Applied Statistics and Data Analysis

3. A review of inference concepts
b. Statistical inference

Paolo Vidoni

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

paolo.vidoni@uniud.it

Based mainly on Chapter 4 of the course textbook *A review of inference concepts*

Table of contents

Summary

- **Introduction to statistical inference** (*for individual revising*)
- **Basic concepts of point estimation** (*for individual revising*)
- **Basic concepts of interval estimation** (*for individual revising*)
- **Basic concepts of hypothesis testing** (*for individual revising*)
- **Basic concepts of model selection**
- **Contingency tables**

Introduction to statistical inference

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

- Making **inferences** about a population or about an interest phenomenon, based on a **random sample**, is a major task in statistical inference.
- Methods for analyzing data characterized by an inherently random variability so that the **conclusions** drawn are **generally valid**, even though obtained from a single set of data.
- For the most part this involves the use of **parametric statistical models**, which are suitable families of probability distributions, specified by one or more **unknown parameters**, describing hopefully how the data might have been generated.
- The aim is to **infer the values of the unknown model parameters** that are consistent with observed data, and to provide a measure of the **accuracy of the inferential conclusion**.
- The focus of interest may be the **interpretation** of the model parameters, and then of the interest phenomenon, or the **prediction** of future observations or outcomes using the estimated model.

- The available **data** $y = (y_1, \ldots, y_n)$ are analyzed as **observations of a random vector** $Y = (Y_1, \ldots, Y_n)$, following an unknown joint probability distribution.
- A **parametric statistical model** is a family of joint density (probability) functions $f(y_1, \ldots, y_n; \theta)$, $\theta \in \Theta$, which hopefully contains the unknown generating probability distribution or a at least a suitable approximation of it.
- The quantity θ, which specifies the density (probability) functions of the family, is a vector of **unknown parameters**; some of these parameters would answer the questions of interest about the system generating y.
- Statistical model may also depend on some further data x that are usually treated as known and called **covariates** or **predictor variables**.
- If the value of θ were known, a correct statistical model would allow the simulation of random data vectors which resemble the observed data y.

Random sampling

- A **random sample** is a set of units selected from a larger population, and in a (uniform) random sample all the elements of the population have the same chance to be included in the sample.
- A **random sample** describes also repeated observations of a random experiment or of a random phenomenon.
- Random sampling is the backbone of **statistical inference**, the collection of methods that allow to draw conclusions from a sample that are valid for the entire population or for the interest random phenomenon.
- Since the observed sample data y can be always thought as generated by random vector Y, the analyst may assume some specific properties about the distribution of the marginal component random variables (r.v.'s), like **independence and identical distribution** (i.i.d.), **independence** but not identical distribution or some specific form of **dependence**.

Example: temperatures

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

A simple model would treat the data as independent observations from normal distribution $N(\mu, \sigma^2)$, with unknown parameters $\theta = (\mu, \sigma^2)$.

Then the density function of a single measurement Y_i, $i = 1, \ldots, 60$, is

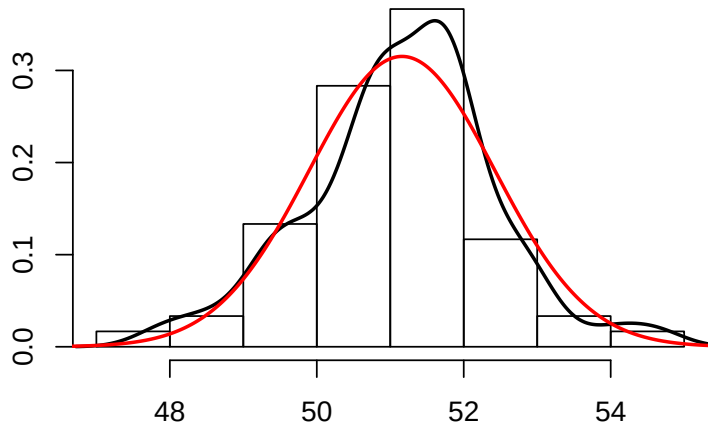
$$f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\}$$

and the joint density of the vector data Y is

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{60} f(y_i; \mu, \sigma^2)$$

Numerical summaries $\bar{y} = 51.16$, $y_{0.5} = 51.20$ may provide a "guess" for μ, whereas $s^2 = 1.60$ a "guess" for σ^2.

Estimates of the the generating probability distribution: histogram, **smooth density estimate**, normal density with $\mu = 51.16$ and $\sigma^2 = 1.60$



The tails seem heavier than those of the normal density; a better model might be to consider the data as independent observations from a Student's t distribution; namely, $Y_1, \ldots, Y_{60}$ i.i.d. r.v.'s such that

$$\frac{Y_i - \mu}{\sigma} \sim t(k),$$

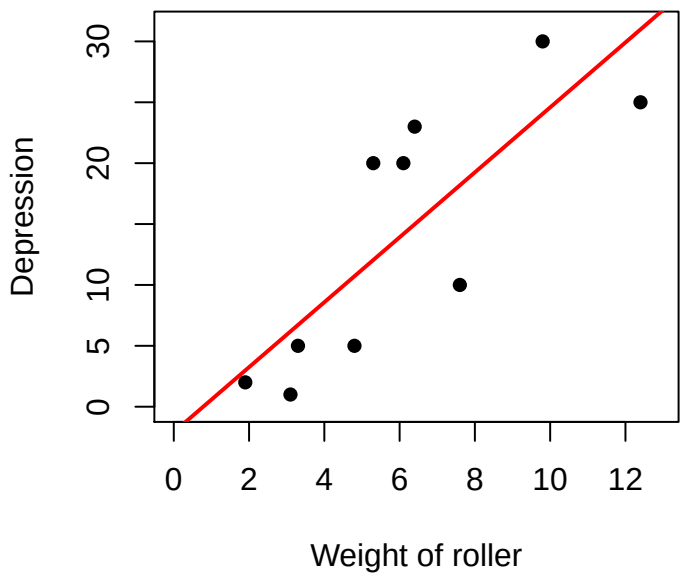
with $\theta = (\mu, \sigma, k)$ unknown parameters.

Example: roller data

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

Vector y includes data on the `depression` measured (**response variable**) and vector x includes the `weights` of the roller (**covariate** or **predictor variable**), which are taken as fixed.

Graphical summary (scatterplot of the data, with the least squares line) suggests that a useful model might be the linear regression model



The **simple linear regression model** has a *linear deterministic part* and an additive *error term* so that

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

with the assumption that, given the covariate x_i, the **error term** is **normally distributed**, $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are **independent**.

This amounts to say that, given x_i (which is taken as fixed), the observed response y_i is viewed as a realization of the r.v. $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, **independent** from the other **response r.v.'s**.

Here $\theta = (\alpha, \beta, \sigma^2)$ and the values $\widehat{\alpha} = -2.087$, $\widehat{\beta} = 2.667$, obtained with the least squares method, are a plausible "guess" for α and β.

Concerning the **interpretation of the model**, a crucial parameter is β, namely the rate of increase of depression with increasing roller weight.

With regard to **prediction**, using the estimated regression model, it is possible to predict the depression corresponding to out-of-sample roller weights, though some care is required.

Inferential questions

Given some data, y, and a statistical model with unknown parameters θ, there could be four basic points to consider:

1. find values for θ which are are most consistent with data y: **point estimation**;
2. find ranges of values (usually, intervals) for θ which are consistent with data y: **interval estimation**;
3. evaluate if some prespecified restriction on θ (hypothesis) is consistent with data y: **hypothesis testing**;
4. evaluate if the model is consistent with the data y for any values of θ: **model selection/checking**.

There is a further point to be considered when the data-gathering process can be controlled; it concerns the organization of this process in order to take on the preceding question as accurately as possible: **experimental/survey design**.

There are two main classes of methods for answering these questions: the **frequentist and the Bayesian approaches**.

The frequentist approach

- Basic inferential methods, like those taught in this courses, are usually frequentist.
- The model parameters θ are interpreted as fixed states of nature, about which the aim is to learn using the available data y.
- Probability is used to investigate what would happen to the inferential analysis under repeated replication of the data.
- Frequentist methods are usually based on the concept of **likelihood function**

$$L(\theta; y) = f(y_1, \ldots, y_n; \theta), \ \theta \in \Theta$$

Function in the argument θ which gives the "probability" of observing the sample y (that was observed) by considering different values for θ.

Given the observed data y, a suitable "guess" $\widehat{\theta}$ for the unknown parameters θ is the value which maximizes the likelihood function (or its logarithmic transformation called **loglikelihood function**)

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\theta; y)$$

- Likelihood-based procedures provide general solutions, with nice theoretical properties, to the above mentioned inferential problems.
- It is possible to derive suitable (**sample**) **statistics** (summaries of the r.v.'s in the sample) for making inference on θ.
- For example, if $y = (y_1, \ldots, y_n)$ are observations from i.i.d. $N(\mu, \sigma^2)$ r.v.'s, the likelihood-based sample statistics for μ and σ^2 are the sample mean and the (uncorrected) sample variance

$$\widehat{\mu} = \bar{Y}, \quad \widehat{\sigma}^2 = S^2(n-1)/n$$

- A further rather general inferential approach is the **least square method**, which is particularly relevant for regression models.
- A simple approach relies on the **method of moments** where, broadly speaking, the sample statistics are defined by considering the sample versions of the interest parameters, whenever available.

  For example, the sample mean is considered for making inference on the population mean, the sample median for the population median, the (corrected) sample variance for the population variance, etc.

Table of contents

Estimators and standard errors

- Sample statistics, used to estimate a certain parameter θ, are generally called **estimators**, and their value computed using the observed data y is just called the **point estimates** of θ and specified as $\widehat{\theta} = \widehat{\theta}(y)$.
- Since an estimator is a summary of the random sample Y, it is itself a r.v. (if θ is scalar), which is also denoted as $\widehat{\theta} = \widehat{\theta}(Y)$ (distinction will be clarified by the context).
- Every estimator follows a sampling distribution, which describes the values assumed by it across (hypothetical) repeated random samples; if not explicitly known, the sampling distributions can be assessed by simulation.
- Two theoretical properties are desirable for an estimator $\widehat{\theta}$:
  - **unbiasedness**: $E(\widehat{\theta}) = \theta$ or at least $|E(\widehat{\theta}) - \theta|$ should be small;
  - **low variance**: $V(\widehat{\theta})$ should be small.

  There is a trade-off between the two properties, so it is usual to seek both.

- Under this respect, a suitable measure of the estimation error is the **mean square error** (MSE)

$$\text{MSE} = E\{(\widehat{\theta} - \theta)^2\} = V(\widehat{\theta}) + |E(\widehat{\theta}) - \theta|^2$$

  The square root of MSE yields the **standard error** $\text{SE} = \sqrt{\text{MSE}}$, which is a measure of the estimation accuracy having the same unit of measurement as the quantity being estimated.
- A **parameter estimate** should always be accompanied by its **estimated standard error**, obtained by substituting θ with $\widehat{\theta}$ in SE.
- If the estimator $\widehat{\theta}$ is unbiased, $\text{MSE} = V(\widehat{\theta})$ and then $\text{SE} = \sqrt{V(\widehat{\theta})}$, namely the *standard deviation of the sampling distribution of* $\widehat{\theta}$.
- It is quite common to look for minimum variance unbiased estimators.
- A further relevant property is **consistency**: $\widehat{\theta} \xrightarrow{p} \theta$, as $n \to +\infty$.
- Under suitable assumptions, the **maximum likelihood estimators** (MLE's) are (asymptotically) unbiased, consistent and achieves, in the large sample limit, a normal distribution with variance equal to the Cramér-Rao lower bound.

Estimation of the mean

- The **sample mean** $\bar{Y}$ is a particular important instance of estimator, widely used to estimate the population mean μ.
- It has some appealing properties, under i.i.d. assumptions, for any distribution of the variable of interest:
  - **unbiasedness**: $E(\bar{Y}) = \mu$;
  - **consistency**: $\bar{Y} \xrightarrow{p} \mu$, as $n \to +\infty$.
- The central limit theorem ensures that the sampling distribution of $\bar{Y}$ can be **approximated** by the **normal distribution** $N(\mu, \sigma^2/n)$; in case of a sample from a normal distribution, $\bar{Y}$ is exactly normal.
- The standard deviation of the sampling distribution of the sample mean is the **standard error of the mean** (SEM) $\sigma/\sqrt{n}$, whose estimate for random sampling is

$$\text{SEM} = \frac{S}{\sqrt{n}}$$

where S^2 is a suitable estimator for σ^2, usually the (corrected) sample variance.

Estimation of the difference of means

- With two *independent* i.i.d. samples of size n_X and n_Y, the comparison is usually in the form of a **sample difference** $\bar{X} - \bar{Y}$, where $\bar{X}$ and $\bar{Y}$ denote the respective sample means.
- If the corresponding standard errors are SEM_X and SEM_Y, then the **standard error for the difference** (SED) is

$$\mathrm{SED} = \sqrt{\mathrm{SEM}_X^2 + \mathrm{SEM}_Y^2}$$

- In case it is reasonable to assume a common standard deviation σ for the two samples, it can be estimated by

$$\mathrm{SED} = S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

where S_p^2 is an estimator for σ^2 based on both samples, usually the **pooled sample variance**

$$S_p^2 = \frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Example: elastic bands

Data from an experiment on the effect of heat on the amount of `stretch` (mm) of elastic bands: 21 bands were randomly divided into two groups, one of $n_X = 10$ and one of $n_Y = 11$.

Bands in the first group were tested for the amount that they stretched under a weight; the other group was placed in hot water for four minutes and then measured for amount of stretch under the same weight.

Two independent i.i.d. samples X and Y: $\bar{x} = 253.5$, $\bar{y} = 244.1$, $s_X = 9.92$, $s_Y = 11.73$, $\text{SEM}_X = 3.14$, $\text{SEM}_Y = 3.54$.

Since the separate standard deviations are similar, the pooled standard deviation estimate $s_p = 10.91$ is an acceptable summary of the variation in the data.

The mean difference is $\bar{x} - \bar{y} = 9.41$, with a $\text{SED} = 4.77$; therefore, the mean change is positive and it corresponds to $9.41/4.77 = 1.97$ times the estimated standard error.

Estimation of a proportion

- The aim is to estimate the probability of occurrence p of a "success" event in a sequence of n i.i.d. $Ber(p)$ r.v.'s, $Y_1, \ldots, Y_n$; an alternative interpretation is in terms of a realization of a $Bi(n, p)$ r.v.
- An unbiased, consistent estimator for p is the observed proportion of the event of interest in the n trials, which corresponds to the sample mean

$$\widehat{p} = \bar{Y}$$

- The associated standard error is

$$\mathsf{SE} = \sqrt{\frac{p(1-p)}{n}}$$

  which is estimated by substituting p with $\widehat{p}$.
- For example, a random sample of $n = 132$ freshmen is selected in order to evaluate the proportion of freshmen that are displaced from their home.

  Since 37 out of 132 freshmen are displaced: $\widehat{p} = 0.28$, $\mathsf{SE} = 0.039$.

Sampling distribution of z- and t- statistics

- Consider an i.i.d. sample $Y = (Y_1, \ldots, Y_n)$ from a $N(\mu, \sigma^2)$ distribution, the ***z*-statistic (standardized sample mean)** is such that

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

  but this is not directly useful unless σ is known.

  When the component r.v.'s do not follows a normal distribution, the above result holds approximately for large n.

- The ***t*-statistic (studentized sample mean)** is obtained by substituting σ with S and it is such that

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

  Given the observed data y, formula $t = (\bar{y} - \mu)/\mathrm{SEM}$ provides a standardized measure (in number of SEMs) of the distance between the sample mean and the true value μ.

Estimation of the variance

- The (**corrected**) **sample variance** S^2 is widely used as estimator for the population variance σ^2.
- Under i.i.d. assumptions it is **unbiased**, $E(S^2) = \sigma^2$, and **consistent**, $S^2 \xrightarrow{p} \sigma^2$, as $n \to +\infty$.
- In the case of i.i.d. observations from a $N(\mu, \sigma^2)$ distribution, S^2 follows a scaled chi-squared distribution

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- With two *independent* i.i.d. samples of size n_X and n_Y, the comparison concerning the distribution variances is usually in the form of a **sample variance ratio** S_X^2/S_Y^2, where S_X^2 and S_Y^2 are the respective corrected sample variances.

  In case of normal samples from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n_X - 1, n_Y - 1)$$

Table of contents

Introduction to confidence intervals

- **Confidence intervals** (**interval estimates**) provide more satisfactory estimation results than point estimates alone, giving an entire set of values (usually an interval) to estimate the population parameter.
- Interval estimation gives also an implicit idea of the accuracy of the estimation procedure.
- A $(1-\alpha)100\%$ confidence interval for a scalar parameter θ is an *observation of a random interval*, based on a suitable sample statistic and designed to have a prescribed probability $1-\alpha$ (**confidence level**) of including the true value of θ.
- The inferential procedure is expected to specify, over repeated samples, intervals that include the true parameter value with a certain proportion of the times, corresponding to the confidence level.
- The confidence level specify a probability referred to the random interval and not to the observed confidence interval.

Confidence interval for the mean

- Consider an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution with σ^2 *known*; by some algebra, using the z-statistics, it follows that

$$P\left(\bar{Y} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

  with $z_{\alpha/2}$ the $\alpha/2$-critical value of a $N(0,1)$ distribution.

- The random interval

$$\left[\bar{Y} \pm z_{\alpha/2}\,\sigma/\sqrt{n}\right]$$

  is the $(1-\alpha)100\%$-level **confidence interval for** μ; levels commonly used are 90%, 95% and 99% .

- The $(1-\alpha)100\%$-level **observed confidence interval**

$$\left[\bar{y} \pm z_{\alpha/2}\,\sigma/\sqrt{n}\right]$$

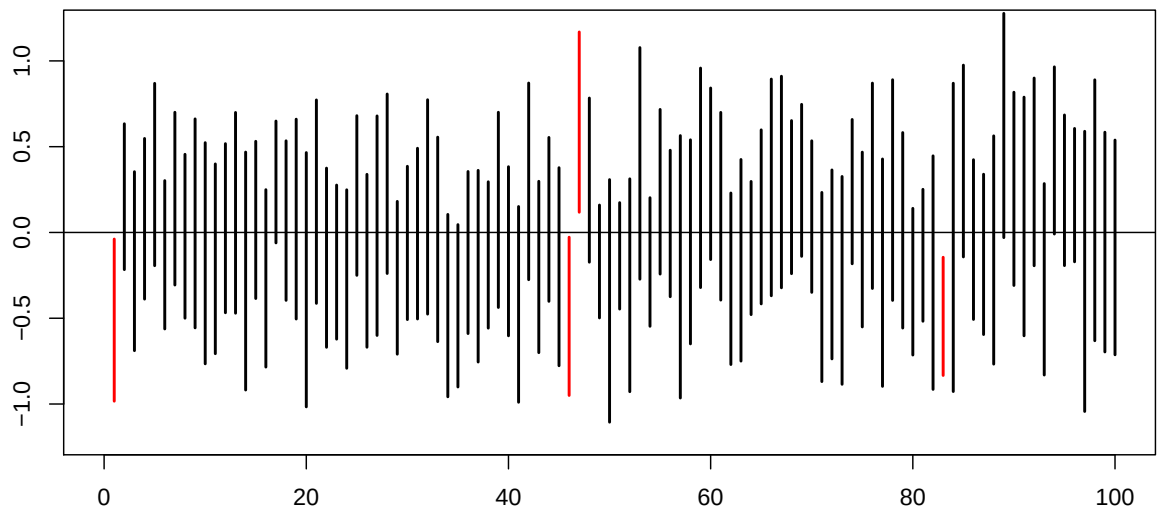
  summarizes the information provided by the observed data on the unknown μ.

- When σ^2 is *unknown*, the $(1-\alpha)100\%$-level **confidence interval for** μ is based on the t-statistics and it corresponds to

$$\left[\bar{Y} \pm t_{n-1;\alpha/2}\, S/\sqrt{n}\right]$$

with $t_{n-1;\alpha/2}$ the $\alpha/2$-critical value of a $t(n-1)$ distribution.

- 95%-level observed confidence intervals based on 100 simulated normal samples of size $n = 15$, with $\mu = 0$ and $\sigma^2 = 1$; four intervals fail to contain the true value



The estimated (by simulations) confidence level is 0.96; increasing the number of simulated samples would get the result closer to 0.95.

Example: cork stoppers

Data from *Applied Statistics Using SPSS, STATISTICA, MATLAB and R* by J.P. Marques de Sá

Data set y with the `total perimeter of the defects` (in pixels) measured in $n = 50$ high quality cork stoppers.

Numerical summaries: $\bar{y} = 365$, $y_{0.5} = 363$, $S^2 = 12167$, $S = 110$, SEM$= S/\sqrt{50} = 15.6$.

Observations interpreted as i.i.d. realizations of a normal distribution.

The 95% confidence interval for μ is

$$[\bar{y} \pm t_{49;0.025}\,\mathrm{SEM}] = [334, 396]$$

with $t_{49;0.025}$ the 0.025-critical value of a $t(49)$ distribution.

The observed confidence interval $[334, 396]$ is obtained using a statistical procedure which is characterized by a risk of 5% of giving wrong results (that is, intervals not containing the true value of μ).

Confidence intervals in general

- In broad generality, $(1-\alpha)100\%$-level confidence intervals for a generic parameter θ have the form

$$[\widehat{\theta} \pm z_{\alpha/2}\,\mathrm{SE}(\widehat{\theta})]$$

  where $\widehat{\theta}$ is an estimator of θ, $\mathrm{SE}(\widehat{\theta})$ is its (estimated) standard error.

- The above formula is valid if $\widehat{\theta}$ is normally distributed; whenever this holds only approximately for large n, the $(1-\alpha)100\%$ confidence level is approximate.

  A relevant example is the confidence interval for a proportion p; indeed, the t-statistic-based confidence level for normal samples is also of the kind above, as $t_{n-1;\alpha/2} \approx z_{\alpha/2}$, for large n.

- For specific models and parameters (variance, ratio of variances, correlation coefficient, etc), there may exist *exact confidence intervals* (exactness refer to the coverage probability) and whenever they exist they represent a better option than approximate intervals.

Table of contents

Introduction to hypothesis testing

- Statistical procedures for hypothesis testing play a fundamental role in statistical inference and they are an essential item in many scientific studies.
- The focus here is mainly on **parametric tests**, which rely on the specification of a parametric statistical model and aim at stating whether a prespecified restriction on the model parameter θ (i.e. a parametric hypothesis) is consistent with data y.
- **Nonparametric tests** are sometimes called distribution-free tests because they are based on fewer assumptions and make no strict assumptions about the probability distributions of the random sample.
- In spite of these differences, the fundamental notions concerning hypothesis testing apply to parametric and to nonparametric tests alike.

Significance level and critical region

- Hypothesis testing is a procedure for validating **a null hypothesis** H_0 made on possible values of θ; the null hypothesis is evaluated against an alternative hypothesis H_1 using the sample data y.
- A fundamental instance corresponds to

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$$

  where the alternative hypothesis is **two-sided**; H_1 could be also defined as **one-sided**: lower, $H_1 : \theta < \theta_0$, or greater $H_1 : \theta > \theta_0$.
- A **test statistic** is a sample statistic, usually (a suitable transformation of) an estimator for the interest parameter θ, such that observed values far from θ_0 (greater or lower) may lead to rejection of the null hypothesis.
- It is essential to derive, at last approximately, the **sampling distribution** of the test statistic **under the null hypothesis**; in standard cases this will be a well-known result, as for example, for the z-statistics or the t-statistics.

- To perform a statistical test it is crucial to select a **significance level** α, common values are 5% and 1%, which is the accepted (by the analyst) probability of the **type I error** (reject H_0 when it is true).
- The knowledge of the distribution of the test statistic under the null hypothesis enables a partition of its possible values into those for which the null hypothesis is rejected (**critical region**) and those for which it is not.
- The critical region is derived so that its probability under H_0 is α, at least approximately.
- Given the observations y, if the observed value of the test statistic is in the critical region, **the null hypothesis** is **rejected**, otherwise it is **accepted** or "not rejected".
- Besides the probability α of the type I error, it is important to evaluate the probability β of the the **type II error** (accept H_0 when it is false); the value of β characterizes the power of statistical tests with a fixed significance level α.

p-value of a test

- The ***p*-value** is the probability, under the null hypothesis, of obtaining a test statistic value, across (hypothetical) repeated random samples, that is at least as extreme (against H_0) as that which was observed.
- The computation of the p-value requires the knowledge of the distribution of the test statistics under H_0, at least approximately, and it depends on the form of the alternative hypothesis (two-sided or one-sided).
- It is a measure of closeness of the data to the null hypothesis, and a small p-value (e.g. < 0.05 or < 0.01) is an evidence against H_0.
- It is common to compare the p-value with the chosen significance level α and to reject H_0, in favor of H_1, if and only if its value is less than α.
- The two testing approaches (critical region and p-value) are equivalent from the decision making perspective (given α, both lead to the same decision), however the second one is usually preferred since it provides a measure of the evidence against or in favor of H_0.

The ASA's statement on p-values

The American Statistical Association (ASA) published a clear guidance (`www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf`) on the proper use and interpretation of the p-value and, more generally, on good statistical and scientific practice.

Four principles (out of six) made by the ASA:

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.

Testing the means

One-sample data

- The purpose is to asses whether the mean of a population (from which the **i.i.d. sample** was collected) has a certain value or not

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

  Here the alternative hypothesis is **two-sided**; H_1 could be also **one-sided**: lower, $H_1 : \mu < \mu_0$, or greater $H_1 : \mu > \mu_0$.

- In case of **normal observations**, the test is called (**one-sample**) ***t* test** since it is based on the ***t*-statistic**, so that, under H_0,

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

- Given a significance level α, the **critical region** for the t test is

$$R_\alpha = \{y : |t| \geq t_{n-1;\alpha/2}\}$$

  that is, H_0 is rejected if the difference, in absolute value, between $\bar{y}$ and μ_0 is at least equal to $t_{n-1;\alpha/2}$ times the SEM.

- If H_1 is one-sided, the critical region is $R_\alpha = \{y : t \leq t_{n-1;\alpha}\}$, lower alternative, or $R_\alpha = \{y : t \geq t_{n-1;\alpha}\}$, greater alternative.
- Given the observed value t of the test statistic T, the ***p*-value** is

$$p = 2\min\{P_{H_0}(T \leq t), P_{H_0}(T \geq t)\} = P_{H_0}\{|T| \geq |t|\}$$

  whereas, for a one-sided lower alternative, $p = P_{H_0}(T \leq t)$ and, for a one-sided greater alternative, $p = P_{H_0}(T \geq t)$.
- When the population **variance** σ^2 is **known** (or the sample size n is sufficiently large), the ***z*-statistic** is considered and the test is called (**one-sample**) ***z* test**; then, under H_0,

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- In the case of **non-normal observations** and a **large sample size**: z test using the *standard error evaluated under* H_0.
- In the case of **non-normal observations** and a **small sample size**: *ad hoc* exact tests.

Example: maximum temperature

Data from *Applied Statistics Using SPSS, STATISTICA, MATLAB and R* by J.P. Marques de Sá

Data set y with `maximum temperature` (°C) registered in 1981 at $n = 25$ weather stations in Portugal.

Numerical summaries: $\bar{y} = 39.8$, $y_{0.5} = 40$, $S = 2.739$, SEM$= 0.548$.

Observations interpreted as i.i.d. realizations of a normal distribution.

A "typical" year has an average maximum temperature of 37.5 °C and then the aim here is to perform a t test with $\alpha = 0.05$ on

$$H_0 : \mu = 37.5 \qquad H_1 : \mu \neq 37.5$$

The reference sampling distribution is $t(24)$ so that $t_{24;0.025} = 2.064$ and $R_{0.05} = \{y : |t| \geq 2.064\}$.

Since the observed value of the test statistic is $t = 4.199$, the null hypothesis is rejected at the level $\alpha = 0.05$ of significance.

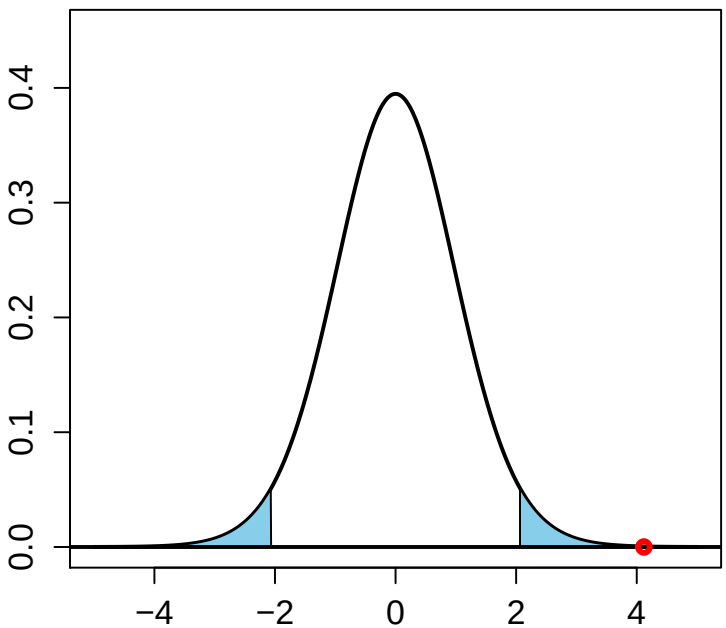
The p-value leads once again to the rejection of H_0:

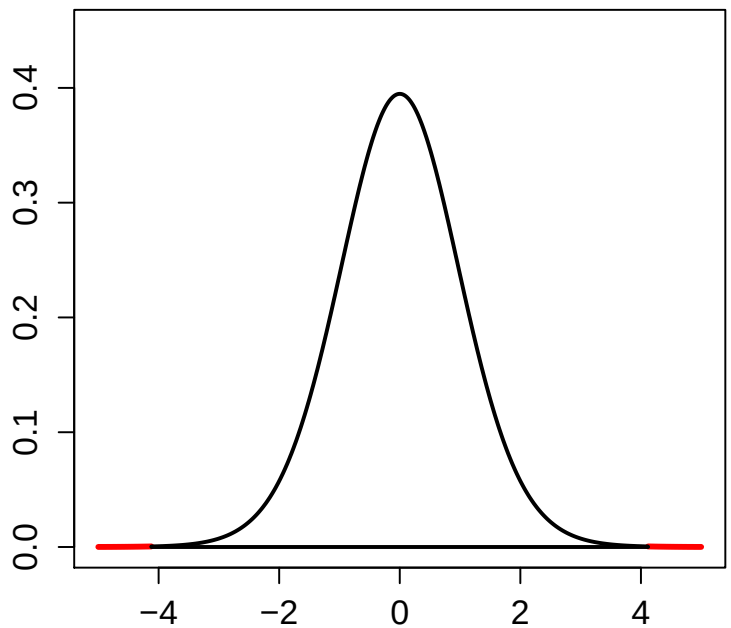
$$p = P_{H_0}\{|T| \geq |4.199|\} = P_{H_0}\{T \leq -4.199 \text{ or } T \geq 4.199\} = 0.0003$$

The 95% confidence interval for μ is $[\bar{y} \pm t_{24;0.025}\text{SEM}] = [38.67, 40.93]$, which does not contain the null value $\mu_0 = 37.5$.

Left panel: observed value of the test statistic and critical region, giving an overall area equal to the level $\alpha = 0.05$ of significance.

Right panel: p-value corresponding to the tails area.





Different ways to report results

Confidence intervals and, to a lesser extent, hypothesis testing provide a way to report summary results in a more interpretative way. For instance, for the `maximum temperature` data, the following statements may be considered.

1. The mean maximum temperature is 39.8, with $\mathrm{SEM} = 0.548$ and $n = 25$.
2. The observed value of the t-statistic is $t = 4.199$, on 24 d.f., namely the difference $\bar{y} - \mu_0$ is 4.199 times the standard error.
3. A 95% confidence interval for the mean is $[38.67, 40.93]$.
4. The null hypothesis, that the true mean maximum temperature is 37.5, is rejected (p=0.0003).

Alternatives 3. and 4. state differently and interpret the information in 1. and 2.; alternative 3. is probably the most informative.

Example: physical activity

The percent of adults ($\geq$18 years old) in US who met the guidelines for `aerobic physical activity` in 2014 is 49.2% (National Health Interview Survey, US, 2014).

A city's council wants to know if the proportion in their city is different from 49.2%: random sample of $n = 200$ adults, 108 meet the guidelines.

Observations interpreted as i.i.d. realizations of a $Ber(p)$ distribution; the aim here is to perform a test with $\alpha = 0.05$ on

$$H_0 : p = 0.492 \qquad H_1 : p \neq 0.492$$

Since p is the mean of a Bernoulli r.v. and the sample size is large: z test statistic, which is approximately normal distributed.

$\widehat{p} = 0.54$, estimated SE $= \sqrt{\widehat{p}(1-\widehat{p})/200} = 0.035$ and, considering the SE under H_0, $z = (\widehat{p} - 0.492)/\sqrt{0.492(1-0.492)/200} = 1.358$.

The (approximate) p-value is $p = P_{H_0}\{|Z| \geq |1.358|\} = 0.175$; the null hypothesis is not rejected: there is not sufficient evidence to state that the proportion of citizens meeting the guidelines is different from 0.492.

Testing the means

Two-sample data

- For testing the equality of the mean of **two independent i.i.d. normal samples**, in case of (unknown) **equal variances**, a well-known test is the (**two-sample**) ***t* test** based on

$$T = \frac{\bar{X} - \bar{Y}}{\mathrm{SED}}$$

  where $\mathrm{SED} = S_p \sqrt{n_X^{-1} + n_Y^{-1}}$, with S_p^2 the pooled estimate of σ^2.
- The computation of the critical regions and of the p-values consider that, under H_0, the test statistic T follows a $t(n_X + n_Y - 2)$ distribution; there are general formulas for large samples, employing the normal distribution.
- If variances are heterogeneous (i.e. **unequal variances**), the t-statistic based on the pooled variance estimate is inappropriate.

- In this case, the **Welch test**, based on $T = (\bar{X} - \bar{Y})/\text{SED}$, with $\text{SED} = \sqrt{\text{SEM}_X^2 + \text{SEM}_Y^2}$, gives an adequate approximate solution; under H_0, T has a t distribution with suitable degrees of freedom.
- **Paired data** arise when *the same units are measured under two different conditions*, generating two different observations x_i and y_i for each unit, $i = 1, \ldots, n$.

  In this case, the methods for two independent samples cannot be used; a simple solution consists in using one-sample tests applied to the individual differences $d_i = x_i - y_i$ (e.g. ***t* test for paired data**).
- In the case of two **non-normal independent samples** and **large sample sizes**: z test based on $\bar{X} - \bar{Y}$, using the SED *evaluated under* H_0.
- In the case of two **non-normal independent samples** and a **small sample size**: *ad hoc* exact tests.
- Specific tests may be considered for two **non-normal dependent samples**; a well-known large sample test for Bernoulli observations is the **McNemar's test**.

Example: white and red wines

Data from *Applied Statistics Using SPSS, STATISTICA, MATLAB and R* by J.P. Marques de Sá

Data x and y, with $n_X = 30$ and $n_Y = 37$, correspond to the `aspartame content` (mg/l) in two independent samples of white and red wines.

Summaries: $\bar{x} - \bar{y} = 27.06 - 20.86 = 6.203$, $s_X = 10.51$, $s_Y = 10.97$.

Observations interpreted as independent i.i.d. realizations of two normal distributions; the variances are considered as equal, although a formal statistical test would be required.

The point is whether the mean `aspartame content` can distinguish white wines from red wines: two-sample t test with $\alpha = 0.05$ on

$$H_0 : \mu_X - \mu_Y = 0 \qquad H_1 : \mu_X - \mu_Y \neq 0$$

Since $\text{SED} = 2.645$, the observed value of the test statistic is $t = 2.345$ and the p-value is 0.022.

Note that for $\alpha = 0.05$ the null hypothesis is rejected, but not for lower values of the significance level, such as $\alpha = 0.01$.

Example: temperatures

Data from *Applied Statistics Using SPSS, STATISTICA, MATLAB and R* by J.P. Marques de Sá

Data x and y correspond to the `maximum temperature` (°C) registered in 1980 and in 1981 at $n = 25$ weather stations in Portugal.

The objective is to compare the maximum temperatures in year 1980 with those of year 1981; since the measurements are performed in the same stations, the data sets x and y define pair data.

Numerical summaries: $\bar{x} = 37.44$, $\bar{y} = 39.80$, $s_X = 2.20$, $s_Y = 2.739$; with regard to the differences $d = x - y$, $\bar{d} = -2.36$, $\text{SEM}_D = 0.412$.

Observations interpreted as two dependent normal i.i.d. samples: t test for paired data on

$$H_0 : \mu_X - \mu_Y = 0 \qquad H_1 : \mu_X - \mu_Y \neq 0$$

The observed value of the test statistic is $t = -5.731$ and the p-value is $6.632 \cdot 10^{-6}$: the null hypothesis is rejected; strong evidence on the fact that the mean maximum temperature in 1981 exceeds that one in 1980.

Example: labor training program

Two groups of individuals are randomly selected: $n_X = 297$ who had participated in labor training programs and $n_Y = 128$ who had not.

Evaluate if the proportion of high school dropouts is the same in the two reference populations: 217 and 65 dropouts observed, respectively.

Observations viewed as independent i.i.d. samples from a $Ber(p_X)$ and a $Ber(p_Y)$ distribution, respectively; the objective is to perform a test on

$$H_0 : p_X - p_Y = 0 \qquad H_1 : p_X - p_Y \neq 0$$

thus, to assess whether the probability distribution of the dichotomous random variable `school dropout` is the same in the two populations.

As the sample sizes are large: z test based on $Z = (\widehat{p}_X - \widehat{p}_Y)/\text{SED}$, where $\text{SED} = \sqrt{\widehat{p}(1-\widehat{p})(n_X^{-1} + n_Y^{-1})}$, with $\widehat{p}$ is the estimated proportion based on the pooled sample.

Since $\widehat{p}_X = 0.731$, $\widehat{p}_Y = 0.508$ and $\widehat{p} = 0.664$, then $z = 4.46$ and the (approximate) p-value is $8.189 \cdot 10^{-6}$, leading to H_0 rejection.

Testing the medians

Nonparametric tests

- Although the (one-sample) t tests are fairly robust against departures from the normal distribution, especially in larger samples, a nonparametric alternative is the **Wilcoxon signed-rank test**.

  It assumes only that the distribution is continuous and symmetric and the hypotheses concern the median instead of the mean.

- The Wilcoxon signed-rank can be considered as a nonparametric test for comparing the median of two **paired samples**, when the population cannot be assumed to be normally distributed.

  It is the same as the previous test, applied on the differences of the paired observations.

- A nonparametric alternative to the (two-sample) t test for independent i.i.d. samples is the **Wilcoxon rank-sum test**.

  It assumes only that the distributions are continuous and the hypotheses usually concern the difference in medians; it is equivalent to the **Mann-Whitney *U* test**.

Testing the variances

One-sample and two-samples

- *One normal i.i.d. sample*: the aim is to asses whether the variance of a population has a certain value or not

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

The test statistic is $(n-1)S^2/\sigma_0^2$, with null distribution $\chi^2(n-1)$.

- *Two independent normal i.i.d. samples*: the purpose is to decide whether the two reference populations have the same variance

$$H_0 : \sigma_X^2/\sigma_Y^2 = 1 \qquad H_1 : \sigma_X^2/\sigma_Y^2 \neq 1$$

A well-known test is called ***F* test** for the homogeneity (equality) of variance (homoscedasticity) and it is based the ratio of the two (corrected) sample variances S_X^2/S_Y^2, which under H_0 follows an $F(n_X - 1, n_Y - 1)$ distribution, since $\sigma_X^2 = \sigma_Y^2$.

- The F test is extremely sensitive to non-normality; the **Levene's test** or the **Bartlett's test** are more robust alternatives.

Example: white and red wines

In order to decide whether the mean `aspartame content` can distinguish white wines from red wines, a two-sample t test was considered, assuming equality of the population variances.

However, to evaluate the homogeneity of variance, an F test with $\alpha = 0.05$ can be performed by considering the hypotheses

$$H_0 : \sigma_X^2/\sigma_Y^2 = 1 \qquad H_1 : \sigma_X^2/\sigma_Y^2 \neq 1$$

Since $s_X^2 = 110.43$, $s_Y^2 = 120.34$ and $n_X = 30$, $n_Y = 37$, the observed value of the test statistic is $s_X^2/s_Y^2 = 0.9178$ and, using the $F(29, 36)$ null distribution, the p-value corresponds to

$$p = 2 \min\{P_{H_0}(F \leq 0.9178), P_{H_0}(F \geq 0.9178)\} = 0.819$$

The hypothesis of homoscedasticity is accepted, having a strong support from the observed data.

Correlation test

- Given a random sample from the bivariate r.v. (X, Y), it might be of some interest to test whether X and Y are correlated; namely

$$H_0 : \rho_{XY} = 0 \qquad H_1 : \rho_{XY} \neq 0$$

  It should be emphasized that a correlation between two variables does not necessarily imply that one causes the other (causation).

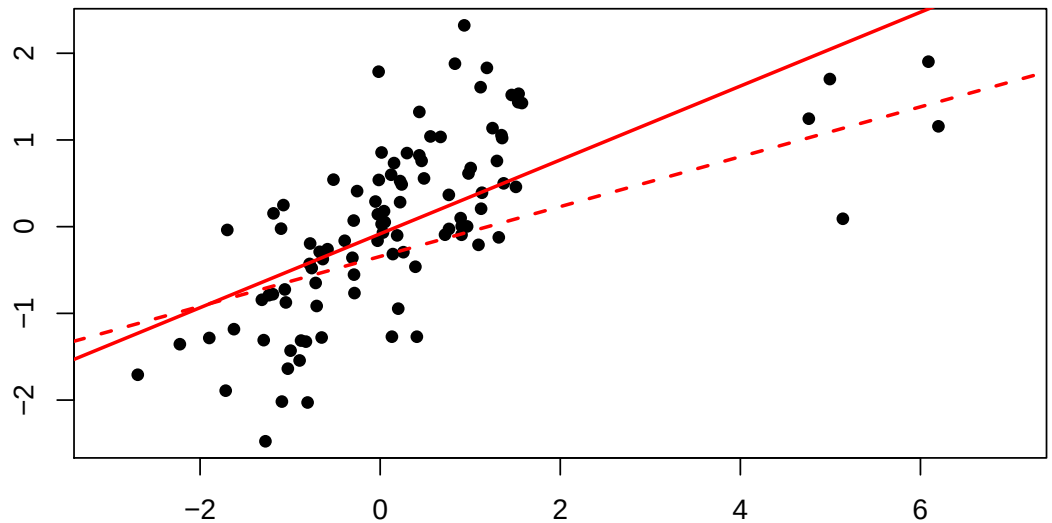
- A **correlation test** can be obtained under the assumption that the random sample derives from a *bivariate normal distribution* (it may be enough to check that both X and Y are normally distributed).
- A well-known test is based on the sample version of the **Pearson correlation coefficient**

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  The test statistics is obtained by transforming r_{XY} so that, under H_0, it follows a suitable t distribution.

- There are some nonparametric variants of the Pearson correlation test, having the advantage of not depending on the normal distribution; however, their interpretation may be not quite clear.
- A popular test is based on the **Spearman's rank correlation coefficient**, which is obtained by replacing the observations by their rank and computing the Pearson correlation coefficient.
- A further nonparametric test involves the **Kendall's τ coefficient**, which is based on counting the number of concordant and discordant pairs (two pairs are concordant if the difference in the x-coordinate is of the same sign as the difference in the y-coordinate).
- The null distribution of the above mentioned test statistics may be calculated exactly for small samples; for larger samples, it is common to use suitable approximations.
- The null distributions and the p-values, and even the confidence intervals for ρ_{XY}, can be computed using suitable simulation-based approximate methods.

The figure below shows that the five outliers in the upper right corner change the least square line



The correlation tests based on the Pearson and on the Spearman correlation coefficient both suggest rejection of the null hypothesis; the rank correlation is more robust, though the p-values are always very small.

Full sample: sample correlation 0.64 (p-value $< 10^{-12}$), 95% confidence interval $[0.51, 0.75]$; rank correlation 0.75 (p-value $< 10^{-16}$).

Without outliers: sample correlation 0.73 (p-value $< 10^{-16}$), 95% confidence interval $[0.61, 0.81]$; rank correlation 0.73 (p-value $< 10^{-16}$).

Table of contents

Model checking

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

- The aim of model checking is to verify if the assumed statistical model is seriously wrong, taking into account that "all models are wrong but some are useful." (George Box)
- "The type of statistical inference used may be less important to the conclusions than choosing a suitable model or models in the first place." (The GAMLSS team)
- Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably." (OxfordDictionaries.com)

  Although an underfitted model is clearly not adequate, an overfitted model contains too many parameters, since it tends to improperly model also the residual random variation observed in a data set.
- "Entities should not be multiplied beyond necessity". (Occam's Razor)

- Although "all models are wrong, but some are useful", if a model is wrong statistically, then the statistical conclusions drawn from it are usually unreliable and questionable.
- **Graphical checks**; when the model is wrong, they frequently indicate how it is wrong: **quantile-quantile plot**, plots of the **standardized residuals** (useful for regression-type models).
- **Parametric tests** may be viewed as a procedure for comparing two alternative *nested* models: the null hypothesis defines a simplified (restricted) version of the statistical model taken into account.
- The **goodness of fit tests** are statistical tests which considered as null hypothesis the compatibility of the observed sample with a parametric statistical model: e.g. **Kolmogorov-Smirnov test**, **chi-square goodness of fit test**, **Shapiro-Wilk test for normality**.

Akaike's information criterion

- The **Akaike's information criterion** (**AIC**) enables model comparison, seeking the "best" model from a set of (two or more) models, which need not necessarily be nested.
- A (theoretical) measure of goodness of fit for discriminating among alternative models is the **Kullback-Leibler divergence**, which, in the continuous case, is

$$K(\widehat{f}, f_T) = \int \{\log f_T(y) - \log f(y; \widehat{\theta})\} f_T(y) dy$$

where f_T is the true density of Y and $\widehat{f} = f(y; \widehat{\theta})$ is the estimated density under the assumed model.

- According to the AIC, *the selected model has the lowest value of*

$$\text{AIC} = -2\ell(\widehat{\theta}; y) + 2\text{dim}(\theta)$$

with $\ell(\widehat{\theta}; y) = \log f(y; \widehat{\theta})$ the maximized loglikelihood based on data y and $\dim(\theta)$ the dimension of the unknown model parameter θ.

- The AIC is an estimate, based on the observed sample y, of the expected value of $K(\widehat{f}, f_T)$, with respect to the distribution of the estimator $\widehat{\theta}$; it requires the assumption that the model is the true one.
- It specifies a *trade-off* between the goodness of fit of the model (measured by the maximized loglikelihood) and the complexity of the model (described by the dimension of θ).
- An alternative criterion recognizes that $K(\widehat{f}, f_T)$ depends on the model only via $-\int \log f(y;\widehat{\theta})\}f_T(y)dy$; a *cross-validation* estimate for this quantity is

$$\text{CV} = -\sum_{i=1}^{n} \log f(y_i;\widehat{\theta}_{-i})$$

  where $\widehat{\theta}_{-i}$ is the MLE based on the data y with y_i omitted.
- The **cross validation score criterion** points to the model which minimizes CV and measures the average ability to predict data to which it was not fitted.

Bayesian information criterion

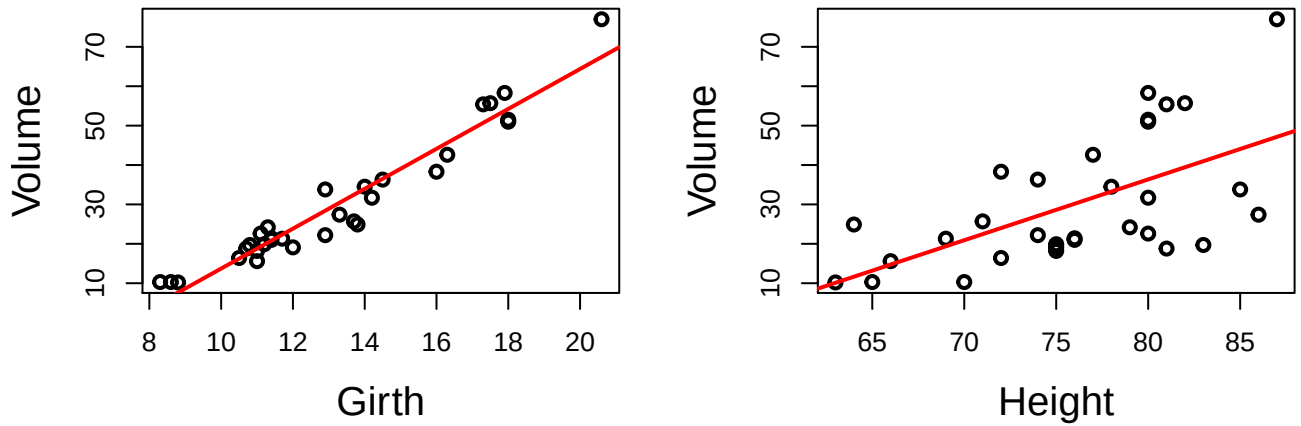
- An obvious Bayesian approach is to consider all possible models, and then to compute the marginal posterior probability for each model; sensitivity to the priors put on the model parameters.
- In the Bayesian framework the goal of summarizing the evidence for or against two alternative models can be achieved by the **Bayes factor**, which unavoidably depends on the choice of the prior.
- A Bayesian criterion, similar to the AIC, is the **Bayesian information criterion** (**BIC**) which selects the model minimizing

$$\text{BIC} = -2\ell(\widehat{\theta}; y) + \log n \, \text{dim}(\theta)$$

- The BIC does not consider the prior and it is obtained by using a suitable approximation for the marginal likelihood $\int f(\theta) f(y|\theta) d\theta$ and by dropping the prior, in a somewhat artificial way.
- The term penalizing model complexity is larger in BIC than in AIC; beyond the Bayesian interpretation, the BIC (as the AIC) belongs to the class of information (or predictive) criteria for model selection.

Example: black cherry trees

Data set with measurements of the `girth` (diameter of the tree, in inches, measured at a fixed distance above the ground), `height` (ft) and `volume of timber` (cubic ft) in $n = 31$ felled black cherry trees



Two linear regression models (the residuals ε are i.i.d. $N(0, \sigma^2)$ r.v.'s):

$$\text{Model 1:} \quad \texttt{volume} = \alpha + \beta \cdot \texttt{girth} + \varepsilon$$

$$\text{Model 2:} \quad \texttt{volume} = \alpha + \beta_1 \cdot \texttt{girth} + \beta_2 \cdot \texttt{height} + \varepsilon$$

Model 1: $\text{logLik} = -87.82$, $\text{AIC} = 181.64$, $\text{BIC} = 185.95$, $\text{CV} = 92.36$

Model 2: $\text{logLik} = -84.45$, $\text{AIC} = 176.91$, $\text{BIC} = 182.65$, $\text{CV} = 90.62$

Table of contents

Bivariate tables

- Two-way tables, called **contingency tables**, display the observed frequencies associated to bivariate sample data.
- Given an i.i.d. n-dimensional sample from a bivariate r.v. (X, Y), where X has $r > 1$ categories and Y has $c > 1$ categories, the observed counts n_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, related to the combination of categories, can be summarized in

	y_1	y_2	$\ldots$	y_c	
x_1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{1+}
x_2	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r+}
	n_{+1}	n_{+2}	$\ldots$	n_{+c}	n

  where n_{i+} and n_{+j} are, respectively, the row and the column sums.
- The marginal r.v.'s X and Y are categorical with a finite number of possible values; also discrete and continuous r.v.'s can be suitably categorized.

Chi-squared test of independence

- The underlying statistical model is a **multinomial distribution** with $r \times c$ cells with probabilities $p_{ij} = P(X = x_i, Y = y_j)$, $i = 1, \ldots, r$, $j = 1, \ldots, c$, and the number of independent trials is n.
- The multinomial distribution is a generalization of the binomial distribution for experiments with more than two possible outcomes.
- There are several tests for contingency tables, but the most commonly used one is the **chi-squared test of independence**.
- The test is suitable to assess the null hypothesis that paired observations on two variables are independent of each other, that is

$$\begin{aligned} H_0 : &\quad p_{ij} = p_{i+}p_{+j}, \text{ for each } (i,j) \\ H_1 : &\quad p_{ij} \neq p_{i+}p_{+j}, \text{ for at least one } (i,j) \end{aligned}$$

with $p_{i+} = \sum_{j=1}^{c} p_{ij} = P(X = x_i)$, $p_{+j} = \sum_{i=1}^{r} p_{ij} = P(Y = y_j)$.

An equivalent expression is $H_0 : p_{ij}/p_{+j} = p_{i+}$ for each (i,j) (so that $p_{ij}/p_{i+} = p_{+j}$): namely, $P(X = x_i|Y = y_j) = P(X = x_i)$ (and $P(Y = y_j|X = x_i) = P(Y = y_j)$).

- Under H_1, the cell probabilities p_{ij} are estimated by the **observed proportions**

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}$$

whereas, under H_0, they are estimated by the **expected proportions under independence**

$$\widehat{p}^0_{ij} = \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{e_{ij}}{n}$$

- The chi-squared statistic compares the observed n_{ij} with the expected (under independence) $e_{ij} = n_{i+}n_{+j}/n$, and computes the score

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, for **large tables**, the test statistic X^2 follows, approximately, the chi-squared distribution $\chi^2((r-1)\cdot(c-1))$.

  Large values of X^2 point against H_0, so that, given the observed value x^2 of X^2, the (approximate) p-value is $p = P_{H_0}(X^2 \geq x^2)$.

- The chi-squared distribution is a good approximation for the true distribution of X^2, provided that $e_{ij} \geq 5$ for all the cells.

  For **sparse tables**, the **exact Fisher test** can be considered; alternatively, the p-value can be obtained by simulation, generating many samples under H_0 and computing the distribution of X^2.
- The requisite of random sampling is really important, and lack of thereof may invalidate the procedure; nonrandom sampling may occur in case of **repeated observation** of the same units over time, **clustering** and **self-selection** of individuals.
- In case of small p-value, it might be useful to investigate the reasons that lead to it by analyzing the (**Pearson**) **standardized residuals**

$$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

  In large tables, under H_0, such quantities are roughly standard normally distributed, so residuals with absolute value larger than 2 point to departure from the independence hypothesis.

Example: steel rods

Four machines produce steel rods, whose diameter can be not defective (`ok`), too short (`short`) or too long (`long`).

A sample of $n = 500$ steel rods is randomly selected and two categorical variables, `type of machine` and `diameter`, are observed; the available sample counts n_{ij} form the two-way table

type of machine	diameter			
	short	ok	long	Total
machine 1	10 (15.84)	102 (96.48)	8 (7.68)	120
machine 2	34 (26.40)	161 (160.80)	5 (12.80)	200
machine 3	12 (13.20)	79 (80.40)	9 (6.40)	100
machine 4	10 (10.56)	60 (64.32)	10 (5.12)	80
Total	66	402	32	500

where in parenthesis are the expected observations e_{ij} under independence.

The presence of dependence between `type of machine` and `diameter` implies that the proportions of rod types produced by different machines are different.

The observed value of X^2 is 15.58, with p-value 0.016, giving a moderate evidence against the independence hypothesis.

The standardized residuals are reported below

type of machine	diameter		
	short	ok	long
machine 1	-1.4673552	0.56197944	0.1154701
machine 2	1.4791480	0.01577201	-2.1801663
machine 3	-0.3302891	-0.15613491	1.0277402
machine 4	-0.1723281	-0.53865504	2.1566757

There are two large residuals in the cells with coordinates (2,3) and (4,3).

The evidence is that the proportion of rods with diameter too long is too low for machine 2, whereas the proportion of rods with diameter too long is too high for machine 4.

Comparing multinomial populations

- The $r \times c$ contingency table and the chi-squared test is a convenient formalism also whenever there are r i.i.d. independent samples obtained from r distinct multinomial populations.
- In this case the rows represent different populations and report the observed frequencies of the c categories of the interest r.v. Y.
- The row totals n_{i+}, $i = 1, \ldots, r$, are assumed to be fixed and correspond to the dimension of the r samples, so that $n = \sum_{i=1}^{r} n_{i+}$.
- The **chi-squared test** X^2 may be considered as-well for assessing whether the r populations follows the same multinomial distribution

$$
\begin{aligned}
H_0 : &\quad p_{1j} = p_{2j} = \cdots = p_{rj}, \text{ for each } j = 1, \ldots, c \\
H_1 : &\quad \exists\, i, j \text{ such that } p_{ij} \neq p_{kj} \text{ for a category } j
\end{aligned}
$$

- For a 2×2 contingency table, the chi-squared test compares two independent Bernoulli populations and it is equivalent to the test assessing whether the "success" probabilities are reasonably the same.

Example: labor training program

Data on the high school dropouts: two samples of, respectively, 128 individuals who had participated in labor training programs and 297 individuals who had not are summarized in a 2×2 contingency table

	high school graduate		
program	yes	no	Total
yes	63 (43.07)	65 (84.93)	128
no	80 (99.93)	217 (197.07)	297
Total	143	282	425

In parenthesis are the expected observations e_{ij} under the null hypothesis that the proportion of dropouts is the same in the two populations.

The observed value of the test statistic X^2 is 19.893 and, since under H_0 it is approximately $\chi(1)$ distributed, the p-value is lower than 0.00001, giving substantial evidence against the null hypothesis.

The conclusion is in accordance with that of the z test based on the difference between the observed proportions (the value of X^2 is the square of that of Z), thought it does not reflect the sign of the difference.

Concerning statistical reasoning

- "No matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong". (Richard Feynman)

- "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."
(John M. Tukey)