Hypothesis Testing

Giovanni Saraceno

Contents


```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

We introduced the following techniques in statistical inference

- point estimation: how to attempt to find the value of an unknown parameter.
- confidence intervals estimation: how to determine fo an unknown parameter, an interval that contains its true value with high probability.

Now, we focus on **Hypothesis Testing**. Differently from point and confidence interval estimation, through hypothesis testing we want to verify if the estimated parameter differ *significantly* from an expected value. We answer to the question: "There exists any effect?"

The researcher has some theory about the world (some phenomenon under study), and wants to determine whether the data support that theory. In other words, hypothesis testing wants to answer to how to proceed for the acceptance or rejection of a particular hypothesis about an unknown parameter.

The first step is the statement of two hypotheses - the **null hypothesis** and the **alternative hypothesis** - about a parameter of interest.

The null hypothesis H_0 is a specific statement, that is the parameter of the population is equal to a determined value which expresses the absence of any effect (no preference, no correlation, no difference). We are interested in rejecting the null hypothesis since we can learn something about the alternative hypothesis.

The alternative hypothesis H_A represents all the remaining cases with respect to the values expressed in the null hypothesis. This hypothesis should coincide with the research question that the researcher hopes to be true.

We evaluate the strength of evidence by assuming that the null hypothesis is true and determining how improbable it would be to observe sample results or statistics as extreme as, or more extreme than, those in the original sample. The objective of a hypothesis test is not to demonstrate that the alternative hypothesis is (probably) true; the goal is to demonstrate that the null hypothesis is (probably) false.
Data are tested against the null hypothesis. Then, if data are compatible with the statement in the null hypothesis, we say that "we cannot reject" and never "we accept" the null hypothesis. Note that, rejecting H_0 means that we can exclude the tested value without any additional information on the true value of the parameter. Estimation aims to find estimated values and their standard errors.

Let's consider again $X \sim f(x;\theta)$ where $\theta \in \Theta$, the statistical test wants to find a partition of the parametric space $\Theta = \{\Theta_0, \Theta_1\}$ such that:

$$H_0 : \theta \in \Theta_0 \qquad H_A : \theta \in \Theta_1$$

In the case we are testing against a specific value $H_0 : \theta = \theta_0$, the alternative hypothesis is bilateral or two-tailed, that is it allows values greater and lower than θ_0.

After looking at the estimated value from the sample we make a decision about whether to believe that the null hypothesis is correct, or to reject the null hypothesis in favor of the alternative. In order to do this we need to calculate a *test statistic*.

After computing the statistic, for deciding if the data are compatible with the null hypothesis we compute the probability of how much data are in disagreement, assuming that H_0 is true. To compute this probability we determine the distribution of the test statistic under H_0, or *null distribution*. If this probability is low, we can consider the null hypothesis incompatible in favor of the alternative hypothesis. This probability is called *p-value*. In other words, the p-value is the probability of obtaining the sample data if the null hypothesis is true.
Remark: in the communication of the results, it is always necessary to provide the value of the test statistics and the corresponding confidence interval (or at least the standard errors) additionally to the obtained p-value.

It may happen that we wrongly reject the null hypothesis or we do not have enough evidence for rejection. Indeed, if we do not reject H_0, do not imply that H_0 is true, since we are considering the information only in the given sample. Specifically we can have

- **Type I errors**: we reject H_0 which is true,
- **Type II errors**: we do not reject H_0 when it is not true.

The probability of type I errors is called *significance level* and it is indicated with α. The probability of a Type II error is denoted by β and $1-\beta$ indicates the *power* of the statistical test. Or equivalently, the power is the probability that the test statistic rejects H_0 when it is false. In summary

	retain H0	reject H0
H0 true	$1-\alpha$ (probability of correct retention)	α (type I error)
H0 false	β (type II error)	$1-\beta$ (power of the test)

Optimal statistical tests aim to minimize the type I errors while maximizing the power.

Let's consider the data set generated in the previous generated example, saved in `"first_dataframe.csv"`.

```
dat <- read.csv("../Exploratory Data Analysys/first_dataframe.csv")
dat$Sex <- as.factor(dat$Sex)
dat$Group <- as.factor(dat$Group)
```

Test for the Mean

Consider the `Response_Time` variable and display its mean and median

```
c(mean(dat$Response_Time), median(dat$Response_Time))
```

```
## [1] 2104.120 2107.057
```

We have seen that for the sample mean

$$Z = \sqrt{n}\left(\frac{\bar{X} - \mu_X}{\sigma_X}\right) \to N(0,1).$$

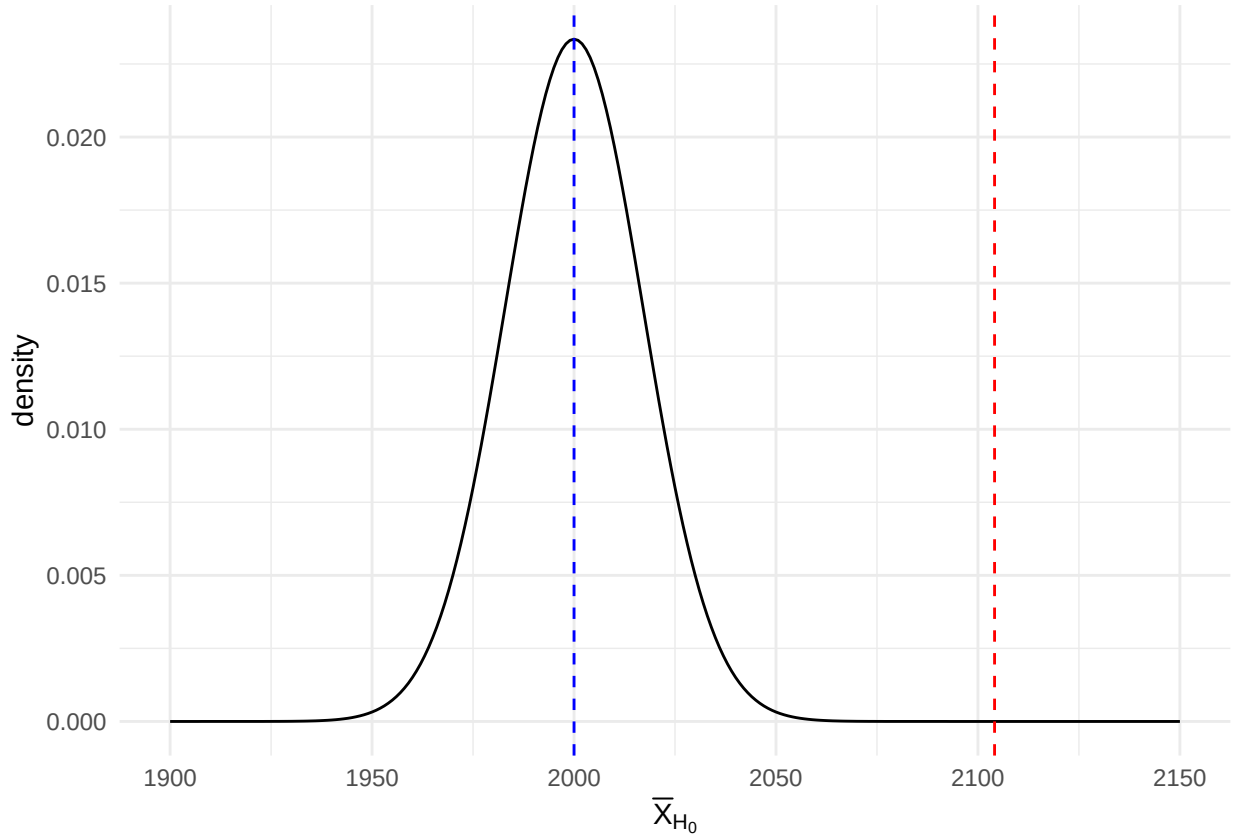
Using this information, we constructed the α confidence intervals, with $\alpha \in (0,1)$.

```r
mean_R <- mean(dat$Response_Time)
sd_R <- sd(dat$Response_Time)
n <- nrow(dat)
alpha <- 0.05
lb <- round(mean_R - qt(1-alpha/2, n-1)*sd_R/sqrt(n), digits = 3)
rb <- round(mean_R + qt(1-alpha/2, n-1)*sd_R/sqrt(n), digits = 3)
print(paste0("The confidence interval for the estimate ", mean_R, " is [", lb, ";", rb, "]"))
```

```
## [1] "The confidence interval for the estimate 2104.1201142988 is [2070.55;2137.69]"
```

Assume that the average response time is 2000 ms. Then we can construct a test where $H_0 : \mu = 2000$ vs $H_1 : \mu \neq 2000$. We can represent the distribution of the null hypothesis with the estimated value of the mean

```r
x <- seq(1900, 2150, 1)
y <- dnorm(x, mean = 2000, sd= sd_R/sqrt(n))
 ggplot() +
   geom_line(mapping = aes(x = x, y = y)) +
   theme_minimal() +
   geom_vline(xintercept = mean_R, color = "red", linetype = "dashed") +
   geom_vline(xintercept = 2000, color = "blue", linetype = "dashed") +
    xlab(expression(bar(X)[H[0]])) +
   ylab("density")
```

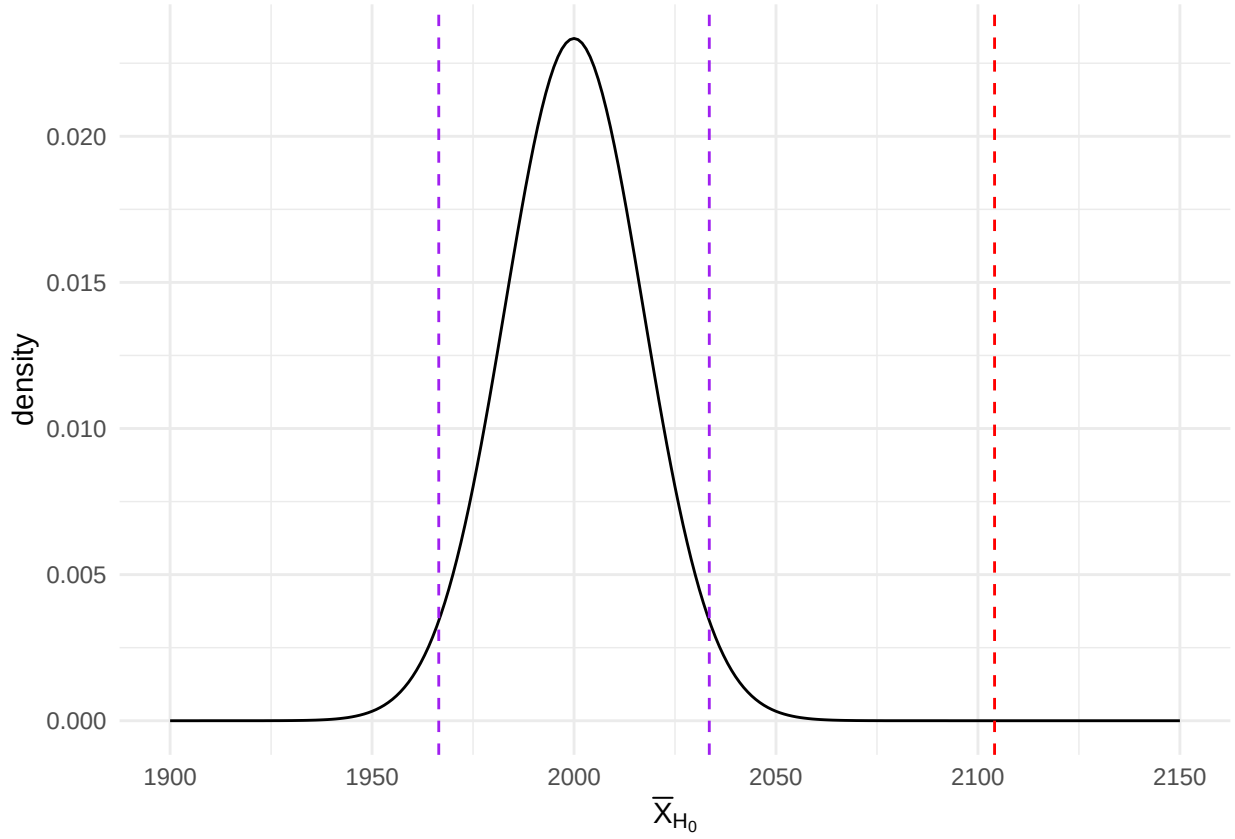


By using the distribution of the test statistic, it is possible to identify a range of values that are unlikely to occur if H_0 is true. According to the significance level, we define a **critical region**, values of the test the lead to the rejection of the null hypothesis, and an **acceptance region**.
Suppose that we consider a significance level $\alpha = 0.05$, the critical region of a two-tailed test will corresponds to the most extreme values (tails of the distribution). In this example

```
alpha <- 0.05
quantiles <- qnorm(c(alpha/2, 1- alpha/2),
      mean = 2000,
      sd = sd_R/sqrt(n))
quantiles
```

```
## [1] 1966.511 2033.489
```

```
ggplot() +
   geom_line(mapping = aes(x = x, y = y)) +
   theme_minimal() +
   geom_vline(xintercept = mean_R, color = "red", linetype = "dashed")  +
    xlab(expression(bar(X)[H[0]])) +
   ylab("density")+
   geom_vline(xintercept = quantiles[1], color = "purple", linetype = "dashed")+
   geom_vline(xintercept = quantiles[2], color = "purple", linetype = "dashed")
```



We can note that our statistic (sample mean) falls into the right critical region, so we reject the null hypothesis. To sum up:

- choose an α level,
- come up with some test statistic that does a good job (in some meaningful sense),
- figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true,
- calculate the critical region that produces an appropriate α level.

In addition we can compute the "famous" p-value. Note that a good critical region almost always corresponds to those values of the test statistic that are least likely to be observed if the null hypothesis is true. Then

the p-value can be defined as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.

```
2*pnorm(mean_R,
      mean = 2000,
      sd = sd_R/sqrt(n),
      lower.tail = FALSE)
```

```
## [1] 1.102789e-09
```

The obtained value is very low, so we reject the null hypothesis.
Remark: the p-value is not the probability that the null hypothesis is true or false.

In some circumstances, it is useful to consider *one-tailed tests* where the alternative hypothesis includes only one side of the value specified in the null hypothesis. For example, in the study of the response time assume that we are interested in showing some evidence that the response time of the population under study is higher than the average. Then we can consider

$$H_0 : \mu = 2000 \text{ vs } H_1 : \mu > 2000.$$

```
t.test(dat$Response_Time, mu=2000, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  dat$Response_Time
## t = 6.0938, df = 499, p-value = 1.104e-09
## alternative hypothesis: true mean is greater than 2000
## 95 percent confidence interval:
##  2075.963      Inf
## sample estimates:
## mean of x
##   2104.12
```

Then H_0 is not rejected

What is the meaning of not rejecting H_0? Unfortunately, we cannot state that H_0 is true, since ther is the possibility that the true parameter is equal to a value very close to the tested value, or that the result of the test depend on the low sample size. How do we interpret the result?

- We can state that the data are compatible with the null hypothesis. There are no evidences that suggest to consider more complicated models for understanding the tested parameter.
- The null hypothesis provide a conservative statement, while usually the alternative implies an effect, that is, the evidence are not so strong for rejecting H_0.

Remark: Do not terminate a study with the results of a hypothesis testing. In order to make inference on the population, it is important to combine hypothesis testing with the results from the estimation of parameters and confidence intervals. For example, if we reject the null hypothesis but the confidence interval is large, then we may not have enough information.

Example Let's consider an example where a metallurgical industry produces plates with a thickness of 14mm and a tolerance of 0.5mm. Every shift (every 6 hours), 10 plates are sampled and their thickness is measured.

```
thickness <- c(13.88, 14.03, 14.11, 13.77, 14.04, 14.05, 13.94, 13.95, 13.94, 13.91)
alpha <- 0.05
Var <- 0.01
```

```r
x_bar <- mean(thickness)
n <- length(thickness)
```

Consider that after each maintenance, the system is calibrated so that the mean thickness of the plates is 14mm. Therefore, a value of $\bar{X}$ different from this could indicate a malfunction. In this case, it is useful to perform a hypothesis test considering $H_0 : \mu = 14$ and $H_1 : \mu \neq 14$. If the null hypothesis is verified, then it can be concluded (with a certain level of confidence) that the system is functioning correctly. First, we calculate the test statistic

```r
mu0 <- 14
sd_thick <- sd(thickness)
z_oss <- (x_bar - mu0)/(sd_thick/sqrt(n))
```

We know that Z described above follows $N(0, 1)$. The considered hypothesis test is a two-tailed test, so we can accept H_0 with confidence level α if the observed value falls within the α confidence interval.

```r
alpha <- 0.05
interval <- c(x_bar + qt(alpha / 2, n-1)*sd_thick/sqrt(n), x_bar + qt(1 - alpha / 2, n-1)*sd_thick/sqrt
interval
```

```
## [1] 13.89136 14.03264
```

```r
x_bar
```

```
## [1] 13.962
```

The true value is within the interval, so we cannot reject the null hypothesis. Alternatively, we can calculate the p-value. In a two-tailed hypothesis test, this is given by:

```r
pval <- 2*pt(z_oss, df=n-1)
pval
```

```
## [1] 0.2545875
```

The p-value does not allow to reject the null hypothesis. The result confirms that the system is functioning correctly for the considered sample. We can do the same using the function `t.test`

```r
t.test(thickness, mu = 14)
```

```
##
##  One Sample t-test
##
## data:  thickness
## t = -1.2169, df = 9, p-value = 0.2546
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  13.89136 14.03264
## sample estimates:
## mean of x
##    13.962
```

Let us suppose we want to verify whether the system is currently calibrated for thicknesses greater than 14mm. In this case, the null hypothesis is $H_0 : \mu = 14$ (equivalently $\mu \leq 14$) and $H_1 : \mu > 14$, or $H_0 : \mu = 14$ (equivalently $\mu \geq 14$) and $H_1 : \mu < 14$. For example, in the first case:

```r
pval <- pt(z_oss, df=n-1)
pval
```

```
## [1] 0.1272937
```

Example. A wholesale supplier of bathroom fixtures wants to maintain internal control over sales. To do so, invoices must be accompanied by a transfer receipt to remove goods from the warehouse. At the end of each month, a sample of invoices is taken to evaluate the average amount reported. Over the last 5 years, the average invoice amount has been \$120. Given that transport costs are influenced by delivery distance, it is important to monitor the average amount. Consider the following sample:

```
fatture <- c(108.98, 152.22, 111.45, 110.59, 127.46, 107.26, 93.32,
             91.97, 111.56, 75.71, 128.58, 135.11)
mx <- mean(fatture)
s2 <- var(fatture)
```

We assume the distribution of the amounts, described by the random variable X, can be approximated by a normal distribution, $X \sim N(\mu_X, \sigma_X^2)$. We know that:

$$t = \sqrt{n}\frac{X - \bar{X}}{\sqrt{S^2}} \sim t_{n-1}$$

follows a t-distribution with $n-1$ degrees of freedom. Next, we compare the obtained sample mean with the known value $\mu_0 = 120$. It is appropriate to perform a two-tailed hypothesis test with $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. We calculate the observed t-value

```
n <- length(fatture)
alpha <- 0.01
z <- qt(1 - alpha / 2, n - 1)
interval <- c(mx - z * sqrt(s2 / n), mx + z * sqrt(s2 / n))
interval
```

```
## [1]  94.2040 131.4977
```

The value μ_0 is inside the 95%-confidence interval. Equivalently the interval of the test statistic

```
# the interval of the test statistic
t_oss <- (mx - 120) * sqrt(n / s2)
t_oss
```

```
## [1] -1.190761
```

```
c(z , z)
```

```
## [1] 3.105807 3.105807
```

The statistic is inside the acceptance region. Alternatively we can calculate the p-value

```
pval <- 2 * pt(-t_oss, n - 1, lower.tail = FALSE)
pval
```

```
## [1] 0.2588093
```

Both results suggest that we fail to reject the null hypothesis. The t.test function allows you to specify the type of test (two-tailed or one-tailed) with the alternative argument.

```
# Two-tailed test
t.test(fatture, mu = 120, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  fatture
## t = -1.1908, df = 11, p-value = 0.2588
## alternative hypothesis: true mean is not equal to 120
## 99 percent confidence interval:
```

```
##   94.2040 131.4977
## sample estimates:
## mean of x
##  112.8508
```

```
# One-tailed tests
t.test(fatture, mu = 120, alternative = "greater", conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  fatture
## t = -1.1908, df = 11, p-value = 0.8706
## alternative hypothesis: true mean is greater than 120
## 99 percent confidence interval:
##  96.53186      Inf
## sample estimates:
## mean of x
##  112.8508
```

```
t.test(fatture, mu = 120, alternative = "less", conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  fatture
## t = -1.1908, df = 11, p-value = 0.1294
## alternative hypothesis: true mean is less than 120
## 99 percent confidence interval:
##      -Inf 129.1698
## sample estimates:
## mean of x
##  112.8508
```

In all cases, the confidence interval and p-value match the manually computed values, confirming that the null hypothesis cannot be rejected.

Test for Proportions

Recall that, given the random variable $X \sim Ber(p)$ and the sample $X_1, \ldots, X_n$ from X, the sample mean or *sample proportion* represents the proportion of successes in the random sample.
The sampling distribution of $\hat{p}$ can be approximated by a Normal distribution

$$\hat{p} \approx N(p, p(1-p)/n)$$

with n sufficiently large.

Consider the variable `Sex` in the dat data set. We want to compute the proportion of male, and then test is the sample proportion is equal to the value 0.5. We can do this using the `prop.test` function

```
prop.test(table(dat$Sex), conf.level = 0.95, alternative = "less")
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  table(dat$Sex), null probability 0.5
## X-squared = 0, df = 1, p-value = 0.5
## alternative hypothesis: true p is less than 0.5
```

```
## 95 percent confidence interval:
##  0.0000000 0.5366809
## sample estimates:
##   p
## 0.5
```

Example: Political Candidate A politician wants to run for office in a district with 100,000 voters. Before announcing the candidacy, they want to assess their likelihood of success. To do this, a survey company is hired to contact 2,500 voters. Out of these, 1,328 declare themselves in favor of the candidate, which corresponds to a percentage of:

$$\frac{1328}{2500} \cdot 100\% = 53\%.$$

We want to infer the unknown percentage p of voters favoring the candidate. The assumptions are: - All respondents have the same probability of being included in the sample. - Responses are independent (no influence among respondents).

Under these conditions, the number of supporters y among the n surveyed voters can be modeled as a random variable $Y \sim Bin(n, p)$. An estimate is given by:

```
n <- 2500
y <- 1328
p_hat <- y / n
p_hat
```

```
## [1] 0.5312
```

In general, the confidence interval is constructed using the normal approximation

$$\frac{\hat{p} - p}{\sqrt{p(1-p)}/\sqrt{n}} \sim N(0, 1),$$

hence:

```
alpha <- 0.1
Var <- p_hat * (1 - p_hat) / n
interval <- c(p_hat - qnorm(1 - alpha / 2) * sqrt(Var),
              p_hat + qnorm(1 - alpha / 2) * sqrt(Var))
interval
```

```
## [1] 0.5147835 0.5476165
```

Given this result, the politician may conclude that there is a reasonable expectation of winning. It is also useful to perform a hypothesis test with $H_0 : p = p_0$, where $p_0 = 0.5$ (indifference among voters), and $H_1 : p > p_0$

```
p0 <- 0.5
p_hat
```

```
## [1] 0.5312
```

```
 c(-Inf,
  p0 + qnorm(1 - alpha / 2) * sqrt(Var))
```

```
## [1]      -Inf 0.5164165
```

The sample proportion is in the critical region. Alernatively, we can compute the p-value

```r
p_obs <- (p_hat - p0) / sqrt(Var)
pnorm(p_obs, lower.tail = FALSE)
```

```
## [1] 0.0008857304
```

Both results suggest rejecting the null hypothesis in favor of the alternative hypothesis. The test can also be performed using an exact binomial test with `binom.test` or using `prop.test`, which is based on the chi-squared distribution

```r
binom.test(y, n, alternative = "greater", conf.level = 0.9)
```

```
##
##  Exact binomial test
##
## data:  y and n
## number of successes = 1328, number of trials = 2500, p-value =
## 0.0009647
## alternative hypothesis: true probability of success is greater than 0.5
## 90 percent confidence interval:
##  0.5181948 1.0000000
## sample estimates:
## probability of success
##                 0.5312
```

```r
prop.test(y, n, correct = FALSE, alternative = "greater", conf.level = 0.9)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  y out of n, null probability 0.5
## X-squared = 9.7344, df = 1, p-value = 0.0009043
## alternative hypothesis: true p is greater than 0.5
## 90 percent confidence interval:
##  0.5183932 1.0000000
## sample estimates:
##      p
## 0.5312
```

Example: Dr. Spock and the Jury Selection Dr. Benjamin Spock was a prominent pediatrician who, in 1968, was tried in a federal court in Boston for conspiracy against the Military Service Act due to his involvement in the anti-Vietnam War movement. The jury selection raised controversy: only 102 out of 350 potential jurors were women, even though 53% of eligible voters were female.

The judge claimed the selection was random. Let us assess this claim using a hypothesis test. Let N denote the total number of eligible voters, and D the number of women among them. Under random selection, the probability of choosing a woman in any draw is $p_0 = D/N$. Assuming N is large, this probability remains approximately constant across selections. Thus, we can model X, the number of women among the selected jurors, as $X \sim Bin(n, p)$, where n is the number of selected jurors.

```r
n <- 350
d <- 102
pn <- d / n
pn
```

```
## [1] 0.2914286
```

```r
p0 <- 0.53
```

We test $H_0 : p = p_0$ (the selection followed the law), $H_1 : p \neq p_0$ (the selection was biased).We can use binom.test

```r
binom.test(d, n, p0)
```

```
##
##  Exact binomial test
##
## data:  d and n
## number of successes = 102, number of trials = 350, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.53
## 95 percent confidence interval:
##  0.2443353 0.3420915
## sample estimates:
## probability of success
##              0.2914286
```

```r
prop.test(d, n, p0, corr = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  d out of n, null probability p0
## X-squared = 79.971, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.53
## 95 percent confidence interval:
##  0.2462908 0.3410950
## sample estimates:
##         p
## 0.2914286
```

In all cases, the obtained p-value is very low, suggesting rejection of the null hypothesis. This indicates that it is difficult to believe the jury selection was random.

Conclusion of the Story:
On July 10, 1968, Dr. Spock was sentenced to two years in prison and fined $5,000. However, the verdict was overturned by the United States Court of Appeals in 1969 due to insufficient evidence.

Comparison of Means of Two Normal Populations

When the difference in population means is analyzed, we must think about the type of sampling design we have:

(i) Design with independent random samples,
(ii) Paired sampling design.

These two sampling designs result in differences in the methods used to compare the two populations.
Consider two independent random variables X_1 and X_2, and corresponding samples of size n_1 and n_2, respectively. We assume that $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$. To test whether $\mu_1 > \mu_2$, we set up the hypotheses $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. This is equivalent to:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs } \mu_1 - \mu_2 > 0.$$

Hence, we can consider the results seen for testing the mean estimate of a normally distributed random variable.

As in the case of a single mean μ, if we use an estimate of σ_1^2 and σ_2^2 instead of the true population value inside the confidence interval formulation, we must consider the quantile of the student t-distribution instead of the standard normal one. In this case the degrees of freedom $df = n_1 + n_2 - 2$ for the first case, while the in the second case has a more complicated formula that we will not consider here.

Example: Wage Comparison Between Unionized and Non-Unionized Women The Wall Street Journal on July 26, 1994, stated:

"Women who are union members earn $2.50 per hour more than women who are not union members."

Based on this statement, it seems advantageous for women in the U.S. to be part of a union. Suppose we have samples of wages for women who are union members and those who are not:

```
iscritte <- c(22.40, 18.90, 16.70, 14.05, 16.20, 20.00, 16.10, 16.30, 19.10, 16.50, 18.50, 19.80, 17.00
non_iscritte <- c(17.60, 14.40, 16.60, 15.00, 17.65, 15.00, 17.55, 13.30, 11.20, 15.90, 19.20, 11.85, 1
length(iscritte)
```

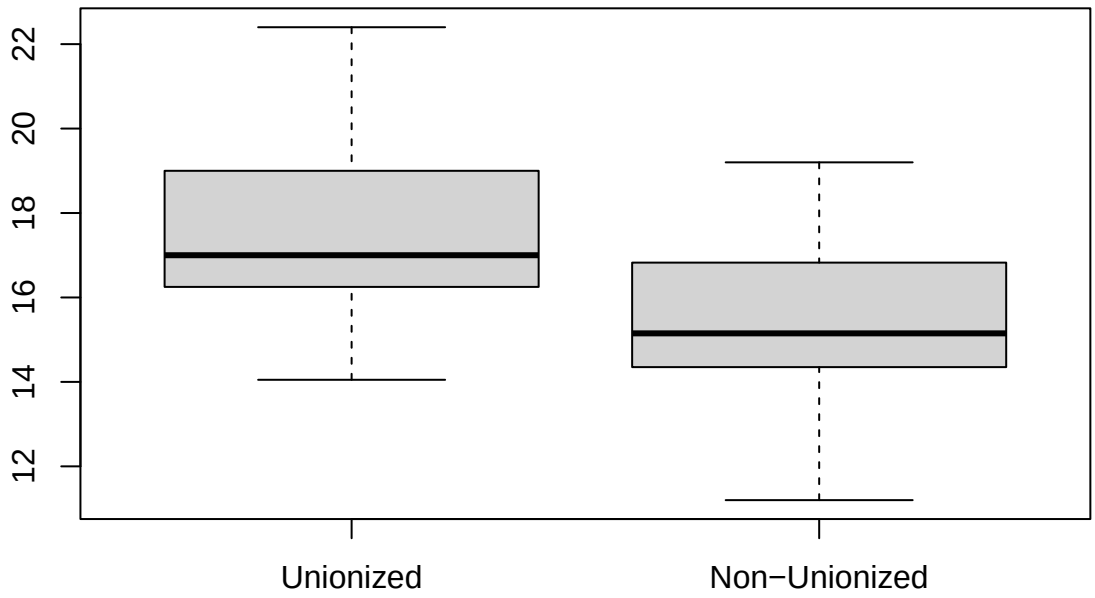
```
## [1] 15
```

```
length(non_iscritte)
```

```
## [1] 20
```

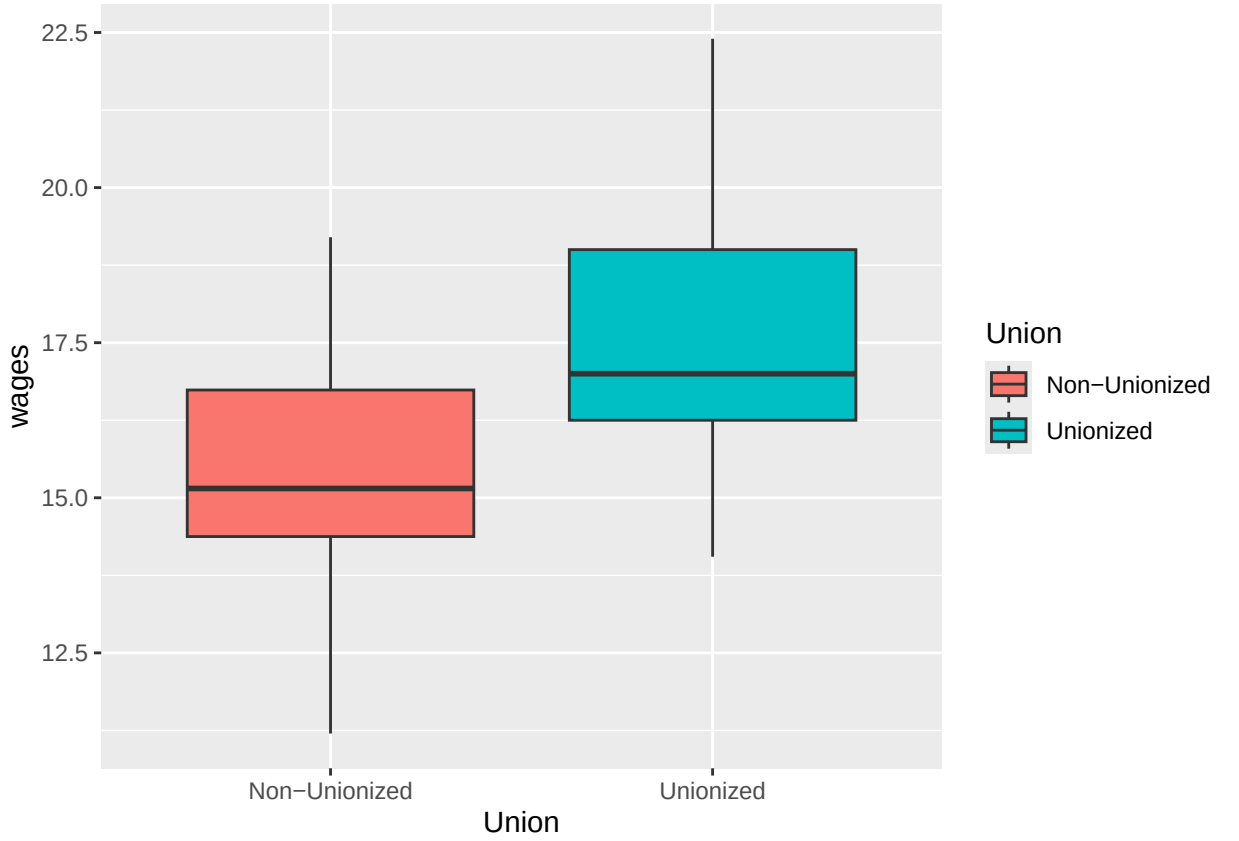
We note that the two samples have different sizes. Descriptive statistics and visualizations can be helpful for initial comparisons:

```
boxplot(iscritte, non_iscritte, names = c("Unionized", "Non-Unionized"))
```




```
union_dat <- data.frame(wages=c(iscritte, non_iscritte),Union = factor(rep(c("Unionized", "Non-Unionize
ggplot(union_dat, aes(x = Union, y=wages, fill=Union))+
```

```
  geom_boxplot()
```



The boxplot comparison suggests that unionized women earn more on average than non-unionized women, but this visualization does not provide any measure of statistical significance or reliability.

Assume that the wages of both groups follow a normal distribution with means μ_1 and μ_2, respectively, and the same variance $\sigma_1 = \sigma_2$. We also assume the two populations are independent. The sample means are:

```
mu1 <- mean(iscritte)
mu2 <- mean(non_iscritte)
c(mu1, mu2)
```

```
## [1] 17.53667 15.36000
```

To test whether $\mu_1 > \mu_2$, we set up the following hypotheses: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. This is equivalent to:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs } \mu_1 - \mu_2 > 0.$$

Since we assume equal variances, we calculate the pooled variance:

```
s1 <- var(iscritte)
s2 <- var(non_iscritte)
n <- length(iscritte)
m <- length(non_iscritte)
s <- ((n - 1) * s1 + (m - 1) * s2) / (n + m - 2)
```

The test statistic is given by:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_1}{\sqrt{s(\frac{1}{n} + \frac{1}{m})}}.$$

Calculate the test statistic and p-value

```
mu1 - mu2
```

```
## [1] 2.176667
```

```
alpha <- 0.05
c(-Inf, qt(1 - alpha, n + m - 2, lower.tail = FALSE)* sqrt(s * (1 / n + 1 / m)))
```

```
## [1]      -Inf -1.213325
```

```
# Or computing the p-value
t_obs <- (mu1 - mu2) / sqrt(s * (1 / n + 1 / m))
pt(t_obs, n + m - 2, lower.tail = FALSE)
```

```
## [1] 0.002327084
```

Alternatively, use the built-in t-test function:

```
t.test(iscritte, non_iscritte, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  iscritte and non_iscritte
## t = 3.036, df = 33, p-value = 0.002327
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.963342      Inf
## sample estimates:
## mean of x mean of y
##  17.53667  15.36000
```

The test suggests rejecting the null hypothesis in favor of the alternative. The confidence interval for the difference in means did not include 0, supporting the conclusion that unionized women earn significantly more than non-unionized women, on average.

Remark: Notice that in this case we are assuming that the two populations have equal variance. In case the sample sizes are large, the t-test still works even if the standard deviations differ up to 3 times. If this difference is larger than 3 times, or the groups are unbalanced (the sample sizes are very different), the t test should not be used. In the case of different variances, we can perform the t test with the Welch approximation.

Paired samples Let's now see the last case, i.e., we have a paired sampling design.

Given $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, when paired sampling is used, we We can consider

$$D = X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma^2_{X_1 - X_2}).$$

For example, considering the data set `dat` previously loaded, we can compare the Response_Time collected in occasion 0 and occasion 1. This data are paired, since they refer to the same observation.

```
t.test(dat$Response_Time[which(dat$Time==1 & dat$Group=="Control")],
       dat$Response_Time[which(dat$Time==2& dat$Group=="Control")],
       conf.level = 0.95,
       var.equal = FALSE,
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dat$Response_Time[which(dat$Time == 1 & dat$Group == "Control")] and dat$Response_Time[which(
## t = -12.353, df = 24, p-value = 6.847e-12
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -238.4414 -170.1697
## sample estimates:
## mean difference
##       -204.3055
```

Comparison of Variances

Consider the following simulated data:

```r
n <- 38
m <- 50
x <- rnorm(n, sd = 5)
y <- rnorm(m, sd = 3)
s2x <- var(x)
s2y <- var(y)
c(s2x, s2y)
```

```
## [1] 26.046961  8.045141
```

Suppose $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. In this example, we are interested in comparing the variability of the two datasets. After calculating the sample variances, we want to perform the following hypothesis test

$$\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2 \\ H_1 : \sigma_x^2 \neq \sigma_y^2 \end{cases}$$

Since variances are strictly greater than zero, we focus on the ratio σ_x^2/σ_y^2. Under the null hypothesis $\sigma_x^2 = \sigma_y^2$, we know

$$\frac{(n-1)s_x^2}{\sigma_x^2} \sim \chi_{n-1}^2, \quad \frac{(m-1)s_y^2}{\sigma_y^2} \sim \chi_{m-1}^2.$$

Thus, under H_0, the ratio of variances follows an F-distribution

$$\frac{s_x^2}{s_y^2} \sim F(n-1, m-1).$$

The confidence interval for the variance ratio is calculated as

```r
s2x / s2y
```

```
## [1] 3.237602
```

```r
alpha <- 0.05
interval <- c(
  (s2x / s2y) / qf(1 - alpha / 2, n - 1, m - 1),
  (s2x / s2y) / qf(alpha / 2, n - 1, m - 1)
)
interval
```

```
## [1] 1.778867 6.050048
```

We can perform an F-test using the `var.test` function

```r
var.test(x, y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 3.2376, num df = 37, denom df = 49, p-value = 0.0001428
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.778867 6.050048
## sample estimates:
## ratio of variances
##           3.237602
```

For a one-sided test to check if $\sigma_x^2 > \sigma_y^2$, specify the alternative hypothesis

```r
var.test(x, y, alternative = "greater")
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 3.2376, num df = 37, denom df = 49, p-value = 7.139e-05
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.959994      Inf
## sample estimates:
## ratio of variances
##           3.237602
```

The F-test results include:

- Test statistic (F): The ratio of variances.
- Degrees of freedom: $n-1$ and $m-1$ for the numerator and denominator.
- p-value: Indicates the strength of evidence against H_0.

Comparison of two Proportions

Consider the following data, representing the number of graduates in 2001 in Economics at Ca' Foscari and Bocconi. The graduates are classified based on the time elapsed between graduation and their first job.

```r
laureati <- data.frame(Ca.Foscari = c(480, 129), Bocconi = c(1338, 438))
rownames(laureati) <- c("meno di un anno", "piu' di un anno")
laureati
```

```
##                 Ca.Foscari Bocconi
## meno di un anno        480    1338
## piu' di un anno        129     438
```

We add the total number of graduates for both universities and each category to the table.

```r
laureati$Totale <- laureati$Ca.Foscari + laureati$Bocconi
rbind(laureati, apply(laureati, 2, sum))
```

```
##                 Ca.Foscari Bocconi Totale
## meno di un anno        480    1338   1818
## piu' di un anno        129     438    567
## 3                      609    1776   2385
```

The event of interest is finding a job after graduation. Based on the available information, we can assume that the random variable describing this event follows a binomial distribution, where success corresponds to finding a job within a year of graduation. First, we calculate the empirical success probabilities for each university.

```
n1 <- sum(laureati$Ca.Foscari)
prop1 <- laureati$Ca.Foscari / n1
prop1
```

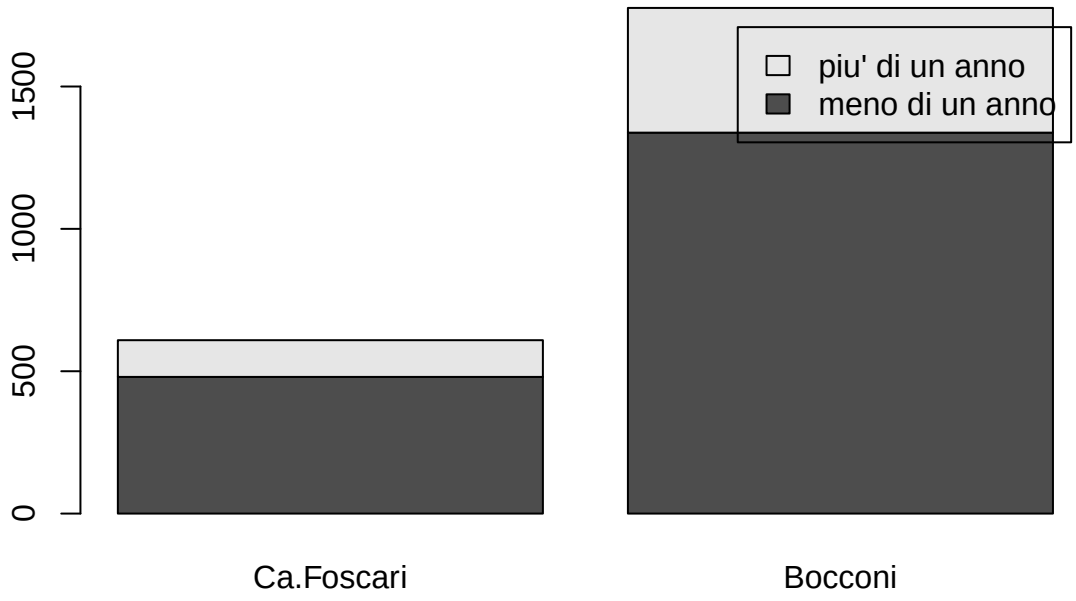
```
## [1] 0.7881773 0.2118227
```

```
n2 <- sum(laureati$Bocconi)
prop2 <- laureati$Bocconi / n2
prop2
```

```
## [1] 0.7533784 0.2466216
```

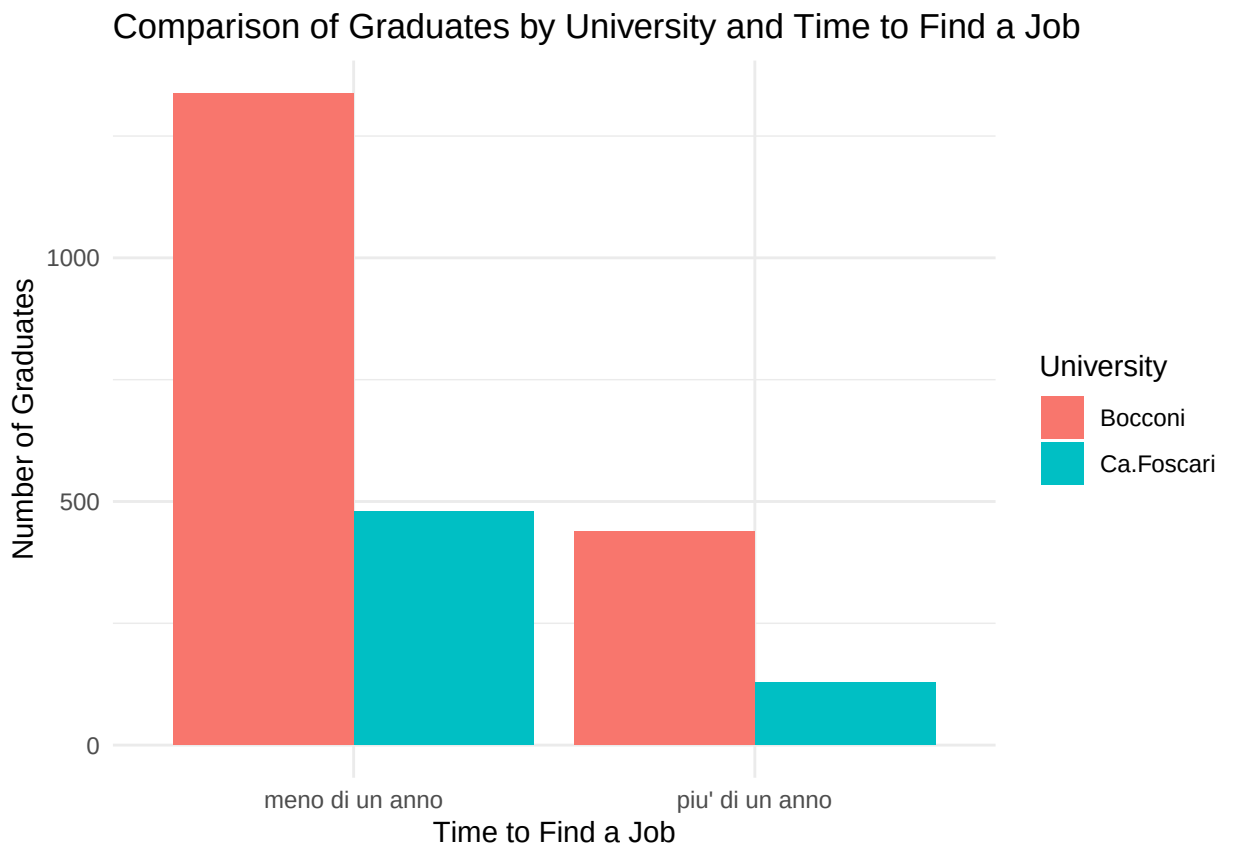
It is helpful to represent this data visually using bar plots to compare the quantities. For instance

```
barplot(as.matrix(laureati[, -3]), legend.text = c("meno di un anno", "piu' di un anno"))
```



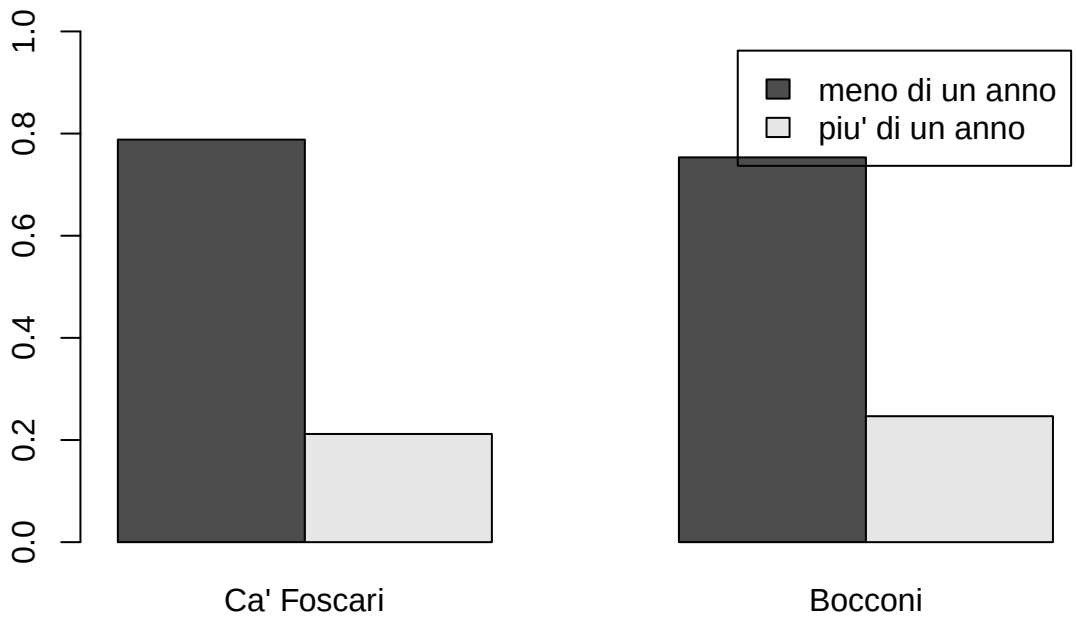

```
laureati$Category <- rownames(laureati)
laureati_long <- pivot_longer(laureati, cols = c("Ca.Foscari", "Bocconi"), names_to = "University", valu

# Generate the bar plot
ggplot(laureati_long, aes(x = Category, y = Count, fill = University)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Comparison of Graduates by University and Time to Find a Job",
```

```
    x = "Time to Find a Job",
    y = "Number of Graduates",
    fill = "University"
  ) +
  theme_minimal()
```



This plot is not very informative due to the difference in total graduates between the two universities. It is more useful to represent the proportions

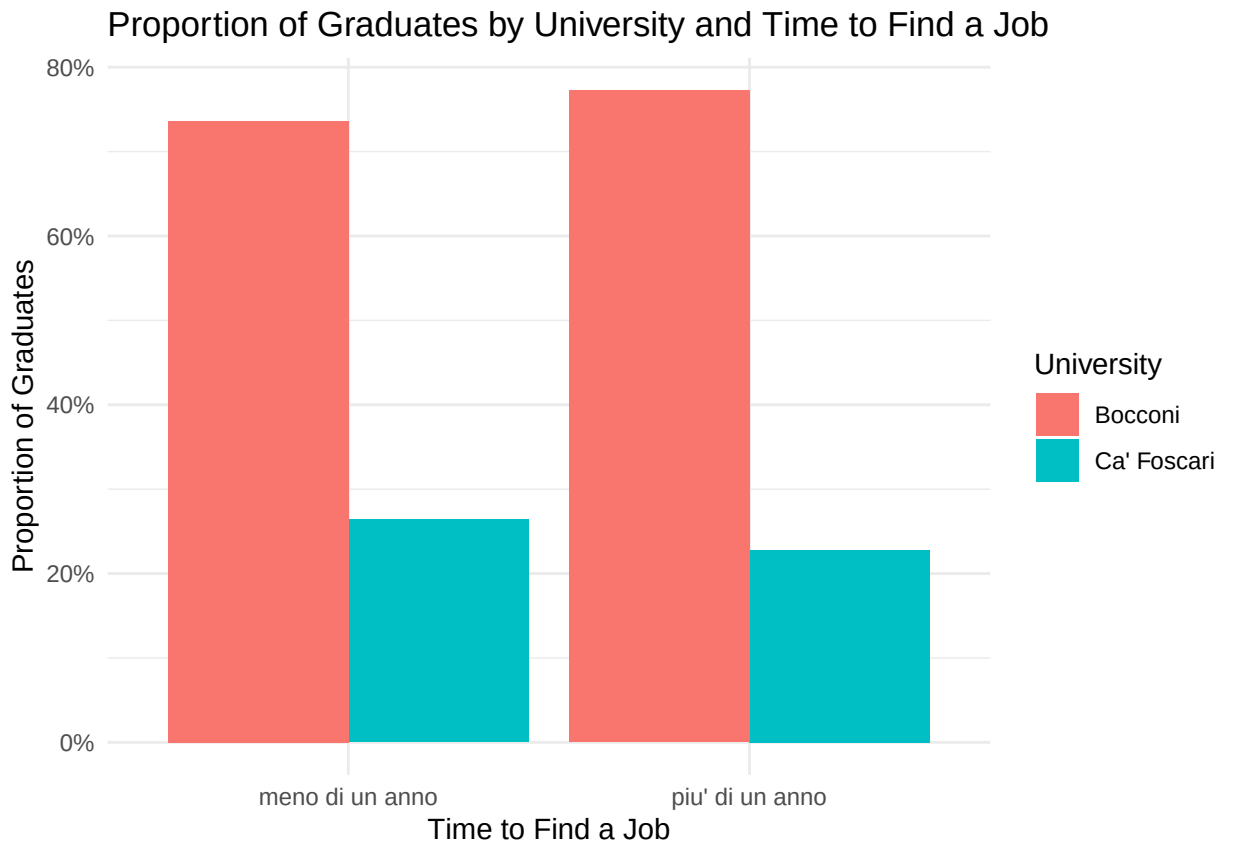
```
barplot(cbind(prop1, prop2), beside = TRUE, ylim = c(0, 1),
        legend.text = c("meno di un anno", "piu' di un anno"),
        names.arg = c("Ca' Foscari", "Bocconi"))
```



Note that the *beside* option creates subgroups.

```
# Calculate proportions for each university
laureati <- laureati %>%
  mutate(
    Prop_Ca_Foscari = Ca.Foscari / Totale,
    Prop_Bocconi = Bocconi / Totale
  )

# Reshape the data to long format for ggplot
laureati_long <- laureati %>%
  select(Category, Prop_Ca_Foscari, Prop_Bocconi) %>%
  pivot_longer(cols = c("Prop_Ca_Foscari", "Prop_Bocconi"),
               names_to = "University",
               values_to = "Proportion") %>%
  mutate(University = recode(University,
                             "Prop_Ca_Foscari" = "Ca' Foscari",
                             "Prop_Bocconi" = "Bocconi"))

# Generate the bar plot
ggplot(laureati_long, aes(x = Category, y = Proportion, fill = University)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::percent) + # Display proportions as percentages
  labs(
    title = "Proportion of Graduates by University and Time to Find a Job",
    x = "Time to Find a Job",
    y = "Proportion of Graduates",
```

```
    fill = "University"
  ) +
  theme_minimal()
```



The statistics suggest that graduates from Ca' Foscari find jobs more quickly. To assess the reliability of this result based on the sample data, we can use confidence intervals and hypothesis tests. Let $Y_1 \sim \text{Bin}(n_1, p_1)$ and $Y_2 \sim \text{Bin}(n_2, p_2)$ be independent random variables describing the number of graduates at Ca' Foscari and Bocconi, respectively. Using the normal approximation, we know that $\frac{Y_1}{n_1} \sim N(p_1, \frac{p_1(1-p_1)}{n_1})$, and similarly for $\frac{Y_2}{n_2}$. To compare the two variables, we consider the difference

```
pn1 <- prop1[1]
pn2 <- prop2[1]
# Pooled variance
s <- pn1 * (1 - pn1) / n1 + pn2 * (1 - pn2) / n2
```

A 95% confidence interval is given by:

```
alpha <- 0.05
intervallo <- c(pn1 - pn2 - qnorm(1 - alpha / 2) * sqrt(s),
                pn1 - pn2 + qnorm(1 - alpha / 2) * sqrt(s))
intervallo
```

```
## [1] -0.003345431  0.072943355
```

Finally, we perform the following hypothesis test:

$$\begin{cases} H_0: & p_1 - p_2 = 0 \\ H_1: & p_1 - p_2 > 0 \end{cases}$$

```r
alpha <- 0.05
intervallo <- c(-qnorm(1 - alpha / 2) * sqrt(s),
                qnorm(1 - alpha / 2) * sqrt(s))
intervallo
```

```
## [1] -0.03814439  0.03814439
```

```r
pn1 - pn2
```

```
## [1] 0.03479896
```

The sample difference is inside the acceptance region, then we should not reject the null hypothesis. However, it is very close to the border of the interval, suggesting that the choice strictly depends on the chosen confidence level of the test. Using the `prop.test` function:

```r
prop.test(c(laureati[1, "Ca.Foscari"], laureati[1, "Bocconi"]),
          c(n1, n2), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(laureati[1, "Ca.Foscari"], laureati[1, "Bocconi"]) out of c(n1, n2)
## X-squared = 2.8414, df = 1, p-value = 0.04593
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.001684628 1.000000000
## sample estimates:
##    prop 1    prop 2
## 0.7881773 0.7533784
```

Using $\alpha = 0.05$, the test suggests rejecting the null hypothesis in favor of the alternative, indicating that the probability of finding a job is higher for graduates from Ca' Foscari. Would our evaluation change if $\alpha = 0.01$? This shows that the p-value should never be used as a threshold for firmly choosing for rejection or not.

Exercises

Exercise 8 Suppose we measured the average lifespan of a sample of 15 fluorescent light bulbs, given by:

```r
tem_med <- c(2928, 2997, 2689, 3081, 3011, 2996, 2962, 3007, 3000, 2953, 2792, 2947, 3094, 2913, 3017)
```

To market the bulbs domestically, the mean lifespan printed on the box must have an error of 1%, while for international markets the error must be 5%. Is the value $\mu_0 = 3010$ hours compatible with these two markets? Construct an appropriate hypothesis test and provide an evaluation.

Exercise 9 Suppose that thanks to a new production process, a rope manufacturing factory obtained the following results from 25 breaking strength tests (expressed in Newtons):

```r
rottura <- c(1975, 1869, 1879, 1790, 1860, 1895, 1810, 1831, 1759, 1585, 1553, 1774, 1640, 1761, 1946,
```

Traditional ropes have a breaking strength of 1730N. Does the new production process significantly improve the rope quality? Construct a hypothesis test to answer this question.

Exercise 10 A company produces metal tubes of standard length. Twenty years ago, the production quality was tested, and it was found that the tube lengths followed a normal distribution. The company claims that the standard deviation of tube lengths is at most 1.2cm. A customer decides to test this claim by selecting 25 tubes:

```r
tubi <- c(10.19, 8.74, 12.83, 7.45, 10.11, 9.811, 9.48, 7.56, 9.21, 9.78, 11.98, 10.01, 10.78, 7.06, 12
```

- Does the sample standard deviation weaken the company's claim?
- Compute the 95% confidence interval.
- Perform an appropriate hypothesis test to provide an answer.