Contents

1 Inferential statistics (Recap)

The concept of statistical model

Aim of statistical inference

Statistics aims to **extract information from data**, and in particular on the process that generated the data.

Two intrinsic difficulties:

- It may be hard to infer what we wish to know from the data available;
- Most data contain some **random variability**: by replicating the data-gathering process several times we would obtain different data on each occasion.
  - We search for conclusions drawn from a single data set that are **generally valid**, and not the result of random peculiarities of that data set.

1.0.0.1 Role of statistical models

Statistics is able to draw conclusions from random data mainly though the use of **statistical models**.

A statistical model can be thought as a *mathematical cartoon* describing how our data might have been generated, if the unknown features of the data-generating process were actually known.

If the unknowns were known, a good model *can generate data resembling* the main features of *observed data*.

The purpose of **statistical inference** is to use the statistical model to go in the *reverse direction*: to infer the model unknowns that are consistent with the observed data.

Mathematical aspects

Notation:

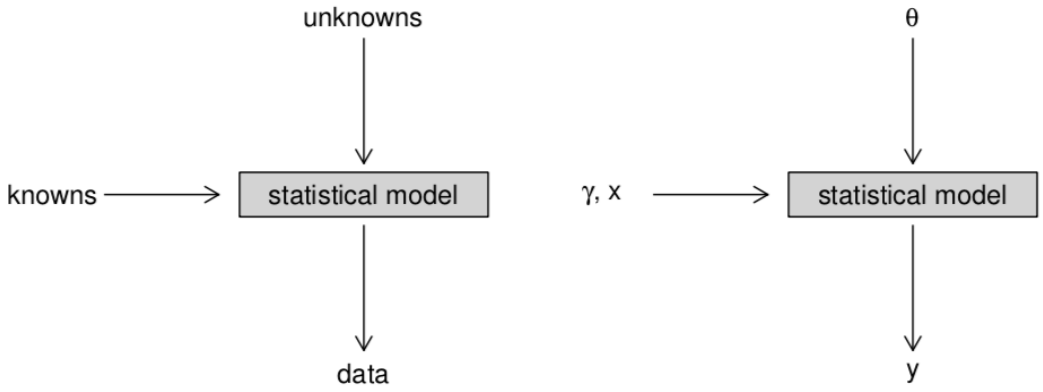
- $\mathbf{y}$ random vector containing the observed data
- $\boldsymbol{\theta}$ vector of parameters of unknown value

We assume that knowing the parameters would answer the question of interest about the process generating the data.

The model specifies how data akin to **y** may be simulated, implicitly defining the **distribution of y** and how it depends on θ.

Moreover, a statistical model may depend on some known parameters γ and some further data **x**, treated as known and denoted as *covariates* or *predictor variables*.

Visually

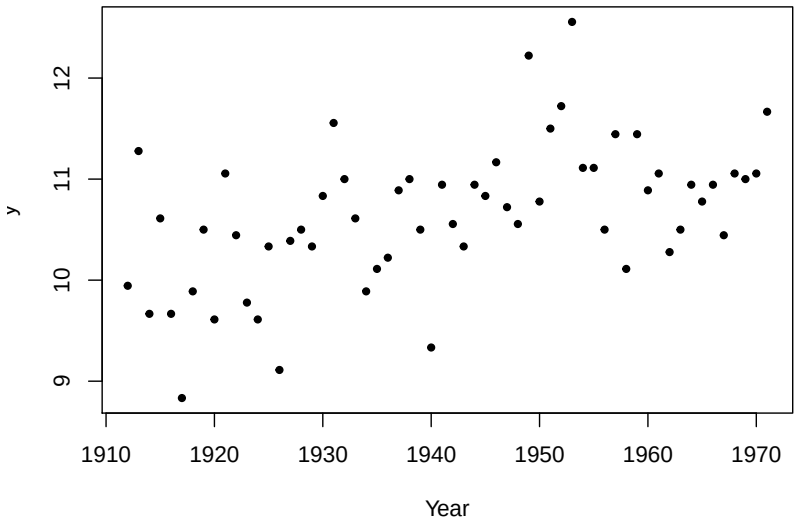


(Taken from Wood's book, page 20)

An example

Consider the following record of 60 mean annual temperatures in New Haven, expressed in Celsius degrees

```
y <- (nhtemp - 32) / 1.8
plot(1912:1971, y, pch = 20, xlab = "Year", ylab = "y")
```



Example: model 1

A first model simply assumes that the data are a random sample from a normal distribution namely they are the observation of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

It follows that the distribution for the entire data vector is the product of the single contributions

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma}\phi\{(y_i\mu)/\sigma\},$$

where ϕ is the $\mathcal{N}(0, 1)$ p.d.f.

Example: model 2

A second model retains the random sample assumption, but replaces the normal distribution with a heavier-tailed t5 distribution, assuming

$$\frac{Y_i - \mu}{\sigma} \sim t_5.$$

The distribution of the data becomes

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma} f_{t_5}\{(y_i\mu)/\sigma\},$$

where f_{t_5} is the t_5 p.d.f.

Example: model 3

The third model relaxes the assumption of identical distribution, assuming a linear trend over time: after setting $t_i = year_i 1912$, $i = 1, \ldots, 60$; we then take

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The independence between observations still holds, so that

$$f(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sigma}\phi\{(y_i\beta_0 - \beta_1 t_i)/\sigma\}.$$

Example: model 4

The last model maintains the trend assumption, but also includes autocorrelation for the error term, meaning that we state

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i \quad \epsilon_i = \rho\epsilon_{i-1} + \nu_i$$

with $\nu_i \sim \mathcal{N}(0, \sigma^2)$ and the autocorrelation $\rho \in (0, 1)$.

The model also requires to specify the distribution for the first observation, here taken as $Y_1 \sim \mathcal{N}(\beta_0, \sigma^2/(1\rho^2))$, so that all the variables in the sample have the same variance.

The model is an instance of a **linear regression model with autocorrelated errors**. The r.v. of the sample are not longer independent, yet the distribution of $\mathbf{Y}$ can be found with some algebra.

Example: model 4 (cont.d)

It is possible to verify that $\mathcal{Y}$ is multivariate normal, with mean vector given by the linear trend

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 t_i,$$

and covariance matrix

$$\Sigma = \frac{\sigma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{pmatrix}$$

so that $f(\mathbf{y}) = \phi_n(\mathbf{y}; \mu, \Sigma)$, being ϕ_n the multivariate normal p.d.f.

Example: model parameters

It is useful to write down the vector parameters θ for each of the four model specifications proposed:

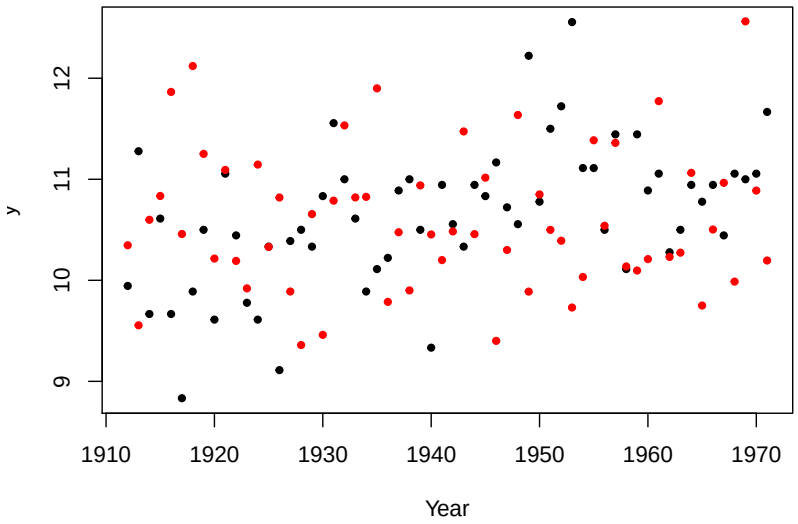
- Model 1: $\theta = (\mu, \sigma^2)$
- Model 2: $\theta = (\mu, \sigma^2)$
- Model 3: $\theta = (\beta_0, \beta_1, \sigma^2)$
- Model 4: $\theta = (\beta_0, \beta_1, \rho, \sigma^2)$

Mote that the meaning of each parameter depends on the chosen model: $\sigma^2 = var(Y_i)$ in Model 1, but $\sigma^2 = 0.6var(Y_i)$ in Model 2.

Example: simulation from model 1

For model 1, the parameters μ and σ^2 are readily estimated by $\bar{y}$ and s^2. A dataset can then be simulated using such values

```
set.seed(2018); ysim <- rnorm(length(y), m = mean(y), s = sd(y))
plot(1912:1971, y, pch = 20, xlab = "Year", ylab = "y")
points(1912:1971, ysim, col = 2, pch = 20)
```



Example: model 1 checking

In order to evaluate whether the simulated dataset is similar to the observed one, we should focus on some important features.

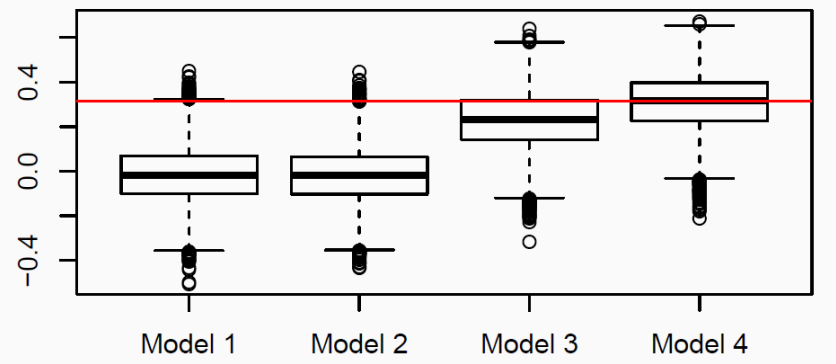


Figure 1: Simulated sample autocorrelation.

For example, climate changes over time may suggest that the temperature of a given year may be positively correlated with the temperature of the subsequent year, an example of *positive autocorrelation*.

We can quantify this point by computing the sample autocorrelation

$$r_1 = \frac{\sum_{i=1}^{n1}(y_i \bar{y})(y_{i+1} \bar{y})}{\sum_{i=1}^{n}(y_i \bar{y})^2},$$

which is computed by the R function acf.

For the original data set $r_1 = 0.31$, whereas for the simulated data from model 1 $r_1 = 0.12$. This is just a single data set, though.

We simulate $10,000$ samples from each of the four models, and each time we compute the r_1 coefficient. The sample distributions obtained are displayed in the plot below.

What model is better at reproducing autocorrelation?

The problems of statistical inference: an overview

Inferential questions

Given a statistical model for data $\mathbf{y}$, with model parameters $\boldsymbol{\theta}$, there are some basic questions to ask (pasted from the CS book):

1. What values of $\boldsymbol{\theta}$ are most consistent with $\mathbf{y}$? [*Point estimation*]
2. What range of values of $\boldsymbol{\theta}$ are consistent with $\mathbf{y}$? [*Interval estimation*]
3. Is some prespecified restriction on $\boldsymbol{\theta}$ consistent with $\mathbf{y}$? [*Hypothesis testing*]
4. Is the model consistent with the data for any values of $\boldsymbol{\theta}$ at all? [*Model checking*]

Question 3 can be enlarged to include

- which of several alternative models is most consistent with $\mathbf{y}$? This is point of *model selection*, which partially overlaps with model checking.

The central issue is the acknowledgment of the intrinsic uncertainty inherent in trying to learn about $\boldsymbol{\theta}$.

For settings where some control over the data-gathering process is possible, a further question arises:

5. How might the data-gathering process be organized to produce data that enables answers to the preceding questions to be as accurate and precise as possible?

This is the core of *experimental and survey design* methods.

An often neglected question, of central importance in many traditional fields where statistics is routinely applied (medical sciences, industrial research, biosciences, etc). It is also very relevant for business and web analytics.

Approaches to statistical inference

There two classes of methods providing an answer to questions 1-4, namely the **frequentist** and **Bayesian** approach.

They differ mainly for the role of model parameters θ, which are treated as fixed constants in the former approach and as r.v. in the latter one.

The difference may appear remarkable, and there has been controversy over the years about the merits of each approach.

Occam's razor

In the previous example no model gives a perfect fit for this data set.

More sophisticated models may give better results, but simple models conform to the **Occam's Razor principle**, that for statistical modelling argues in favor of simple models for simple problems, moving to more complex models when simple models are inappropriate.

2 Parameter estimation (Recap)

2.1 Point estimation

The aim of point estimation

Given a model for the data $\mathbf{y}$, with parameter $\boldsymbol{\theta}$, **point estimation** is concerned with finding a reasonable parameter estimate from the data.

There are several methods for doing this.

Here we focus on some general aspects of point estimation.

Example: sample mean and sample variance

A very simple model assumes that the data are observations of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

Straightforward estimates of μ and σ^2 are given by, respectively, the sample mean

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

and the sample variance

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Such estimates are actually sensible anytime we are interested in estimating the mean and variance of an i.i.d. sample.

Estimation properties

To figure out what could be a good estimate, we need to consider **repeated** *estimation under repeated replication of the data-generating process.*

This makes fully sense whenever the available data are a random sample obtained from a large population, like in industrial or social surveys, so that it would perfectly possible to iterate the sampling and obtain further data with the same structure of $\mathbf{y}$.

However, we apply the same logic even when repetition is just the result of an idealization, like in the case of the temperatures recorded in New Haven of the previous example.

The point is: what do we expect to find if we repeat the same analysis to many data sets generated from the same model?

2.1.0.1 Unbiasedness

If we replicate the random data and we repeat the estimation process, the result will be a different value (*estimate*) of $\hat{\boldsymbol{\theta}}$ for each replicate.

The values are observations of a random vector, the **estimator of $\boldsymbol{\theta}$**, which is usually also denoted by $\hat{\boldsymbol{\theta}}$ (the context will make clear whether we are referring to the estimator or to the estimate for a given sample).

Since, the estimator is a r.v., it makes fully sense to compute its mean.

For an **unbiased** estimator

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}.$$

Unbiasedness is a desirable property, and we would also like the estimator to have **low variance**.

2.1.0.2 Mean Squared Error

There is tradeoff between unbiasedness and low variance, so we usually seek to get both (to some extent): ideally we would target a small **Mean Squared Error** (MSE)

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\}.$$

With some algebra, we obtain

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\theta}}) &= \{E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\}^2 + var(\hat{\boldsymbol{\theta}}) \\ &= \text{Squared bias} + \text{Variance}.\end{aligned}$$

Example: normal random sample

For a normal random sample, it is straightforward to verify that

$$E(\bar{\mathbf{Y}}) = \mu \qquad var(\bar{\mathbf{Y}}) = \frac{\sigma^2}{n} = \text{MSE}(\bar{\mathbf{Y}}).$$

For the sample variance, we use the property that

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

to obtain

$$E(S^2) = \sigma^2 \qquad var(S^2) = \frac{2(n-1)\sigma^4}{n^2} = \text{MSE}(S^2).$$

The unbiasedness of the sample mean and variance is a general property, holding also for non-normal samples.

2.1.0.3 Consistency

A (scalar) estimator is said to be **(weakly) consistent** if, for any $\epsilon > 0$

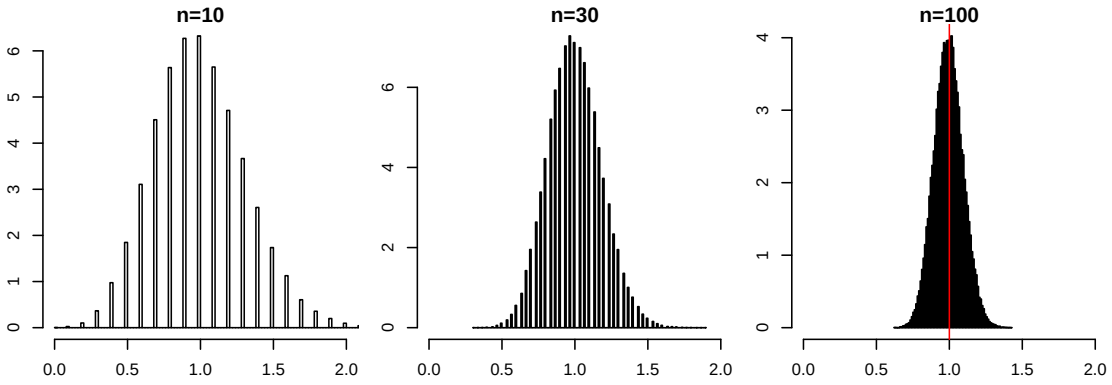
$$Pr(|\hat{\theta} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty.$$

A sufficient condition for this is that $\text{MSE}(\hat{\theta}) \to 0$ for large samples, which requires that both bias and variance become negligible.

The law of large numbers implies that the sample mean is a consistent estimator for the true mean in random samples.

R lab: consistency of the sample mean

```
M <- 10^5; n <- c(10, 30, 100); y <- matrix(NA, M, 3)
for(j in 1:3) {
  for(i in 1:M) {
    y[i,j] <- mean(rpois(n[j], 1))
    }}
for(j in 1:3)
  hist.scott(y[,j], xlim=c(0,2)); abline(v=1,col=2)
```



2.1.0.4 Efficiency

An **efficient** estimator is an estimator that estimates the parameter of interest in some *optimal* manner.

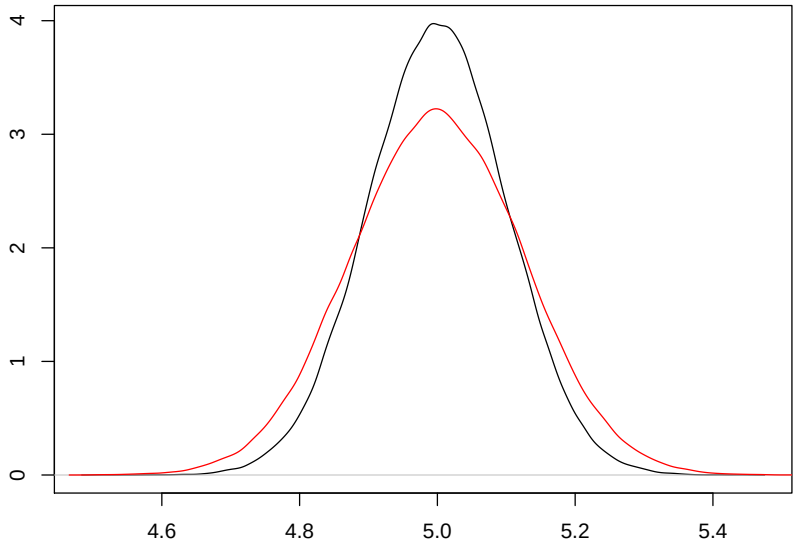
Efficiency are often defined using mean square error.

Among estimators with negligible bias or consistent, efficiency is associated to small variance.

R lab: efficiency of the sample mean

For a normal random sample, both the sample mean and sample median are consistent estimators of μ. The mean is more efficient.

```
M <- 10^5; n <- 100; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {y <- rnorm(n, 5)
  mat[i,] <- c(mean(y), median(y))}
plot(density(mat[,1]), type="l", main="")
lines(density(mat[,2]), col=2)
```



2.1.0.5 Standard Error

An important quantity defined for a (scalar) estimator is given by its **standard error**, defined as

$$\text{SE}(\hat{\theta}) = \sqrt{var(\hat{\theta})}.$$

An example is the standard error of the mean $\text{SE}(\bar{Y}) = \sigma/\sqrt{n}$, which is estimated by $s/\sqrt{n}$.

In applications, the estimated standard error is routinely reported along with the estimate, since it quantifies the **estimation precision**.

2.1.0.6 The delta method

Suppose that we are interested in a parameter which is a function of a scalar parameter θ, namely

$$\psi = g(\theta), \text{ for a continuous and differentiable function } g.$$

If $\hat{\theta}$ is a consistent estimator of θ, then the **continuous mapping theorem** ensures that $g(\hat{\theta})$ is consistent for ψ.

Its standard error is provided by the **delta method**, stating that

$$\text{SE}(\hat{\psi}) \doteq \text{SE}(\hat{\theta})|g^{'}(\theta)|,$$

with the approximation becoming more accurate for large samples.

The result can be extended to settings with multiple parameters.

2.1.0.7 Robust estimation

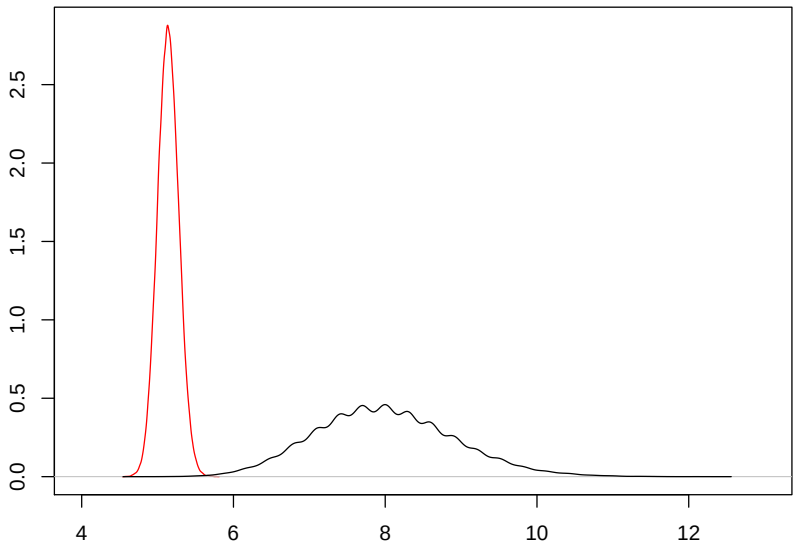
A **robust** estimator has good performances across a wide range of statistical models for the data.

The **sample median** is a robust estimatior of location, not affected by possible outlying data, quite the opposite of the sample mean.

Robust estimators trades some efficiency with resistance to outliers, and they are often a sensible choice for semi-automatic data analyses.

R lab: robustness of the sample median

```
M <- 10^5; n <- 100; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { x <- rbinom(n, size = 1, prob = 0.9)
  y <- x * rnorm(n, 5) + (1 - x) * rnorm(x, 35)
  mat[i,] <- c(mean(y), median(y))}
plot(density(mat[,2]), type="l", main="", xlim=c(4, 13), col = 2)
lines(density(mat[,1]), col=1)
```



2.2 Interval estimation

The aim of interval estimation

Confidence intervals provide more satisfactory estimation results than point estimates alone, giving an entire set of values to estimate the model parameter.

They are built by considering a single parameter at a time.

Extensions to multidimensional *confidence regions* exist, but they are seldom used in practice.

Pivots

Confidence intervals make suitable usage of **pivots**, which are functions of the data and the parameter whose distribution is known.

A notable example is the following one for a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, when the parameter of interest is the mean μ, and σ^2 is not known (so that $\boldsymbol{\theta} = (\mu, \sigma^2)$):

$$T(\mu) = \frac{\bar{Y}\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}, \quad \forall \mu \in \Re, \sigma^2 > 0$$

2.2.0.1 Obtaining a confidence interval

In the normal random sample example, from the previous pivot property it follows that (for $0 < \alpha < 1$)

$$Pr(t_{n-1;\alpha/2} \leq T(\mu) \leq t_{n-1;1-\alpha/2}) = 1\alpha,$$

where $t_{n-1;\alpha}$ is the α quantile of a t_{n-1} distribution; due to symmetry of the latter, $t_{n-1;\alpha} = -t_{n-1;1-\alpha}$.

With some simple algebra, the previous property is equivalent to

$$Pr\left(\bar{Y} - t_{n-1;\alpha/2}\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1\alpha.$$

Definition of confidence interval

Hence the random interval with endpoints

$$\bar{Y} - t_{n-1;\alpha/2}\sqrt{\frac{S^2}{n}}, \quad \bar{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n}}$$

contains μ with probability (1α).

This interval is called a $(1\alpha) \times 100\%$ **confidence interval**.

Common choices are $(1\alpha) = 0.95$ or $(1\alpha) = 0.99$.

Interpretation

Given a particular set of data $y_1, \dots, y_n$ we calculate the confidence interval by replacing $\bar{Y}$ and S^2 with their observed values $\bar{y}$ and s^2

$$\bar{y} - t_{n-1;\alpha/2}\sqrt{\frac{s^2}{n}}, \quad \bar{y} + t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n}}$$

This interval *either does or does not contains the true value of* μ.

The probability interpretation previously introduced refers to an *hypothetical sequence of sets of data* generated from the statistical model.

R lab: confidence interval

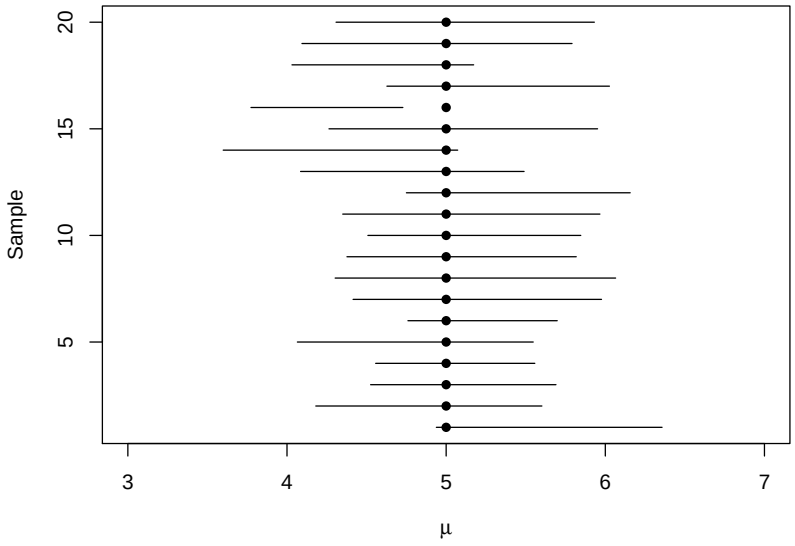
```
M <- 10^5; n <- 10; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
  se_t <- sqrt(var(y) / n) * qt(0.975, n-1)
  mat[i,] <- mean(y) + se_t * c(-1, 1)}

mean(mat[,1] < 5 & mat[,2] > 5)
#> [1] 0.95028
```

R lab: visualizing confidence intervals

We can visualize the first 20 simulated confidence intervals, expecting that (on average) 19 out of 20 will include the true μ

```
plot(rep(5, 20), 1:20, pch = 16, ylab="Sample", xlab=expression(mu))
for(i in 1:20) segments(mat[i,1],i,mat[i,2],i)
```



2.2.0.2 One-sided confidence intervals

If we lift the equi-tailed condition, we can define infinitely many intervals such that

$$Pr(\bar{Y} - t_{n-1;\alpha_1}\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha_2}\sqrt{\frac{S^2}{n}}) = 1\alpha,$$

where $\alpha_1 + \alpha_2 = \alpha$.

Other than the standard choice $\alpha_1 = \alpha_2 = \alpha/2$, other notable choices are $\alpha_1 = 0$ (which makes the lower limit equal to ∞) or $\alpha_2 = 0$ (which makes the upper limit equal to ∞).

They are called **one-sided confidence intervals**, and are sometimes employed in applications.

Approximate confidence intervals & coverage probability

Exact pivots are scarce, but approximate ones are easy to find.

A common one is the **Wald pivot** for a generic parameter of interest ψ, based on a consistent estimator which is approximately normally distributed for large samples

$$Z(\psi) = \frac{\hat{\psi} - \psi}{\text{SE}(\hat{\psi})} \dot{\sim} \mathcal{N}(\prime, \infty), \quad \forall \psi \in \Psi$$

The corresponding confidence interval is

$$\hat{\psi} - z_{1\alpha/2}\text{SE}(\hat{\psi}),\ \hat{\psi} + z_{1\alpha/2}\text{SE}(\hat{\psi})$$

The Central Limit Theorem provides such a solution for random samples, when ψ corresponds to the mean of each variable.

R lab: approximate confidence intervals

```
M <- 10^5; n <- 10; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
  se_z <- sqrt(var(y) / n) * qnorm(0.975)
  mat[i,] <- mean(y) + se_z * c(-1, 1)}

mean(mat[,1] < 5 & mat[,2] > 5)
#> [1] 0.91791
```

```
M <- 10^5; n <- 100; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
  se_z <- sqrt(var(y) / n) * qnorm(0.975)
  mat[i,] <- mean(y) + se_z * c(-1, 1)}

mean(mat[,1] < 5 & mat[,2] > 5)
#> [1] 0.94738
```

Confidence interval for a proportion

The method for approximate intervals can be readily used for confidence intervals on a proportion π, the success probability of a random sample of n binary variables, $Y_i \sim \mathcal{B}(1, \pi)$, $i = 1, \dots, n$.

Here the pivot is

$$Z(\pi) = \frac{\bar{Y} - \pi}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} \dot{\sim} \mathcal{N}(\prime, \infty), \ \forall \pi \in (0, 1),$$

since $\hat{\pi} = \bar{Y}$ and $\text{SE}(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$, which is estimated by plugging-in $\hat{\pi}$ in place of π.

R lab: confidence interval for a proportion

```
M <- 10^5; n <- 50; mat <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rbinom(n, size = 1, prob = 0.25)
   p.hat <- mean(y)
  se_z <- sqrt(p.hat * (1 - p.hat) / n)
  se_qz <- se_z * qnorm(0.975)
  mat[i,] <- mean(y) + se_qz * c(-1, 1)}

mean(mat[,1] < 0.25 & mat[,2] > 0.25)
#> [1] 0.94075
```

3 Hypothesis Testing (Recap)

Fundamentals of hypothesis testing

The idea of hypothesis testing

The basic aim of hypothesis testing within a parametric statistical model $f_\theta(y)$ is **to establish whether the data could be reasonably be generated from** $f_{\theta_0}(y)$, where θ_0 is a specific value of the parameter.

This is simply denoted by the succinct notation

$$H_0 : \theta = \theta_0,$$

with H_0 being termed **null hypothesis**.

Complementary to the choice of H_0, it is required to select a complementary **alternative hypothesis** H_1, specifying the values of the parameter which become reasonable when H_0 does not hold.

Example: testing the mean of a normal sample

Assume the very simple model for independent observations $y_1, y_2, \dots, y_n$ given by $Y_i \sim \mathcal{N}(\mu, 1)$. Then we may want to test

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0$$

which amounts to testing the null hypothesis of data generated from a standard normal distribution, against the possibility that the true mean takes instead a positive value.

This choice of H_1 makes fully sense when we can rule out negative values of μ (**one-sided alternative**). If this is not the case, a better choice would be given by $H_1 : \mu \neq 0$ (**two-sided alternative**).

General formulation

In broad generality, hypothesis on a parameter θ can be cast in the form

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 form a bi-partition of the set containing all the possible values for the parameter θ, that is named the **parameter space** Θ.

In what follows, we will illustrate the main ideas by means of simple, yet important, instances.

Steps of hypothesis testing

The theory of hypothesis testing is rather articulated, so that it may help to go through the main parts of the theory in a systematic fashion.

Some noteworthy concepts are

- Test statistic
- Null and alternative distributions
- p-value
- Significance level, rejection and acceptance regions
- Errors and power

3.0.0.1 Test statistic

A **test statistic** is a statistic (namely, a function of the r.v. representing the available sample) which is used to carry out the test.

Large values (in absolute value) of the test statistic cast doubt on H_0 and on the theory underlying it.

Its choice depends on the problem under study. For the simple normal example mentioned above, a natural choice is to take as test statistic the (standardized) sample mean

$$Z = \frac{\bar{Y}}{\sqrt{\frac{1}{n}}} = \sqrt{n}\bar{Y}$$

3.0.0.2 Null and alternative distributions

The distribution of a test statistic will generally depend on the true value of the parameter under testing.

In the example, if H_0 is true (under H_0), then

$$Z \sim \mathcal{N}(0, 1),$$

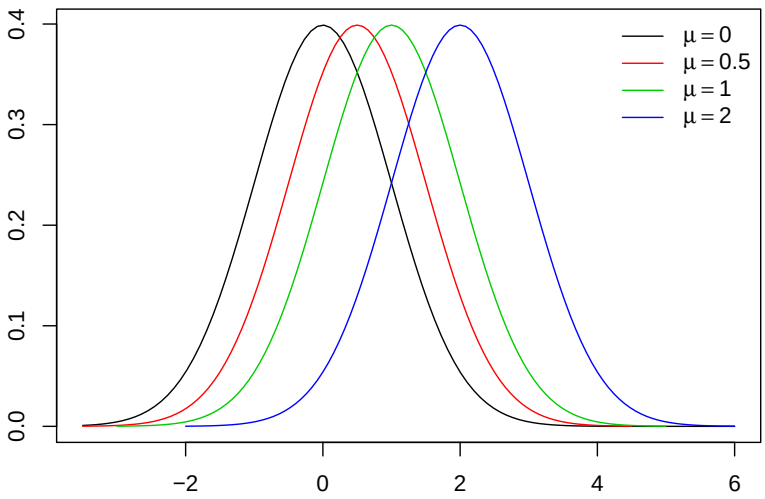
and this is called the **null distribution** of Z.

Instead, if H_1 holds (under H_1), it follows that

$$Z \sim \mathcal{N}(\Delta, 1)$$

where $\Delta = \sqrt{n}\mu > 0$ increases with the value of μ . The distributions valid under H_1 are called the **alternative distributions** of Z.

R lab: visualizing the null and alternative distributions



3.0.0.3 The p-value

The p-value measures the distance between the data and H_0.

Small values of it correspond to a test statistic unlikely to arise under H_0, and suggest that H_0 and the data are inconsistent.

In the example, the idea is that any value larger than the observed z_{obs} (the value of Z computed with the observed data) would cast even greater doubt on H_0.

The p-value is thus defined as the probability (under H_0) of observing a value of the test statistic equal or larger than the observed one

$$p = Pr_{H_0}(Z \geq z_{\text{obs}})$$

Since under H_0 we have $Z \sim \mathcal{N}(0, 1)$, it follows that

$$p = 1\Phi(z_{\text{obs}})$$

R lab: computing the p-value for a sample

In case the null distribution is not known, it would be possible to compute the p-value by simulation whenever it is possible to generate data under H_0.

```
set.seed(13); n <- 10; y_obs <- rnorm(n)
z_obs <- mean(y_obs) * sqrt(n)
print(z_obs)
#> [1] 1.897537

M <- 10^5; z_sim <- numeric(M)
for(i in 1:M) { y <- rnorm(n)
 z_sim[i] <- mean(y) * sqrt(n) }
```

```
c(mean(z_sim >= z_obs), 1 - pnorm(z_obs))
#> [1] 0.02877000 0.02887856
```

Other alternative hypotheses: more details

For the simple example of test on μ and the same $H_0 : \mu = 0$, other two possibilities for H_1 could then be considered.

In either case, the same test statistic Z would still be used, but the computation of the p-value would change, due to the different direction of deviation from H_0.

For $H_1 : \mu < 0$, small values of Z would flag deviation from H_0 (that is, negative values with large absolute value), so that

$$p = Pr_{H_0}(Z \leq z_{\text{obs}}) = \Phi(z_{\text{obs}}).$$

Instead, for $H_1 : \mu \neq 0$, both directions ought to be considered, and

$$p = Pr_{H_0}(|Z| \geq |z_{\text{obs}}|) = 2Pr_{H_0}(Z \geq |z_{\text{obs}}|) = 2(1 - \Phi(|z_{\text{obs}}|)).$$

3.0.0.4 Significance level

We commonly say that a the result of a test is significant at the 5% level whenever the p-value is ≤ 0.05. Other levels of some practical interest are 1% or 0.1%.

An often-followed convention is

Range	Evidence.against.the.null.hypothesis
0.05<p<=0.1	marginal evidence
0.01<p=<0.5	evidence
0.001<p<=0.01	strong evidence
p<=0.001	very string evidence

A test with *fixed significance level* arises when the significance level is fixed in advance, and then it is just reported whether the p-value is smaller than the fixed level.

If this happens, it may be reported that H_0 is **rejected**, otherwise we may say that H_0 is **not rejected** (**or accepted**).

3.0.0.5 Rejection and acceptance regions

If we define the sample space as the set of the values that our available sample may take, the **rejection region** of a test with fixed significance level is the subset of the sample space corresponding to the samples that would lead to a rejection of H_0.

The remaining part of the sample space forms instead the **acceptance region**.

Both these two regions are determined by means of a test statistic.

Rejection and acceptance regions for the example

In the simple normal example previously introduced, for $H_1 : \mu > 0$, it is simple to verify that a rejection region of level α is simply

$$R_\alpha = \{\mathbf{y} : Z \geq z_{1\alpha}\},$$

where $z_{1\alpha}$ is the standard normal $(1-\alpha)$-quantile, i.e. 1.645 for $\alpha = 0.05$.

The acceptance region is just given by

$$A_\alpha = \{\mathbf{y} : Z < z_{1\alpha}\}.$$

(Note: the computation of the p-value, and of R_α and A_α would be exactly the same if the null hypothesis were of the form $H_0 : \mu \leq 0$, maintaining the same alternative hypothesis.)

3.0.0.6 Errors for a fixed-significance level test

When we adopt a test with fixed significance level, we move from using the p-value as a measure of evidence against H_0 to using a test to decide which of H_0 and H_1 is more supported by the data.

Two wrong decisions are possible. We commit a **Type I error** by rejecting H_0 when it is true, or a **Type II error** by accepting H_0 when it is false.

In the example, $Pr_{H_0}(Y \in R_\alpha) = \alpha$, and in fact *the fixed significance level equals the probability of making a Type I error.*

3.0.0.7 Power of a test

For a test with fixed significance level, the **power** is the probability of (correctly) detecting that H_0 is false

$$Pr_{H_1}(Y \in R_\alpha)(= 1 - \beta).$$

The power of a test can be used for comparing alternative tests for the same problem, with tests with higher power being preferable.

The power is often used for designing studies, in particular for choosing the sample size in medical or industrial studies.

Indeed, for fixed significance level, the power increases with the sample size.

Power of two tests for the example

For the simple example (with $H_1 : \mu > 0$), an alternative (but silly) test statistic may be given by taking the same Z as above computed by using only half of the sample (for n even).

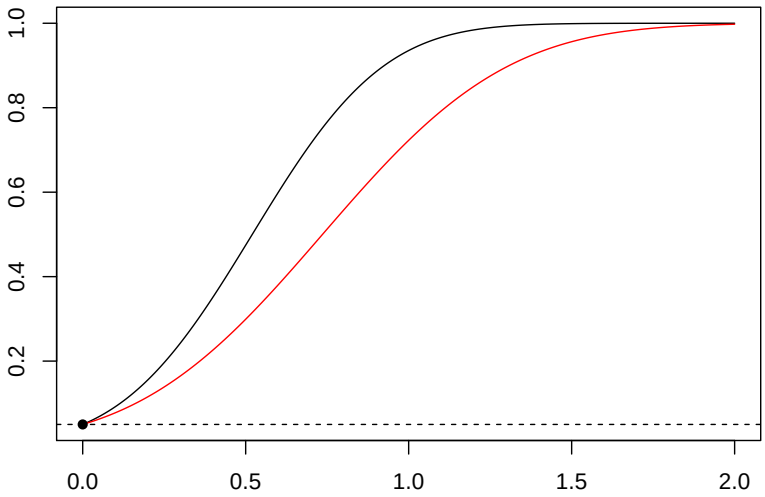
Fixing a significance level of 5%, the two tests have exactly the same probability of a Type I error, so for comparing them we must use their power.

The power is a function of the μ assumed under H_1, and for a certain $\mu \geq 0$ we obtain (since $z_0.95 = 1.645$)

$$Pr_\mu(Z \geq 1.645) = 1 - \Phi(1.645 - \sqrt{n}\mu)$$

R lab: power of two alternative tests

```
mu <- seq(0, 2 , l = 1000); n <- 10; n1 <- 5
plot(mu, 1 - pnorm(1.645 - sqrt(n) * mu), type = "l",
ylab="Power", xlab = expression(mu))
lines(mu, 1 - pnorm(1.645 - sqrt(n1) * mu), col = 2)
abline(h=0.05, lty = 2); points(0, 0.05, pch = 16)
```



3.0.0.8 Comments on the p-value

The usage of p-values is not free of controversies, and in ending the review of the general theory on testing some comments are in order.

1. The **p-value is NOT the probability that H0 is true**, since the latter is not even an event.
2. The results of statistical tests, and p-values in particular, should never taken without considering **context-specific knowledge**. Even a small p-value may not be particularly meaningful if the alternative hypothesis is logically implausible.
3. Hypothesis testing is useful in certain contexts, but it has some important limitations. For (very) large sample sizes, even tiny deviations from the null hypothesis will lead to small p-values. For large sample sizes, there are alternative approaches which are more fruitful, and techniques based on **model selection** are often preferable to statistical tests.

Some commonly used tests

One-sample t-test

Given a normal random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, a classical testing problem on μ is of the form (for two-sided alternative, say)

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

The test statistic is given by

$$T = \frac{\bar{Y}\mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n1}, \quad \text{when } H_0 \text{ is true}$$

with the p-value given by

$$p = Pr_{H_0}(|T| \geq |t_{\text{obs}}|)$$

which can be computed as $p = 2Pr_{H_0}(T \geq |t_{\text{obs}}|) = 2(1 - F_{t_{n1}}(|t_{\text{obs}}|))$.

3.0.0.9 Example

The DAAG book introduces the simple dataset pair65, about an experiment on the effect of heat on the stretchiness of elastic bands: a small sample of differences between two different conditions for 9 bands.

heated	ambient	difference
244	225	19
255	247	8
253	249	4
254	253	1
251	245	6
269	259	10
248	242	6
252	255	-3
292	286	6

Focusing on the 9 differences on the amount of stretch, we test

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

by means of the t.test function, resulting in significance at 5% level

```
#>
#>  One Sample t-test
#>
#> data:  pair65$difference
#> t = 3.1131, df = 8, p-value = 0.01438
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>   1.641939 11.024728
#> sample estimates:
#> mean of x
#>  6.333333
```

Relation between tests and confidence intervals

As displayed for the pair65 data testing, the t.test R function returns also the confidence interval for the parameter under testing, in that case the true mean of the differences in stretchiness.

This is not by chance, since there is a close connection between hypothesis testing on the value of a certain parameter and confidence intervals for that parameter.

For the case of a mean, for example, the basic idea is that If the confidence interval for μ does not contain zero, this is equivalent to rejection of the hypothesis that the true mean is zero.

Nonparametric tests

Nonparametric tests specify only partially a statistical model for the data, so that they may provide more robust inferences than parametric tests with contaminated data, outliers or, more generally, in settings where model specification is hard.

This is sometimes useful, especially when only certain aspects of the data are of interest, or for checking the results obtained with a full model specification.