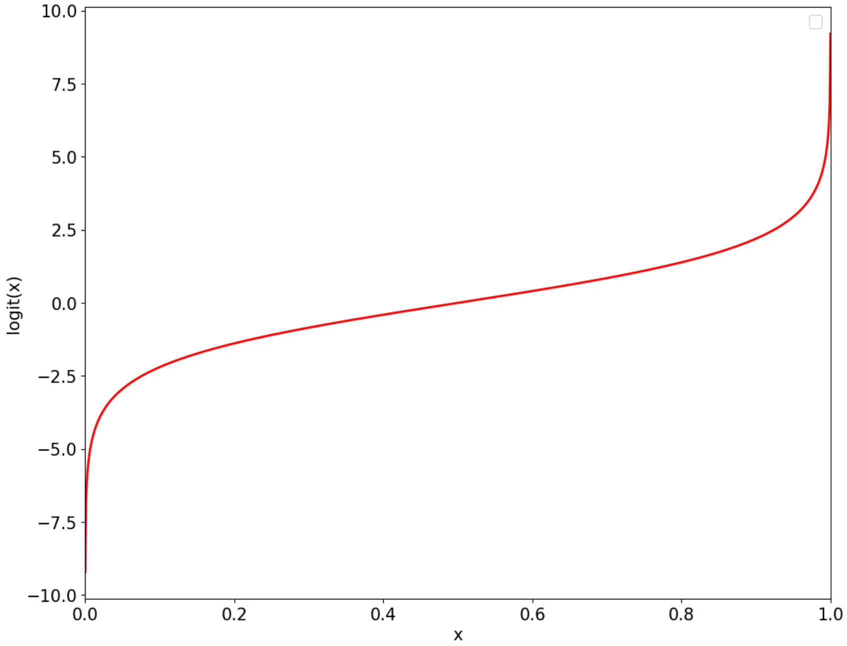


Figure 2.3: The logit function maps $(0, 1)$ to $\mathbb{R}$, thus allowing to use general linear models having a probability as a response variable.

2.4 Preliminary analysis

The preliminary analysis should provide a basic understanding of the distributional properties of the data. The methods than can be applied depend upon the type of the data. The outcome of this analysis should be presented in a suitable way (e.g. tables or graphics).

- **Frequency plot:** (no requirement) This simply consists in plotting the frequency of each class.
- **Location:** Sample mean (requires interval scale)

  $$\bar{x} = \frac{\sum x_i}{n}$$

  median (requires an ordered set)

  $$m = \underset{x}{\operatorname{argmin}}\{\sharp\{i : x_i \leq x\} \geq n/2\}.$$

- **Spread:** Sample variance (requires interval scale)

  $$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}.$$

  It allows to produce confidence intervals for the mean, assuming that the family of distribution of the data allows it. For instance, if the mean is normally distributed we can provide a two-sided 95% confidence interval for the true expectation is

  $$\bar{x} \pm \frac{1.96 \cdot s}{\sqrt{n}}.$$

The mean can be normally distributed either because the data is intrinsically normally distributed or because the dataset is sufficiently large (cf central limit theorem).

More generally, assuming a symmetric distribution, we can build two-sided confidence intervals of level $1-\alpha$ in the following way

$$\bar{x} \pm \frac{Q_{1-\alpha/2} \cdot s}{\sqrt{n}}.$$

With limited knowledge of the distribution of the data, we can still use the Chebyshev inequality

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}.$$

- **Modes:** Some distributions have multiple modes ("*peaks*"). This usually indicates that the distribution is a mixture. Each component of the mixture should be investigated separately as basic statistics such as the mean, the median and the standard deviation do not reflect the nature of the distribution.

- **Correlation:** If data contain multiple variables, some of them may be linearly associated. This can be assessed by measure the correlation which is derived from the variance-covariance matrix. The covariance between x and y is estimated by

$$s_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

and the corresponding correlation coefficient is

$$\rho_{x,y} = \frac{s_{x,y}}{s_x s_y}.$$

If the coefficient is close to zero the variables are non-correlated, if it is close to -1 they are negatively correlated and if it is close to 1 they are positively correlated. As a rule of thumb, an absolute value larger than 0.5 indicates a strong correlation, an absolute value between 0.3 and 0.5 indicates a moderate correlation. In the process of fitting some models, a high correlation between multiple explanatory variables indicates a redundancy that could be problematic. On the contrary, a high correlation between an explanatory variable and a response suggests a linear relationship that could be leveraged.

For categorical variables, we can do little more than calculating the frequency of the classes.

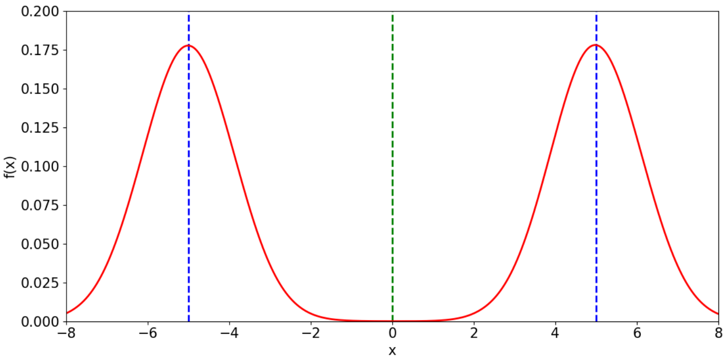


Figure 2.4: The density of a mixture of two Gaussian. The mean and the median (both equal to zero here) are not indicating "typical values".

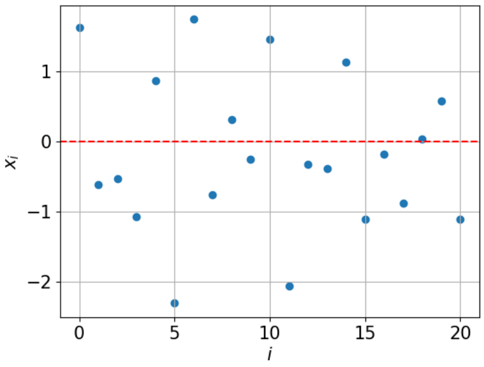
Chapter 3

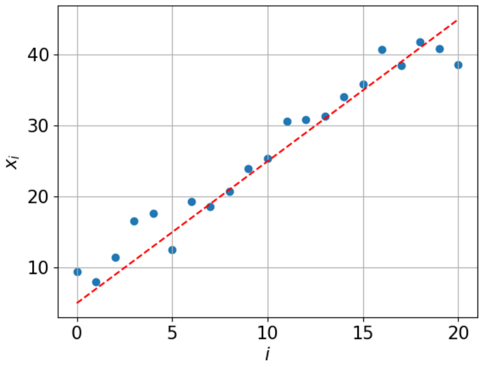
Sample diagnostic

3.1 Independence test

3.1.1 Visual assessment

As a sanity check it can be helpful to plot the data entries (i, x_i) to observe any pattern that would suggest a dependence





3.1.2 Wald–Wolfowitz run test

Binary sequences

In this test we consider a sequence of binary entries (e.g. 0s and 1s). The goal is to check if the entries are independent and identically distributed (i.i.d.). The test consists in counting the number of runs, that is the number of sub-sequences of adjacent equal elements (for instance the sequence 00011011 contains 4 runs). Assuming that we have a sequence of n bits (b_i) containing n_0 zeros and n_1 ones. Assuming that the sequence is i.i.d. all of the bits should have the same probability to be terminating their run (except the last)

$$\mathbb{P}[b_i \text{ terminates its run}] = \underbrace{\frac{n_0}{n}}_{b_i=0} \cdot \underbrace{\frac{n_1}{n-1}}_{b_{i+1}=1} + \underbrace{\frac{n_0}{n}}_{b_i=1} \cdot \underbrace{\frac{n_1}{n-1}}_{b_{i+1}=0} = \frac{2n_0n_1}{n(n-1)}.$$

The number of runs should be asymptotically Gaussian with mean and variance

$$\mu = 1 + \frac{2n_0n_1}{n}, \qquad \sigma^2 = \frac{(\mu-1)(\mu-2)}{n-1}.$$

Thus, a test would consist in comparing the following statistic

$$\frac{\#runs - \mu}{\sigma}$$

to the quantiles of the standard normal distribution.

Ordinal data

The run test can be adapted to test independence in ordinal data. To do so, the entries should be compared to the median. Entries that are smaller than the median will be given a 0 and entries that are larger than the median will be given a 1. The run test can then be performed on the binary sequence that has been obtained.

3.1.3 Serial correlation

Given data entries $x_1, \cdots, x_n$ that arranged in time, the k-th self-correlation r_k as the sample correlation between pairs of entries that are k units apart

$$r_k = \frac{\sum_{i=k+1}^{n} \frac{(x_i - \bar{x})(x_{i-k} - \bar{x})}{n-k}}{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}.$$

Under the hypothesis of independence, r_k is approximately Gaussian with mean $-1/(n-1)$ and variance $1/n$.

3.2 Testing for normality

One way to test if data follow the normal distribution is to check if the estimated moments matches with the moments of the normal distribution. These can be derived from the moment generating function defined as

$$\phi_X(t) = \mathbb{E}[e^{tX}],$$

using the following formula

$$\mathbb{E}[X^n] = \left. \frac{d\phi_X(t)}{dt} \right|_{t=0}.$$

Theorem. *The moment generating function of the standard normal distribution is given by*

$$\phi_X(t) = e^{t^2/2}.$$

Proof. Left as an exercise. □

Thus, one way to check for normality of data entries is to standardise the data (subtract the mean and divide by the standard deviation) and compare the sample moments to the moments of the standard normal distribution.

3.3 Goodness of fit

3.3.1 Chi-squared test

Reminder

If $X_1, X_2, \cdots, X_n$ are i.i.d. from the standard normal distribution

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2.$$

Test for discrete distributions

Let us consider a sample $x_1, x_2, \cdots, x_n$. We want to test if it comes from a discrete law

$$p : i \in I \mapsto p_i.$$

Assuming that the sample is i.i.d. with law p, the number of occurrences of any class $i \in I$

$$n_i = \sharp\{k : x_k = i\}$$

should follow the binomial distribution with n trials and success probability p_i. In particular, when n goes to infinity we should expect a normal distribution

$$\frac{n_i - np_i}{\sqrt{np_i(1-p_i)}} \overset{n\to\infty}{\underset{\text{law}}{\longrightarrow}} \mathcal{N}(0,1).$$

In view of our previous comment, we might expect the following

$$\sum_{i\in I} \frac{(n_i - np_i)^2}{np_i(1-p_i)} \overset{n\to\infty}{\underset{\text{law}}{\longrightarrow}} \chi^2_{\sharp I}.$$

This does however assumes that the n_i are independent without accounting for the constraint

$$\sum_{i\in I} n_i = n.$$

In reality a degree of freedom is lost because of this constraint

$$\sum_{i\in I} \frac{(n_i - np_i)^2}{np_i(1-p_i)} \overset{n\to\infty}{\underset{\text{law}}{\longrightarrow}} \chi^2_{\sharp I-1}.$$

In practice, the law p may not be entirely specified and it may be a parametric distribution with, say, d parameters that has been fitted to the data. In this case we have

$$\sum_{i\in I} \frac{(n_i - np_i)^2}{np_i(1-p_i)} \overset{n\to\infty}{\underset{\text{law}}{\longrightarrow}} \chi^2_{\sharp I-1-d}.$$

In conclusion, in order to test if the sample comes from the distribution p, it suffices to calculate the statistic

$$\sum_{i\in I} \frac{(n_i - np_i)^2}{np_i(1-p_i)}$$

and to compare its value to the quantiles of the Chi square distribution with the adequate number of degrees of freedom.

Exercise

We want to test if a dice with 6 faces is well balanced. The dice was rolled 60 times with the following outcome

$$(n_1 = 10, n_2 = 12, n_3 = 10, n_4 = 10, n_5 = 9, n_6 = 9).$$

Test for continuous distributions

This test can be adapted to continuous distributions. In order to do so, the set of the outcomes need to be partitioned into smaller subsets. As a rule of thumb, there should be at least 5 observations in each of these subsets.

3.3.2 Kolmogorov–Smirnov test

The goal of this section is to check if a sample comes from a continuous law with cumulative distribution function

$$F : x \in \mathbb{R} \mapsto F(x).$$

To do so, we define the empirical distribution function

$$F_n(x) = \frac{\sharp\{x_i : x_i < x\}}{n} = \sum_{i=1}^{n} \frac{\mathbf{1}(x_i \leq x)}{n}.$$

As a consequence of the law of large numbers, we observe that if the sample is i.i.d. With distribution F, we have a pointwise convergence of the empirical distribution when n goes to infinity

$$\forall x \in \mathbb{R}, \qquad F_n(x) \overset{n\to\infty}{\longrightarrow} F(x).$$

This convergence can be strengthened with the Glivenko–Cantelli Theorem

Theorem.

$$\forall \varepsilon > 0, \qquad \mathbb{P}[\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \varepsilon] \overset{n\to\infty}{\longrightarrow} 1.$$

In practice, this theorem does not allow to realise statistical test. A stronger result states that the following statistic

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

follows the Kolmogorov distribution with n trials for which you can find tables of quantiles.

3.4 Test if two samples come from the same law

This time we consider two independent samples of respective size n and m

$$x_{1,1}, x_{1,2}, \cdots, x_{1,n}$$

$$x_{2,1}, x_{2,2}, \cdots, x_{2,m}.$$

Exercise

Use the Glivenko–Cantelli theorem and the triangular inequality to prove that if both samples follow the same distribution F

$$\forall \varepsilon > 0, \qquad \mathbb{P}[\sup_{x\in\mathbb{R}}|F_{1,n}(x) - F_{2,m}(x)| \leq \varepsilon] \underset{m\to\infty}{\overset{n\to\infty}{\longrightarrow}} 1.$$

Two samples Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test can be declined to check if two samples have the same distribution. To do so it suffices to compute the statistic

$$\sup_{x\in\mathbb{R}}|F_{1,n}(x) - F_{2,m}(x)|.$$

The critical values for a test of significance α have the form

$$c(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}}.$$

For instance, in order to realise a test at 95% confidence ($\alpha = 5\%$), we should take $c(\alpha) \simeq 1.358$. Other values of $c(\alpha)$ can be found in statistical tables.

Chapter 4

Least squares estimation and linear regression

4.1 A bit of optimisation

To minimise a function of a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$ you need to find the values $\hat{x}$ that cancel the first order derivatives

$$\nabla f(\hat{x}) = \left(\frac{\partial f}{\partial x_i}\right)\Bigg|_{\hat{x}} = 0$$

and to ensure that the Hessian matrix

$$\mathrm{Hess}_{\hat{x}}(f) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)\Bigg|_{\hat{x}}$$

is positive definite, for instance with the Sylvester's criterion or by finding the eigenvalues. If there are multiple such solutions, so-called local minima, they should be compared in order to find the global minimum.

4.2 Least squares estimation

We consider n random responses

$$Y_1, Y_2, \cdots, Y_n.$$

After observation we obtain the values

$$y_1, y_2, \cdots, y_n.$$

The expected value of the responses is assumed to be connected to a set of unknown parameters

$$\theta = (\theta_1, \theta_2, \cdots, \theta_k)$$

through a set of known functions

$$\mathbb{E}[Y_i] = g_i(\theta).$$

Our goal is to find the "best" θ. Here we consider that the responses are uncorrelated

$$\mathrm{Cov}(Y_i, Y_j) = 0, \ \ \forall i \neq j,$$

and have the same variance. If it exists, the least square estimation of θ is the solution of

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}}\, S(\theta),$$

where

$$S(\theta) = \sum_{i=1}^{n} (y_i - g_i(\theta))^2.$$

Exercise

We assume that $y_1, y_2 \cdots, y_n$ are observations of an i.i.d. sample. Determine the least square estimation of the mean.

Exercise

Two objects have unknown weight θ_1 and θ_2. A number of weighings are performed independently on a set of unbiased scales with identical variance. In $y_1, y_2 \cdots, y_k$ both objects are weighed together. In $y_{k+1}, \cdots, y_n$, only the first object is weighed. Find the least square estimator of (θ_1, θ_2).

4.3 Weighted least squares estimation

If the variance is not constant across sample, but is known relatively,

$$\omega_i = \frac{\mathbb{V}[Y_1]}{\mathbb{V}[Y_i]},$$

the least square estimator is the solution of

$$\hat{\theta} = \operatorname*{argmin}_{\theta} S_W(\theta),$$

where

$$S_W(\theta) = \sum_{i=1}^{n} \omega_i (y_i - g_i(\theta))^2.$$

4.4 Linear model

4.4.1 The model

In this section we regroup the random responses in a vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

To each response corresponds a set of explanatory variables, considered to be deterministic in this setting (for instance the explanatory variables may be set in advance by the experimenter)

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

The linear model assumes that there is a linear relationship between the expected responses that the explanatory variables i.e. that there exist a vector β with k entries such that

$$\mathbb{E}[Y] = X\beta.$$

Remark. It is common to incorporate an intercept in the model. This can be done by including a column of ones in the matrix of explanatory variables

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

4.4.2 Least square estimator

Given an observation of the responses

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

the least square estimator, if it exists, is defined as the solution of

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, S(\beta)$$

where

$$S(\beta) = ||y - X\beta||^2 = (y - X\beta)^\intercal (y - X\beta).$$

Theorem (Normal equation)**.** *The least square estimator must satisfy the following equation*

$$X^\intercal X\hat{\beta} = X^\intercal y.$$

Proof. In this proof we calculate the gradient of S by performing the calculations on the components of the matrices. The proof can be shortened by using the chain-rule and the rule for differentiating multilinear functions (bilinear in this case).

$$\begin{aligned} S(\beta) &= (y - X\beta)^\intercal (y - X\beta) = y^\intercal y - y^\intercal X\beta - \beta^\intercal X^\intercal y + \beta^\intercal X^\intercal X\beta \\ &= y^\intercal y - 2y^\intercal X\beta + \beta^\intercal X^\intercal X\beta \\ &= \sum_{i=1}^{n} \left[y_i^2 - 2y_i(X\beta)_i + (X\beta)_i^2 \right] \\ &= \sum_{i=1}^{n} \left[y_i^2 - 2y_i \left(\sum_{j=1}^{k} x_{ij}\beta_j \right) + \left(\sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \right]. \end{aligned}$$

Now let us calculate the derivatives

$$\frac{\partial S}{\partial \beta_l} = \left(-2 \sum_{i=1}^{n} y_i x_{il} \right) + 2 \sum_{i=1}^{n} x_{il} \sum_{j=1}^{k} x_{ij}\beta_j = -2\left(X^\intercal y\right)_l + 2(X^\intercal X\beta)_l.$$

We conclude the proof by observing that the least square estimator must cancel these. □

So far, the normal equation doesn't ensure that $\hat{\beta}$ is defined unequivocally. This must be confirmed by checking if the Hessian is positive definite. In this case the Hessian is constant as S is a quadratic function of β

$$\text{Hess } S = 2X^\intercal X.$$

Here it is positive but may not be definite.

Theorem. *If $k \leq n$ and rank$(X) = k$ (i.e. if the normal equation is not underdetermined), then $X^\intercal X$ is invertible and the normal equation has a unique solution*

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y.$$

We will keep this assumption in what follows. We can thus define the following random variable corresponding to the observation $\hat{\beta}$

$$\hat{B} = (X^\top X)^{-1} X^\top Y.$$

We can easily check that $\hat{B}$ is unbiased. As X is deterministic and the expectation is a linear operation

$$\mathbb{E}[\hat{B}] = (X^\top X)^{-1} X^\top \mathbb{E}[Y] = (X^\top X)^{-1} X^\top X\beta = \beta.$$

Under certain assumptions, we can produce confidence regions for β. If $\mathbb{V}[Y_i] = \sigma^2$ for all $i \in [\![1, n]\!]$ and $\mathrm{Cov}[Y_i, Y_j] = 0$ for all $i \neq j$, i.e. if $\mathrm{Cov}[Y] = \sigma^2 I_n$ (this is for instance the case if the responses are i.i.d.)

$$\mathrm{Cov}[\hat{B}] = \sigma^2 (X^\top X)^{-1}.$$

Theorem (Gauss–Markov)**.** *$\hat{B}$ is the BLUE (Best Linear Unbiased Estimator) in the sense that among all of the estimators of the form*

$$\tilde{B} = MY,$$

such that $\mathbb{E}[\tilde{B}] = \beta$, the matrix $Cov[\tilde{B}] - Cov[\hat{B}]$ is a positive matrix.

4.4.3 Fitted values and residuals

Now that we have an estimator $\hat{\beta}$, we can use it to estimate what would be the response for a given set of explanatory variables

$$\hat{y}_i = X_i \hat{\beta}.$$

If observations have been performed, the discrepancy between the observed responses and our fitted model is measured by the following residuals

$$\hat{\varepsilon} = y - \hat{y}$$

and the variance σ^2 under the previous assumptions is estimated by

$$S^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{\varepsilon}_i = \frac{||y - \hat{y}||^2}{n-k}.$$

4.4.4 Numerical issues

The normal equation is sensitive to perturbations for instance if X is replaced by a perturbed matrix $X(1+\delta)$, the error is amplified

$$(X + \delta X)^\top (X + \delta X) = X^\top X(1 + 2\delta + \delta^2).$$

There are also risks of underflow and of overflow. For instance if

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix},$$

we have

$$X^\top X = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{pmatrix}.$$

If ε is close to zero but can still be represented on the computer (as a 64 bits float for instance), it might happen that ε^2 is too small to be represented thus making $X^\intercal X$ practically non invertible. This kind of issues and how to handle them through various methods of matrix factorisation is the purpose of numerical analysis.

4.4.5 Weighted least squares

When the responses exhibit some correlations it is preferable to use a weighted version of the least squares. If $\mathrm{Cov}[Y] = \sigma^2 V$ for a known matrix V, the estimation would be

$$\hat{\beta} = (X^\intercal V^{-1} X)^{-1} X^\intercal V^{-1} Y.$$

Chapter 5

Parameters estimation

Here we assume that we have an i.i.d. sample

$$\mathbf{X} = (X_1, X_2, \cdots, X_n)$$

from a law that depends on a parameter θ or a set of parameters $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. We write Θ the set of the possible values for the parameters.

To make the dependence on θ more explicit, we may write the probability mass function $p(\cdot|\theta)$ for a discrete random variable, or the density (resp. the c.d.f.) $f(\cdot|\theta)$ (resp. $F(\cdot|\theta)$).

In practice we observe a realisation of **X**

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

5.1 Point estimation

The goal is estimating the value of θ from the sample, that is finding a function

$$T : \mathbf{x} \mapsto T(\mathbf{x}) \in \Theta,$$

that we expect to be "close" to the actual parameter θ. Given a set of random variables $\mathbf{X}$, we call $T(\mathbf{X})$ an estimator of θ. Given an observation of these variables, we call $T(\mathbf{x})$ an estimation (remark: we may also use the notation $\hat{\theta} = T(\mathbf{x})$).

We denote by f_T (resp. F_T, p_T) the density (resp. the c.d.f. , the p.m.f.). The distribution corresponding to $T(\mathbf{X})$ is called the sampling distribution.

Exercise

Find the sample law of the sample mean $\bar{X}$ if

- X is an i.i.d. sample following the Bernoulli distribution with success probability θ,
- X is an i.i.d. following the normal distribution with mean μ and variance σ^2.

5.2 Bias

(In this section we only consider cases with a single parameter, however similar notions exist for multiple parameters.)

So far, we just said that an estimator is a function of the observations. We have some intuitions about what constitutes a "good estimator". Although there is no universal definition of such thing, a reasonable approach may consists in taking unbiased or asymptotically unbiased estimators

$$\mathbb{E}[T(\mathbf{X})] = \theta$$

with a relatively small variance. By virtue of the Bienayme–Chebyshev inequality it means that the estimator is concentrated around the parameter that we wish to estimate

$$\forall \varepsilon > 0, \;\; \mathbb{P}\left[|T(\mathbf{X}) - \theta| > \varepsilon\right] \leq \frac{\mathbb{V}[T(\mathbf{X})]}{\varepsilon^2}.$$

However it might sometimes be preferable to use a biased estimator, as will see, unbiased estimators are intrinsically limited but firstly, we need to state the following definitions. The Fisher information of the i.i.d. sample $\mathbf{X}$ is defined as

$$I(\theta) = n \cdot \mathbb{E}\left[\left(\frac{\partial \ell(\theta|X_1)}{\partial \theta}\right)^2\right],$$

where

$$\ell(\theta|X_1) = f(X_1|\theta)$$

is the log-likelihood of X_1. It is a quantity that we will see again in the next chapter.

Theorem (Cramér–Rao)**.** *Let $T(\mathbf{X})$ be an unbiased estimator of θ. If the following sufficient conditions are satisfied*

- *The Fisher information is well defined,*

- *the integration with regards to* $\mathbf{x}$ *and the derivative with regards to* θ *can be swapped*

$$\frac{\partial}{\partial \theta} \int T(\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} = \int T(\mathbf{x}) \left[\frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) \right] d\mathbf{x}.$$

the variance of $T(\mathbf{X})$ *is bounded below as follows*

$$\mathbb{V}[T(\mathbf{X})] \geq \frac{1}{I(\theta)}.$$

In practice, the regularity conditions of the Cramér–Rao bound are often satisfied and we will not require to verify them in most problems treated in this module, although we will see cases where they are not (typically when the support of f depends on θ).

While this theorem is very convenient to prove that have a so-called minimum variance unbiased estimator, we should bear in mind that there may be no estimators that match the Cramér–Rao bound. Besides, while the conditions of the inequality are often satisfied, it is not difficult to construct instances of unbiased estimators that do not satisfy them.

The limitation set by the Cramér–Rao bound leads to think that, in some situations, we may prefer a biased estimator. If we do so, it is not appropriate to use the variance as a benchmark, as it measures the concentration around the mean of the estimator instead of the parameter to estimate. Instead, we should consider the mean squared error

$$MSE(T(\mathbf{X})) = \mathbb{E}[(T(\mathbf{X}) - \theta)^2].$$

An estimator $T_1(\mathbf{X})$ is said to be more efficient than another estimator $T_2(\mathbf{X})$ if

$$MSE(T_1(\mathbf{X})) \leq MSE(T_2(\mathbf{X})).$$

Exercise

Consider a sample $X_1, \cdots, X_n$ following the uniform distribution over $[0, \theta]$. Show that $2\bar{X}$ is an unbiased estimator of θ and calculate its variance. Does it satisfy the Cramér–Rao bound?

5.3 The Delta-method

(Again, in this section we only consider cases with a single parameter.)

As shown by the central limit theorem, $\bar{X}$ is asymptotically Gaussian, when the sample size grows to infinity. Linear combinations of independent Gaussian random variables are also Gaussian. But it is not uncommon to estimate parameters by applying non-linear transformations. The Delta-method is a powerful that shows that, under reasonable conditions, such estimators tend to be Gaussian.

Theorem. *Suppose that θ and σ^2 are finite and that $\hat{\theta}_n$ is such that*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

then for all C^2 function g such that $g'(\theta) \neq 0$

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \cdot g'(\theta)^2).$$

Proof. First of all let us write a Taylor approximation of $g(\hat{\theta}_n)$

$$g(\hat{\theta}_n) = g(\theta) + (\hat{\theta}_n - \theta)g'(\theta) + (\hat{\theta}_n - \theta)^2 O(1).$$

Thus we have

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) = \sqrt{n}\left[(\hat{\theta}_n - \theta)g'(\theta) + (\hat{\theta}_n - \theta)^2 O(1)\right].$$

Subsequently, we can ignore the remainder since $\sqrt{n}(\hat{\theta}_n - \theta)^2$ goes to zero almost surely. Finally we can prove the theorem by checking the convergence of the moment generating function.

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[\exp t\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right)\right] &= \lim_{n\to\infty}\mathbb{E}\left[\exp\left(tg'(\theta)\right)\sqrt{n}\left(\hat{\theta}_n - \theta\right)\right] \\ &= \exp\left(\frac{(tg'(\theta))^2\sigma^2}{2}\right).\end{aligned}$$

□

Example

Due to the central limit theorem, we can estimate the mean of an i.i.d. sample of the exponential distribution with rate λ

$$\sqrt{n}\left(\bar{X} - \frac{1}{\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right)$$

by applying the Delta method we can show that we can also directly estimate the sampling distribution of the corresponding estimator of λ

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \lambda\right) \xrightarrow{d} \mathcal{N}\left(0, \lambda^2\right)$$

5.4 Sufficiency

A statistic $T(\mathbf{X})$ is said to be sufficient for a parameter θ, if the law of $\mathbf{X}$ conditional on $T(\mathbf{X})$ doesn't depend on θ. In other words, a sufficient statistic contains all the information needed to estimate θ.

Example

Consider an i.i.d. sample $X_1, \cdots, X_n$ following the Bernoulli distribution with success probability θ. The statistic $T(X) = \sum_{i=1}^{n} X_i$ is sufficient for θ.

The following theorem provides a criterion for identifying sufficient statistics.

Theorem (Fisher–Neymann)**.** *A statistic $T(\mathbf{X})$ is sufficient for θ if and only if the density of $\mathbf{X}$ can be decomposed in the following way*

$$f_{\mathbf{X}}(x|\theta) = g(T(x), \theta) \cdot h(x).$$

In practice, it is possible to derive sufficient statistics for numerous distributions such as the ones belonging to the exponential family.

Theorem (Exponential family)**.** *If the density of a single observation can be factorised as*

$$f(x|\theta) = e^{K(x)p(\theta)+S(x)+q(\theta)}$$

then, the sum

$$\sum_i K(X_i)$$

is a sufficient statistic for θ derived from a sample.

Proof. The demonstration is left as an exercise. □

Exercise

List some examples of distributions that belong to the exponential family.

Chapter 6

Maximum likelihood estimation

Given a sample $X_1, \cdots, X_n$, assumed to be i.i.d. from a parametric model, the likelihood function measures how likelihood this sample would be under given values of the parameter. If the law has a density it is defined as

$$\mathcal{L}(\theta|\mathbf{X}) = f(\mathbf{X}|\theta) = \prod_{i=1}^{n} f(X_i|\theta),$$

if the law is discrete

$$\mathcal{L}(\theta|\mathbf{X}) = p(\mathbf{X}|\theta) = \prod_{i=1}^{n} p(X_i|\theta).$$

We will see that it can even be adjusted to handle censoring, that is when only partial information is available about the sample. This function has several uses, including the estimation of parameter through it maximisation, or the comparison of different models.

6.1 Maximum likelihood estimator

The maximum likelihood estimator, if it is well defined, is the random variable defined as the solution of

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta|\mathbf{X})$$

In practice, it is preferable to maximise the log-likelihood

$$\ell(\theta|\mathbf{X}) = \log \mathcal{L}(\theta|\mathbf{X}),$$

in order to leverage the linearity of the differential. This is also preferable from a numerical standpoint as it is a way to avoid underflow. Assuming that it is C^2, it can be maximised by finding critical points that cancel the first order derivatives

$$\nabla \ell(\hat{x}) = \left(\frac{\partial f}{\partial x_i}\right)\bigg|_{\hat{x}} = 0$$

and to ensure that the Hessian matrix

$$\mathrm{Hess}_{\hat{x}}(f) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)\bigg|_{\hat{x}}$$

is negative definite. This gives local maxima that should then be compared in order to find the global maximum.

Exercise

- Find the maximum likelihood estimator of θ if $X_1, \cdots, X_n$ is an i.i.d. sample with distribution $\mathcal{B}(\theta)$.
- Find the maximum likelihood estimator of μ and σ^2 if $X_1, \cdots, X_n$ is an i.i.d. sample with distribution $\mathcal{N}(\mu, \sigma^2)$.
- Find a method of moments and the maximum likelihood estimators of σ^2 if $X_1, \cdots, X_n$ is an i.i.d. sample following the Rayleigh distribution

$$f(x|\theta) = \frac{x}{\sigma^2} \cdot e^{-x^2/(2\sigma^2)}, \quad x \geq 0.$$

6.2 Handling censored data

What if we cannot obtain full information about the realisation of a random variable? A common scenario in survival analysis involves considering the times until a patient dies. It may happen that the patient leaves the study before dying, or dies from a cause not investigated in the study. This is known as right-censoring. While we do not know the exact time of death, we can still ascertain that the patient survived until leaving the study, and we may wish to use this information.

In the survival analysis example we gave, we may consider that we have a sample of times $\mathbf{T} = (T_1, \cdots, T_n)$. Among these, some are right-censored and we may replace the density (or the probability mass function) by the survival function (i.e. the complementary cumulative distribution). Hence the likelihood

$$\mathcal{L}(\theta|\mathbf{T}) = \prod_{i \text{ censored}} (1 - F(T_i|\theta)) \cdot \prod_{i \text{ non censored}} f(T_i|\theta).$$

Similarly we may consider left-censoring or interval censoring.

Exercise

Derive the maximum likelihood estimator for a sample $T_1, \cdots, T_n$, containing censored entries and following the exponential distribution with rate λ.

6.3 Relative likelihood

Relative likelihood is used to compare the plausibility of different parameter values for a statistical model, given the observed data. For instance, if θ_1 and θ_2 are possible values of θ, the likelihood ratio for θ_1 versus θ_2 is $\mathcal{L}(\theta_1|\mathbf{X})/\mathcal{L}(\theta_2|\mathbf{X})$. If this ratio is greater than 1, the parameter θ_1 is consider to be more plausible than θ_2.

As a reference, we may take the maximum likelihood estimator and define the relative likelihood function

$$\theta \mapsto R(\theta) = \frac{\mathcal{L}(\theta|\mathbf{X})}{\mathcal{L}(\hat{\theta}|\mathbf{X})}.$$

The level set of this function, also called likelihood regions

$$LR(\alpha) = \{\theta \in \Theta : R(\theta) \geq \alpha\}, \quad 0 \leq \alpha \leq 1$$

are a convenient way to summarise the plausible values of θ.

- Values inside the $\alpha = 50\%$ likelihood region are considered as very plausible,
- Values inside the $\alpha = 10\%$ likelihood region are considered as plausible,
- Values outside the $\alpha = 10\%$ likelihood region are considered as implausible,
- Values outside the $\alpha = 1\%$ likelihood region are considered as very implausible.

As usual, in practice it may be more convenient to work with the logarithm likelihood function, in which case we should use the following thresholds for the logarithm of the likelihood ratio

- $\log(50\%) \simeq -0.69$,
- $\log(10\%) \simeq -2.30$,
- $\log(1\%) \simeq -4.61$.

Example

Consider $X_1, \cdots, X_n$ i.i.d. following the exponential distribution with rate parameter λ. The log-likelihood function is

$$\ell(\lambda|\mathbf{X}) = n(\log(\lambda) - \lambda\bar{X})$$

and the maximum likelihood estimator is $\hat{\lambda} = 1/\bar{\mathbf{X}}$. In particular, we can evaluate the log-likelihood in $\hat{\lambda}$

$$\ell(\hat{\lambda}|\mathbf{X}) = n(\log(\hat{\lambda}) - 1).$$

Now, let us consider a perturbation of the estimator

$$\tilde{\lambda} = \hat{\lambda} + \varepsilon,$$

where the perturbation ε is assumed to be close to 0. In order to estimate the log-relative likelihood, we can perform a Taylor expansion

$$\ell(\tilde{\lambda}|\mathbf{X}) - \ell(\hat{\lambda}|\mathbf{X}) = -\frac{n}{2} \cdot \frac{\varepsilon^2}{\hat{\lambda}^2} + no(\varepsilon^2).$$

We can use this to approximate the likelihood regions

$$LR(\alpha) = \{\lambda \in \mathbb{R}_+ : R(\lambda) \geq \alpha\} \simeq \left[\hat{\lambda}\left(1 - \sqrt{\frac{-2 \cdot \log \alpha}{n}}\right), \hat{\lambda}\left(1 + \sqrt{\frac{2 \cdot \log \alpha}{n}}\right)\right].$$

As we have more data, i.e. n gets larger, we have smaller likelihood regions.

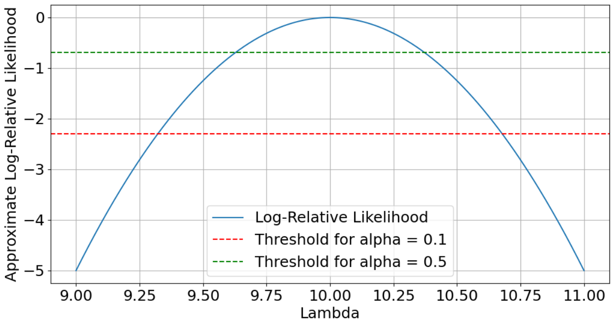


Figure 6.1: Log-relative likelihood for $n = 10^4$ and $\hat{\lambda} = 10$.

6.4 Notable properties of maximum likelihood estimators

- If $T(\mathbf{X})$ is a sufficient statistic for θ, the maximum likelihood estimator of θ is a function of $T(\mathbf{X})$. This can easily be checked with the Fisher–Neymann theorem

  $$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta|\mathbf{X}) = \underset{\theta}{\operatorname{argmax}}\ g(T(\mathbf{X},\theta)) \cdot h(x) = \underset{\theta}{\operatorname{argmax}}\ g(T(\mathbf{X},\theta)).$$

  This is consistent with the concept of sufficiency, confirming that we can extract the necessary about θ information from sufficient statistics.

- The maximum likelihood estimator is asymptotically unbiased and Gaussian

  $$\hat{\theta} \xrightarrow{law} \mathcal{N}(\theta, I(\theta)^{-1}),$$

  where $I(\theta)$ is the Fisher information of the sample $\mathbf{X}$.

- The maximum likelihood estimator is invariant under re-parametrisation of the distribution i.e. if g is a bijection from Θ and $\hat{\theta}$ is the maximum likelihood estimator of θ then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$. For instance if we consider the exponential distribution, $\bar{X}$ is the maximum likelihood estimator of the mean and $1/\bar{X}$ is the maximum likelihood estimator of the rate.

Chapter 7

Significance tests and hypothesis testing

7.1 Significance test

Suppose that we want to investigate if an hypothesis H_0 (null hypothesis) is valid by observing a random sample $\mathbf{X} = (X_1, \cdots, X_n)$. A significance test is a procedure that evaluates if the sample is consistent with H_0, in which case we fail to reject the hypothesis.

The consistency of the data with H_0 is examined by calculating a p-value, that is the probability to obtain the observed data or more extreme results under the assumption that H_0 is true. The p-value is roughly interpreted as follows

- $p > 0.1$, the data are reasonably consistent with H_0,

- $p \simeq 0.05$, we have moderate evidence against H_0,
- $p < 0.01$, we have strong evidence against H_0.

It's important to note that this choice of thresholds is arbitrary and should depend on the context of the hypothesis test.

Designing a test consists in defining an acceptance region A, such that we accept H_0 if the event $\mathbf{X} \in A$ occurs and we reject H_0 if it doesn't.

In practice we test H_0 against an alternative H_1, which doesn't need to be its "complementary". It is important to bear in mind that the hypothesis H_0 and its alternative don't have a symmetric role. In many scenarios, H_0 represents the default position, for instance if we test a new drug it states that the drug has no effect, the hypothesis H_1 is representative of what we would consider to be a meaningful deviation. In this context it is preferable to have strong evidence before we can accept the new drug.

Example

We have a single observation of a time T assumed to follow the geometric distribution with real parameter θ

$$\mathbb{P}[T = t|\theta] = \theta(1-\theta)^t, \quad t \in \mathbb{N}_0.$$

We want to test the hypothesis

$$H_0 : \quad \theta = \theta_0$$

against

$$H_1 : \quad \theta > \theta_0.$$

We remark that if H_0 is valid, we shouldn't observe times that are "excessively" short, since the expected value should be $(1-\theta_0)/\theta_0$. In particular, we may set a threshold $t^* \in \mathbb{N}$ and reject the hypothesis if T exceeds t^*. One way to set this threshold would be the following

$$\mathbb{P}[T < t^* | \theta = \theta_0] = 0.05 \Leftrightarrow 1 - (1-\theta_0)^{t^*} = 0.05 \Leftrightarrow t^* = \frac{\log 0.95}{\log(1-\theta_0)}.$$

In this case the acceptance region is $A = [0, t^*[$.

In this the hypothesis H_0 is said to be simple, this means that under the hypothesis H_0 the whole sampling distribution is specified. The hypothesis H_1 is composite, it covers a range of possible distributions (of parameters in this specific case).

7.2 Performance of a test

7.2.1 Type I and Type II errors

It is reasonable to say that a "good" test should have a low error rate. For a single test we distinguish two kinds of errors

- Type I (false positive) : H_0 is rejected despite being true,

- Type II (false negative) : H_0 is accepted despite being wrong.

	H_0 **is true**	H_0 **is false**
Reject H_0	False positive	True positive
Fail to reject H_0	True negative	False negative

The probability of Type I error is often designated by

$$\alpha = \mathbb{P}[X \in A^c|H_0]$$

while the probability of Type II error is often designated by

$$\beta = \mathbb{P}[X \in A|H_1].$$

Of course, in general, we cannot minimise simultaneously the probability of both types of errors. Moreover the computation α supposes that H_0 is simple, while the computation of β supposes that H_1 is simple. For instance, a test that always rejects H_0 produces no Type I errors but always produces Type II errors.

The performance of a test can also be summarised in a confusion matrix

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

it is also used to summarise the performance of classification models with more than two classes. From this matrix numerous metrics of performance are derived.

7.2.2 Significance level and power

The significance level of a test is the probability false positive α. This is usually the first quantity that is set when we design a test. Afterward, the test should be designed in order to maximise the power $1 - \beta$, that is the probability of true positive.

Example

Let us consider our previous example but with two simple hypothesis instead. This time we want to test the hypothesis

$$H_0 : \quad \theta = \theta_0$$

against

$$H_1 : \quad \theta = \theta_1,$$

for $\theta_0 < \theta_1$. Following the procedure that we gave, we might set a constraint on t^* to ensure a significance of 0.05

$$\mathbb{P}[T < t^* | \theta = \theta_0] \leq 0.05 \Leftrightarrow t^* \leq \frac{\log 0.95}{\log(1 - \theta_0)}.$$

Under this constraint we now want to minimise the Type II error (or maximise the power)

$$t^* = \underset{t \leq \log 0.95 / \log(1 - \theta_0)}{\operatorname{argmin}} (1 - \theta_1)^t = \frac{\log 0.95}{\log(1 - \theta_0)}.$$

In this simple case the threshold t^* matches exactly with our bound

7.3 Neyman–Pearson lemma

The goal of the following result is to design the Uniformly Most Powerful (UMP) test of significance α, that is the test of significance α that minimises the Type II error.

Theorem. *Consider both H_0 and H_1 are simple i.e. we test*

$$H_0 : \theta = \theta_0$$

against

$$H_1 : \theta = \theta_1.$$

The UMP test of level α is defined by the acceptance region

$$A = \left\{ \mathbf{x} : \frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\theta_1|\mathbf{x})} > k_\alpha \right\}$$

where

$$\mathbb{P}[\mathbf{X} \in A^c | \theta = \theta_0] = \alpha.$$

Example

We consider an i.i.d. sample $\mathbf{X}_1, \cdots, \mathbf{X}_n$ following the distribution $\mathcal{N}(\mu, \sigma^2)$, with known σ^2. We want to test

$$H_0 : \mu = \mu_0$$

against

$$H_1 : \mu = \mu_1.$$

where $\mu_1 > \mu_0$. Find the UMP test of level $\alpha = 0.05$.

Example (a pathological case?)

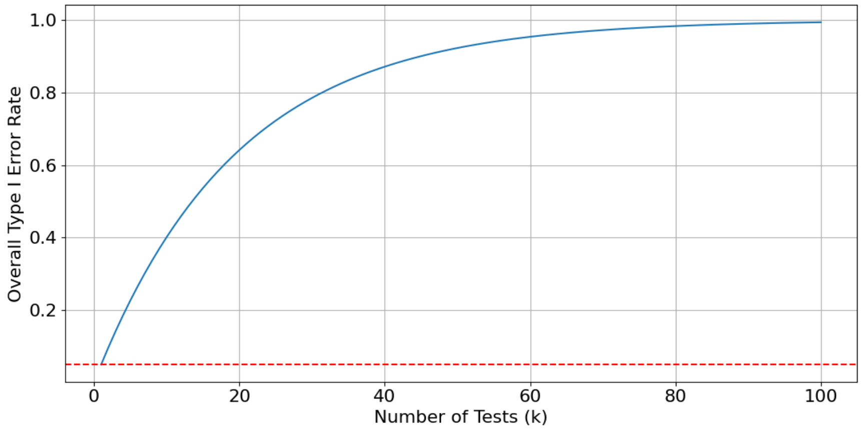
We observe a single time T. We want to test if its density is given by one of the following

$$H_0 : f(t|H_0) = \lambda \cdot \exp(-\lambda t), \quad t \geq 0$$
$$H_1 : f(t|H_1) = \lambda \cdot \exp(-\lambda(t-\delta)), \quad t \geq \delta,$$

where $\delta > 0$.

7.4 Multiple testing

Suppose we consider multiple hypothesis $H_{0,1}, H_{0,2}, \cdots, H_{0,k}$. Each hypothesis is tested by checking if $\mathbf{X}$ belongs to an acceptance region A_k. Naively we may set a significance level α (say 0.05) for each test. In this case the overall Type I error would be $\mathbb{P}[\exists i,\ \mathbf{X} \in A_i^c | H_0] = 1 - \mathbb{P}[\forall i,\ \mathbf{X} \in A_i | H_0]$. If we assume that the tests are conducted independently, this is equal to $1 - (1-\alpha)^k$.



A simple way to solve this issue is to apply a Bonferroni correction, consisting in taking any combination of α_i such that $\sum_{i=1}^{k} \alpha_i = \alpha$. For instance, we may take $\alpha_i = \alpha/k$.

Chapter 8

Confidence regions

Example

Suppose we have an i.i.d. sample $X_1, \cdots, X_n$ from the normal distribution with unknown mean μ and unknown variance σ^2. We usually estimate the mean by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and the variance by

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

A confidence interval of level $1-\alpha$ for μ would be

$$\left[\bar{X} - \frac{t_{\alpha/2,n-1}\cdot s}{\sqrt{n}}, \bar{X} + \frac{t_{\alpha/2,n-1}\cdot s}{\sqrt{n}}\right],$$

where $t_{\alpha/2,n-1}$ is the quantile of probability $\alpha/2$ for the Student distribution with $n-1$ degrees of freedom.

Example

Suppose we have an i.i.d. sample $X_1, \cdots, X_n$ following the uniform distribution over $[0, \theta]$, with $\theta > 0$. Let us consider the statistic

$$M = \max_i X_i.$$

We have

$$\mathbb{P}[M/\theta \leq t_\alpha] = t_\alpha^n$$

thus if we take $t_\alpha = \alpha^{1/n}$ we have $\theta \in [M, M/t_\alpha]$ with probability $1 - \alpha$.

8.1 Pivotal quantity and confidence region

Let us assume that we want to estimate a parameter θ. A pivotal quantity is a function of the observations and of θ for which the distribution doesn't depend on θ. For instance, in the first example

$$T(\mathbf{X}, \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

was a pivotal quantity for μ (it follows the Student distribution with $n - 1$ degrees of freedom). In the second example

$$T(\mathbf{X}, \theta) = \max_i \frac{X_i}{\theta}$$

was a pivotal quantity for θ as the X_i/θ follow the uniform distribution over $[0, 1]$.

We observe that in general taking a function of multiple pivots that doesn't explicitly depends on θ will produce a new pivot.

A confidence region of level $1 - \alpha$ is a subset of Θ derived from a pivot that contains the real parameter θ with probability $1 - \alpha$. Ideally a pivot would have a one-to-one correspondence with the parameter of interest and the diameter of the confidence region should be as small as possible.

In a nutshell: producing a confidence region involves

1. Finding a pivot $T(\mathbf{X}, \theta)$.

2. Finding the distribution of the pivot and deduce a region, say E, such that the $\mathbb{P}[T(\mathbf{X}, \theta) \in E|\theta] = 1 - \alpha$.

3. Taking the pre-image of E by the function $\theta \mapsto T(\mathbf{X}, \theta)$.

It is not always easy to find a pivotal quantity and even once we have a pivot its distribution might not be known. We may sometimes rely on asymptotic results (such as the central limit theorem) to approximate confidence regions.

8.2 Pivotal quantities for location-scale families

In this section, for simplicity we restrict ourselves to real valued random variables but these notions can be extended to vector valued random variables.

8.2.1 Useful properties

- If X and Y are two independent random variables with respective densities f_X and f_Y then the density f_{X+Y} of $X+Y$ is the convolution

$$f_{X+Y}(x) = f_X * f_Y(x) = \int f_X(t) f_Y(x-t)\, dt.$$

- If a is positive, the density f_{aX} of aX is given by

$$f_{aX}(x) = a^{-1} f_X(a^{-1}x).$$

8.2.2 Location families

A real-valued distribution belongs to a scale family if it has a cumulative distribution that can be expressed in the following way

$$F(x|a) = F\left(x-a\middle|0\right),$$

where $F(\cdot|0)$ is called the baseline. In particular its density satisfies

$$f(x|a) = f(x-a|0).$$

Theorem. *Suppose $X_1, \cdots, X_n$ are i.i.d. with a cumulative distribution $F(\cdot|a)$ that belongs to the location family with baseline $F(\cdot|0)$. Then $\bar{X}-a$ is a pivot for a.*

Proof.

$$\mathbb{P}[\bar{X}-a \leq x|a] = \mathbb{P}[\bar{X} \leq x+a|a] = \int_{-\infty}^{x+a} f(y-a|0)^{*n}\, dy = \int_{-\infty}^{x} f(z|0)^{*n} dz$$

□

Remark. We can deduce a confidence region by observing the following

$$\mathbb{P}[l \leq \bar{X} - a \leq L|a] = \mathbb{P}[\bar{X} - L \leq a \leq \bar{X} - l|a].$$

This gives regions of the form $[\bar{X} - L, \bar{X} - l]$.

8.2.3 Scale families

A real-valued distribution belongs to a scale family if it has a cumulative distribution that can be expressed in the following way

$$F(x|b) = F\left(b^{-1}x\middle|1\right),$$

where $F(\cdot|1)$ is called the baseline. In particular its density satisfies

$$f(x|b) = b^{-1}f(b^{-1}x|1)$$

In this case the scale parameter b should be strictly positive. While we will not develop this idea, this concept can be extended to higher dimension data, in this case b should be a positive definite matrix.

Theorem. *Suppose $X_1, \cdots, X_n$ are i.i.d. with a cumulative distribution $F(\cdot|b)$ that belongs to the scale family with baseline $F(\cdot|1)$. Then $b^{-1}\bar{X}$ is a pivot for b.*

Proof. It follows immediately from the properties of section 8.2.1 .

$$f_{b^{-1}X_i}(x|b) = bf(bx|b) = f(x|1).$$

□

Remark. as previously we can use this to draw confidence regions.

Theorem. *Suppose $X_1, \cdots, X_n$ are i.i.d. with a cumulative distribution $F(\cdot|b)$ that belongs to the scale family with baseline $F(\cdot|1)$. Denote S^2 as the sample variance. Then S^2/b^2 is a pivot for b.*

Proof. By definition

$$\frac{S^2}{b^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i}{b} - \frac{\bar{X}}{b}\right)^2$$

□

8.2.4 Location-scale families

A real-valued distribution belongs to a location-scale family if it has a cumulative distribution that can be expressed in the following way

$$F(x|a,b) = F\left(b^{-1}(x-a)\middle|0,1\right),$$

In this case we have the following pivots for a and b

$$\frac{\bar{X}-a}{S}, \ \frac{S^2}{b^2}.$$

The first pivot can be used to produce confidence region for (a, b) while the second won't provide any useful information about a.

8.3 Maximum likelihood procedure

We have seen in chapter 6 that if $\hat{\theta}$ is the maximum likelihood estimator of θ, then

$$\hat{\theta} \xrightarrow{law} \mathcal{N}(\theta, I(\theta)^{-1}),$$

where $I(\theta)$ is the Fisher information of the sample $\mathbf{X}$. This can be used to produce asymptotic confidence regions. For instance, a symmetric confidence region of level $1 - \alpha$ would be

$$[\hat{\theta} - L, \hat{\theta} + L]$$

where

$$L = \sqrt{I(\theta)^{-1}} \cdot Z_{1-\alpha/2}.$$

Chapter 9

Introduction to Bayesian statistics

9.1 The Bayesian paradigm

9.1.1 Setting

We assume that $\mathbf{X} = (X_1, \cdots, X_n)$ is a set of i.i.d. variables drawn from a distribution with probability mass function/density $p_{\mathbf{X}}(\mathbf{x}|\theta)$, which is called the likelihood in the Bayesian setting. Unlike the frequentist approach considered previously, the parameter θ is assumed to have its own distribution, denoted as $\pi(\theta)$. This distribution $\pi(\theta)$, known as the prior distribution, represents our beliefs about θ before observing any realisation of $\mathbf{X}$.

9.1.2 The Bayes formula

After observing a realisation $\mathbf{x}$ of $\mathbf{X}$, the distribution of θ should be updated according to this new information. This is done by using the Bayes formula (proven during tutorial)

$$\pi(\theta|\mathbf{x}) = \frac{p_{\mathbf{X}}(\mathbf{x}|\theta) \cdot \pi(\theta)}{p_{\mathbf{X}}(\mathbf{x})}$$

where $\pi(\theta|\mathbf{x})$ is known as the posterior distribution.

In practice, it is not always necessary to calculate the normalisation factor $p(\mathbf{x})$, thus we usually apply the following version of the Bayes formula:

$$\pi(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta) \cdot \pi(\theta).$$

For instance, the Python package 'scipy' allows defining the prior and the posterior up to a proportionality constant. Analytically, it may be sufficient to know the posterior distribution up to a proportionality constant in order to determine if it belongs to a well-known family.

In fact, the prior distribution does not need to be a well-defined probability distribution in the conventional sense (i.e., a positive function that integrates to 1). Instead, the prior can simply be any non-negative function that expresses the relative plausibilities of different parameter values before observing the data.

As an immediate consequence of the Fisher–Neyman factorization theorem, if T is a sufficient statistic, the posterior distribution can be expressed in terms of T.

9.1.3 Uninformed/uniform prior

In the absence of information about the parameter θ, except from its range Θ, we may assume that the prior distribution is constant over Θ

$$\pi(\theta) \propto \mathbf{1}_\Theta(\theta).$$

Example

We consider an i.i.d. sample $X_1, \cdots, X_n$ from the Bernoulli distribution without parameter $\theta \in [0, 1]$. The posterior distribution is the Beta distribution, given by

$$\pi(\theta|\mathbf{x}) = \frac{\theta^{n\bar{\mathbf{x}}} \cdot (1-\theta)^{n-n\bar{\mathbf{x}}} \cdot \mathbf{1}_{[0,1]}(\theta)}{p_\mathbf{X}}.$$

where the normalisation factor can be expressed with the Beta function

$$p_\mathbf{X}(\mathbf{x}) = \int_0^1 \theta^{n\bar{\mathbf{x}}} \cdot (1-\theta)^{n-n\bar{\mathbf{x}}}\, d\theta = B(n\bar{\mathbf{x}}+1, n-n\bar{\mathbf{x}}+1).$$

The analysis of this distribution requires manipulations involving the Gamma and Beta functions, which are not covered in this module (although you might have encountered them in a complex analysis module). We will just indicate the mode

$$\underset{\theta\in[0,1]}{\operatorname{argmax}}\, \pi(\theta|\mathbf{x}) = \bar{\mathbf{x}}$$

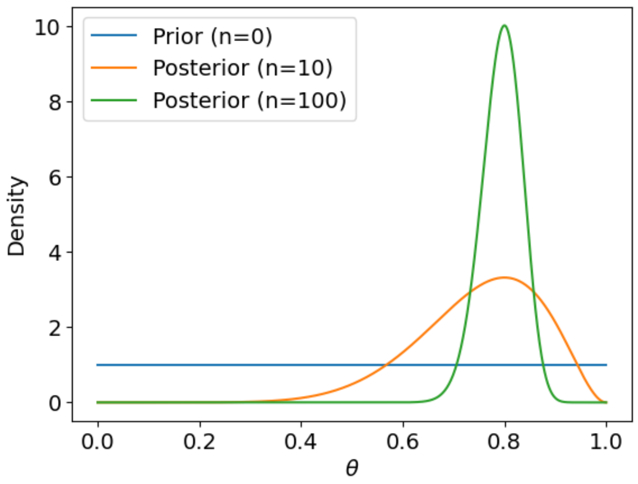
the expectation

$$\mathbb{E}[\theta|\mathbf{x}] = \frac{n\bar{\mathbf{x}}+1}{n+2} \sim \bar{\mathbf{x}}$$

and the variance

$$\mathbb{V}[\theta|\mathbf{x}] = \frac{(n\bar{\mathbf{x}}+1)(n\bar{\mathbf{x}}+2)}{(n+2)(n+3)} - \left(\frac{n\bar{\mathbf{x}}+1}{n+2}\right)^2 = O(1/n).$$

In particular, we observe that the distribution is getting more and more concentrated near $\bar{\mathbf{x}}$, which is consistent with intuition.



9.1.4 Jeffreys' prior

One issue of the uniform prior is that it depends on the parametrisation of the distribution, this is why it can be preferable to consider the Jeffreys prior

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

Indeed, we observe that if we take a function $\psi(\theta)$ which is C^1 and strictly monotonic, if we denote by $\tilde{\psi}$ the Jeffreys prior in this new parametrisation a simple application of the chain rule shows that

$$\tilde{\pi}(\psi) = \left|\frac{\partial\theta}{\partial\psi}\right| \cdot \pi(\theta).$$

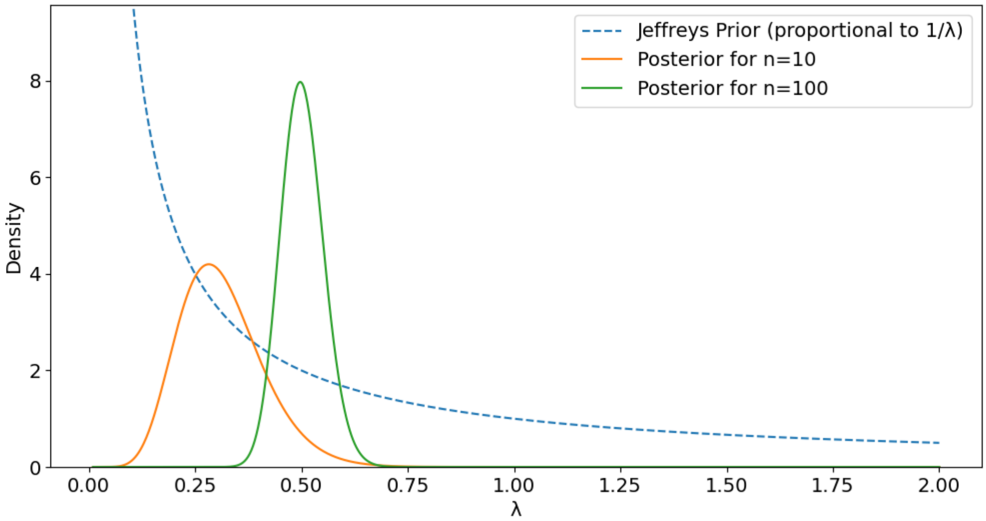
Example

We consider an i.i.d. sample $X_1, \cdots, X_n$ from the exponential distribution with rate parameter λ. The Jeffreys prior is

$$\pi(\lambda) \propto \frac{\mathbf{1}_{(0,\infty)}(\lambda)}{\lambda}.$$

The posterior is a Gamma distribution

$$\pi(\lambda \mid x_1, \ldots, x_n) = \frac{(n\bar{\mathbf{x}})^n}{\Gamma(n)} \cdot \lambda^{n-1} \cdot e^{-n\bar{\mathbf{x}}\lambda}.$$



9.1.5 Gaussian prior

We assume that $\mathbf{X} = (X_1, \cdots, X_n)$ is a set of i.i.d. variables following $\mathcal{N}(\theta, \sigma^2)$ and that θ has the prior distribution $\mathcal{N}(b, d^2)$. This situation might occur if prior to any observations we already have some expert information about the expected parameter. In this case, the posterior is also Gaussian

$$\mathcal{N}\left(\frac{n\bar{\mathbf{x}}/\sigma^2 + b/d^2}{n/\sigma^2 + 1/\sigma^2}, \frac{1}{n/\sigma^2 + 1/d^2}\right).$$

If the prior information is very limited, that is if the variance d^2 goes to infinity, this reduces to a familiar distribution $\mathcal{N}(\bar{\mathbf{x}}, \sigma^2/n)$.