# Estimation and Confidence intervals

Giovanni Saraceno

## Contents



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

So far, we have limited ourselves to simply describing the data, but data is collected to discover something about the population from which the information is drawn. In practice, a common data analysis task involves making inferences about an unknown aspect of a population of interest using observed data that is sampled from that population. Usually, we don't have access to data for the entire population. Questions in data analysis that pertain to how the summaries, patterns, trends, or relationships in a data set can be extended to the broader population are referred to as inferential statistical questions.

Common techniques in **statistical inference** include

- 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
- hypothesis testing: how to proceed to the acceptance or rejection of a particular hypothesis about the parameter.

### Point estimation

Assume that the variable $X$ is a random variable describing some quantity of interest. Specifically, we want to draw inference on the entire population $X$, which encompasses the complete set of individuals or cases we want to study. We assume that the population distribution depends on some unknown parameter(s). *Estimation* is the process for inferring the population parameters starting from the sample data.
For instance, consider 100 observations of the variable $X$ following a Normal distribution. The sample mean $\bar{x}$ and sample variance $s^2$ are sample estimates of the parameters $\mu$ and $\sigma^2$ for the entire population.

```
# For example generate 100 obervations from a Normal distribution with mean 70 and standard deviation 1
set.seed(1234)
x <- rnorm(100, mean = 70, sd=10)
mean(x)
```

```
## [1] 68.43238
```

```
sd(x)
```

```
## [1] 10.04405
```

The parameters $\mu$ and $\sigma$ give information about the location and dispersion of the population. Recall that, for the normal distribution, the values less than one standard deviation from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%. This property of the normal distribution is also know as *68–95–99.7 (empirical) rule*, or the *3-sigma rule*.

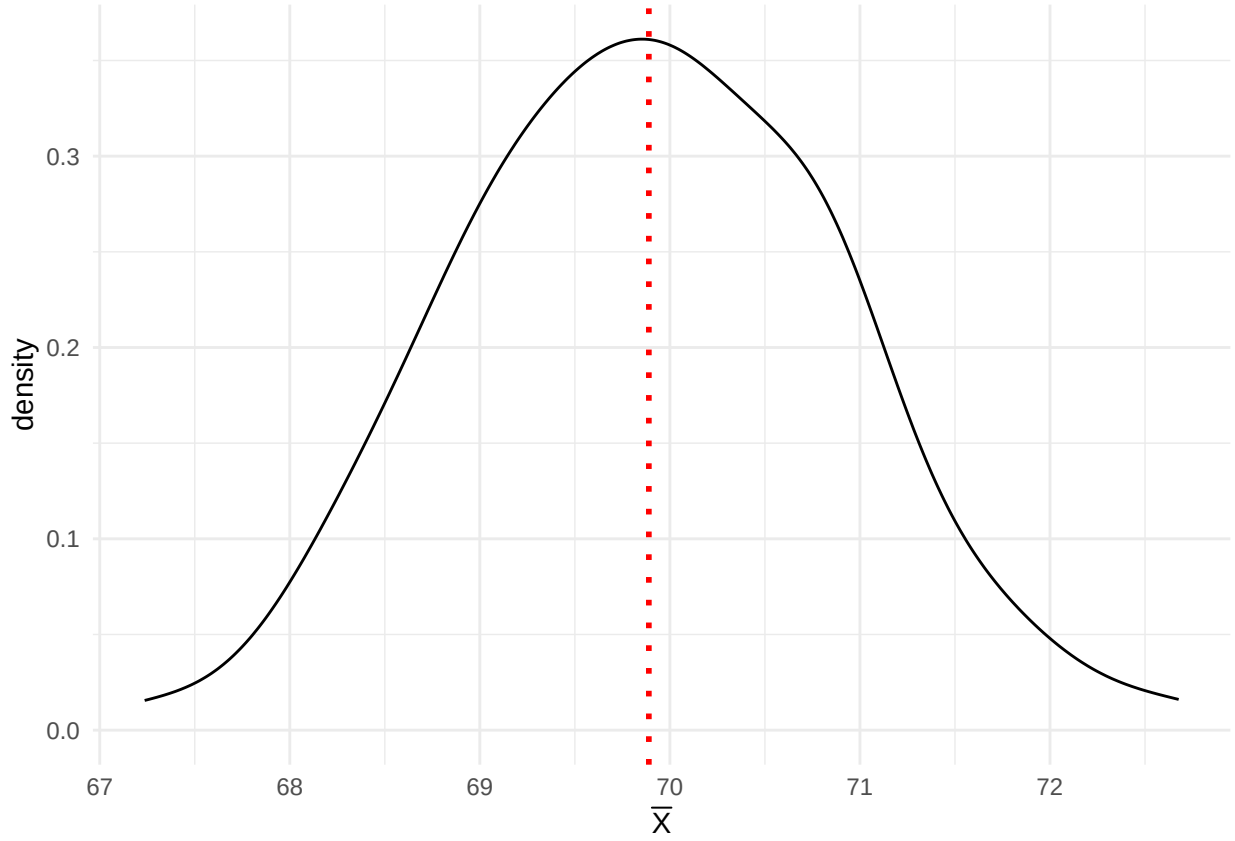
Given a parameter $\theta$, since sampling is influenced by randomness, how much confidence can we have in its estimate? For example, consider to generate from the above normal distribution 5 times (using different seeds) and to compute the sample mean and sample variance in each iteration.

```
for(i in 1:5){
  set.seed(1234 + i)
  x <- rnorm(100, mean = 70, sd=10)
  print(paste0(round(mean(x),2), " (", round(sd(x),2), ")"))
}
```

```
## [1] "70.78 (10.5)"
## [1] "71.45 (8.33)"
## [1] "70.31 (10.45)"
## [1] "70.25 (9.51)"
## [1] "69.27 (10.09)"
```

Resampling data points, the computed estimate has different values. Does this mean that the estimate could be unreliable?
Indeed, estimates vary from sample to sample due to sampling variability. To answer this question, we consider the sampling distribution of the estimate (referred to as the *estimator* $\hat{\theta}$, which is a random variable). For example, we can sample 50 observations 100 times and compute the sample mean in each iteration. Then we can study the distributions of computed means

```
vect_mean <- rep(0,100)
for(i in 1:100){
  x <- rnorm(100, mean = 70, sd=10)
  vect_mean[i] <- mean(x)
}
vect_mean <- data.frame(Mean = vect_mean)
ggplot(vect_mean, aes(x=Mean))+
  geom_density()+
  theme_minimal()+
  geom_vline(xintercept=mean(vect_mean$Mean), col = "red", linetype = "dotted", linewidth =1)+
  xlab(expression(bar(X)))
```

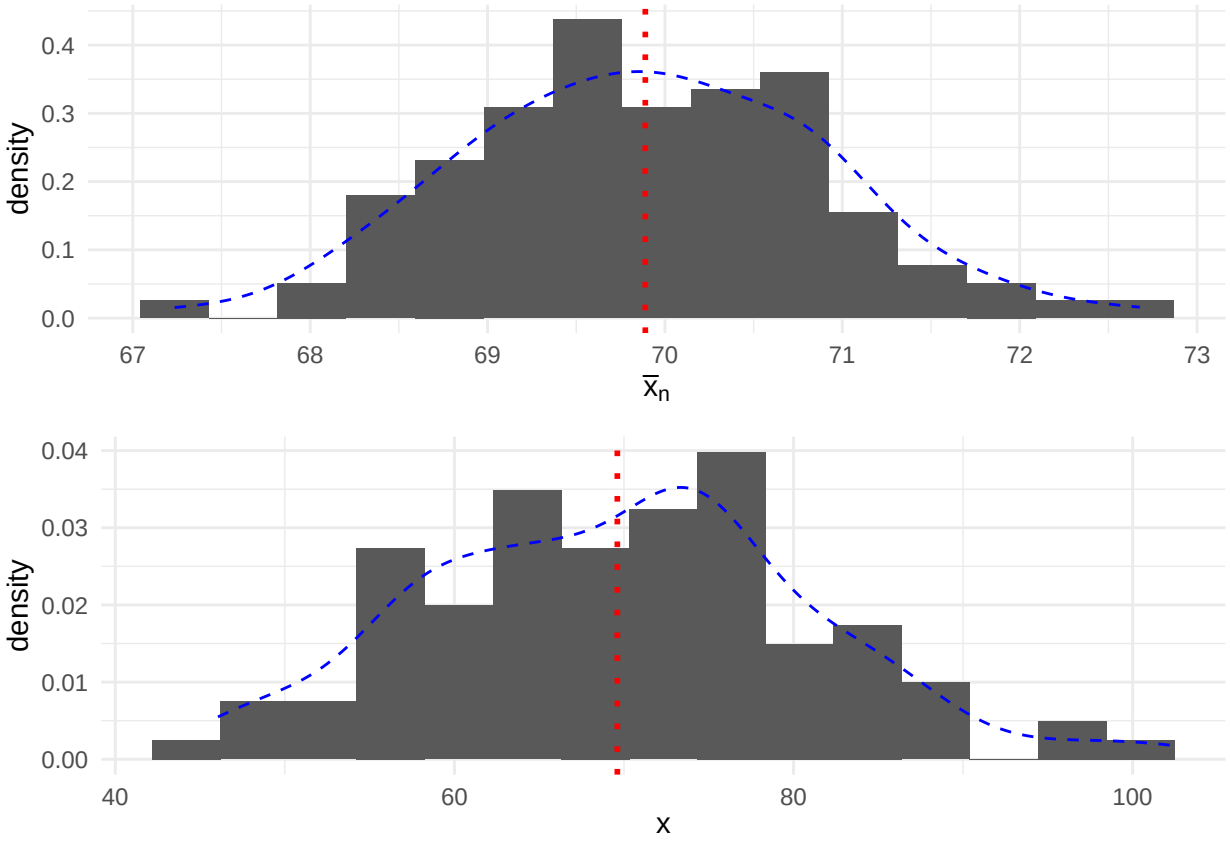



```
summary(vect_mean)
```

```
##       Mean
##  Min.   :67.24
##  1st Qu.:69.21
##  Median :69.94
##  Mean   :69.89
##  3rd Qu.:70.67
##  Max.   :72.68
```

We can also look at the distributions of the original sample and the estimated means.

```
g1 <- ggplot(vect_mean)+
  geom_histogram(aes(x=Mean, y = after_stat(density)), bins=15)+
  theme_minimal()+
  xlab(expression(bar(x)[n])) +
  geom_vline(xintercept = mean(vect_mean$Mean), col = "red", linetype = "dotted", linewidth = 1) +
  geom_density(aes(x = Mean), color = "blue", linetype = "dashed")
g2 <- ggplot()+
  geom_histogram(aes(x=x, y = after_stat(density)), bins=15)+
  theme_minimal()+
  xlab(expression(x)) +
  geom_vline(xintercept = mean(x), col = "red", linetype = "dotted", linewidth = 1) +
  geom_density(aes(x = x), color = "blue", linetype = "dashed")
gridExtra::grid.arrange(g1,g2)
```




The *sampling distribution* of the estimate is the probability distribution of the values of an estimate obtained from sampling a population infinitely many times. Additionally, an estimator is called unbiased if $\mathbb{E}[\hat{\theta}] = \theta$, that is the estimator gives "correct" result on average. Understanding the sampling distribution and its properties enables us to make probabilistic assessments of how well the statistic provides information about the population parameter it is defined for.
We want estimators satisfy two crucial asymptotic properties, that hold for increasing sample size (for the limit of $n \to \infty$). In particular, consider the estimator $\hat{\theta}_n$, where the $n$ indicates the dependence on the sample size, and let $\theta$ be the true parameter. The estimator is said to be *consistent* if it converges in probability to the true parameter value as the sample size grows, that is

$$\hat{\theta}_n \to \theta \text{ as } n \to \infty$$

. This ensures that the estimate will be arbitrarily close to the true parameter value as the sample size grows. An estimator is said to have *asymptotic normality* if

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma^2),$$

that is, for large sample sizes, the distribution of the scaled estimator approaches a normal distribution with mean 0 and some variance $\sigma^2$.
Considering the previous figures, we can note that the sampling distribution has a bell shape, and it has a lower spread than the population or sample distributions, i.e.

```
var(vect_mean$Mean)
```

```
## [1] 1.007442
```

```
var(x)
```

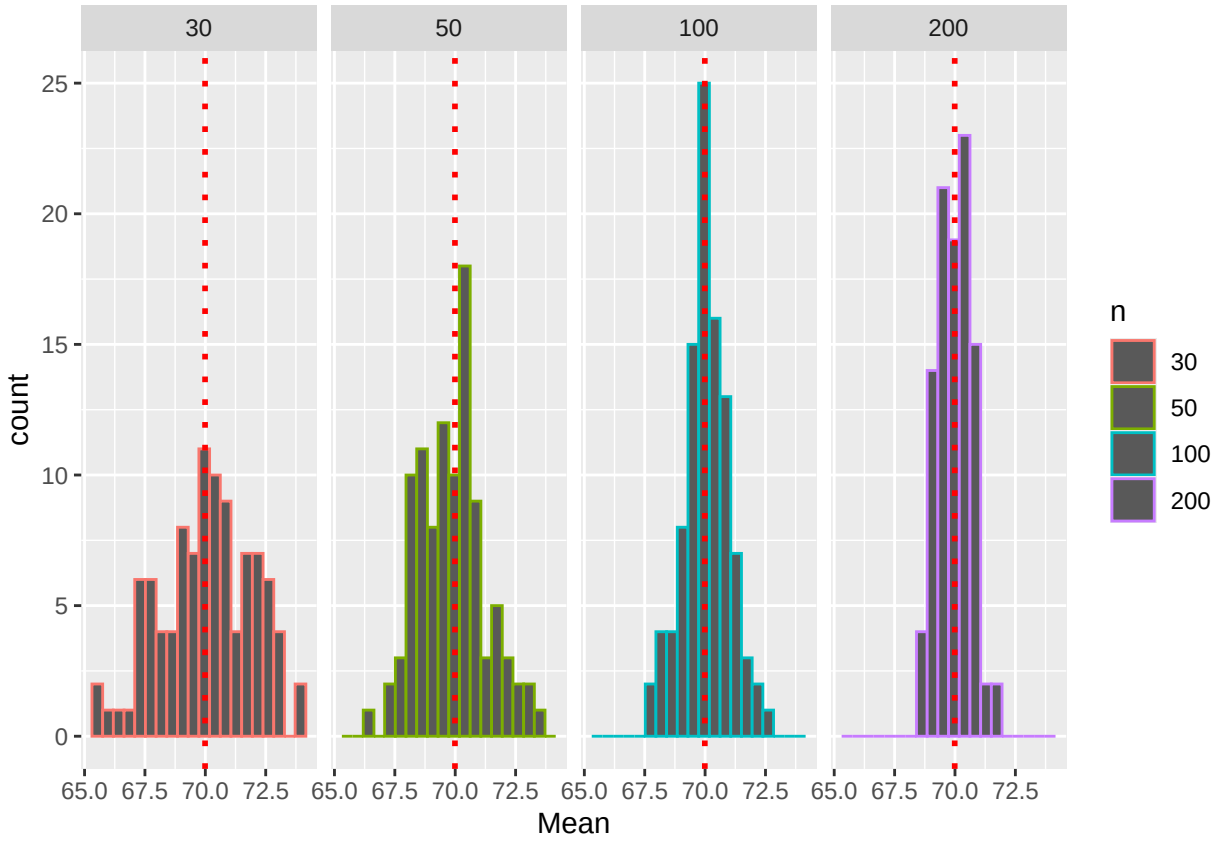
```
## [1] 128.1505
```

The standard deviation of the distribution of an estimate is also called *standard error*, since it reflects the difference between a parameter and the corresponding estimate. For the sample mean of a Normal distribution, it is computed as $\text{se} = s/\sqrt{n}$.

Is there any way to improve the estimate? As suggested by the asymptotic properties, one way to improve a point estimate is to take a larger sample, that is increasing $n$.

```
n_values <- c(30, 50, 100, 200)
vect_mean <- matrix(0,nrow=400, ncol=2)
k=0
for(j in n_values){
  for(i in 1:100){
    x <- rnorm(j, mean = 70, sd=10)
    vect_mean[i+k*100,] <- c(j, mean(x))
  }
  k = k +1
}
vect_mean <- data.frame(vect_mean)
colnames(vect_mean) <- c("n", "Mean")
vect_mean$n <- as.factor(vect_mean$n)
ggplot(vect_mean, aes(x=Mean, color=n))+
  geom_histogram(bins=20) +
  facet_grid(~n)+
  geom_vline(xintercept = mean(vect_mean$Mean), col = "red", linetype = "dotted", size = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




Increasing the size of the sample decreases the variability of the sampling distribution of the estimate. Therefore, a larger sample size results in a more reliable point estimate of the population parameter. The distribution of the sample mean is roughly bell-shaped (normally distributed).

Other examples of estimators for $\mu$ and $\sigma^2$ are the median, and interquartile range or median absolute deviation (MAD), respectively.

**Remark**:

(i) The sample is a subset of individuals drawn from the population.

(ii) $X_i$ are independent and identically distributed (i.i.d.), following the distribution of the model described by the population: $X_i \sim f(x;\theta)$.

(iii) An estimate can be also called a statistic. A *statistic*, denoted as $T_n = T(X_1, X_2, \ldots, X_n)$, is any real-valued function (transformation) of the random sample $X = (X_1, X_2, \ldots, X_n)$ that does not depend on any other unknown quantities. Since a statistic is a random variable, it has a distribution.

(iv) In statistical estimation, the sample is used to compute a single value, often referred to as *(point) estimate*, which serves as the *best guess* of the unknown population parameter.

(v) The *3-sigma-rule* can be also considered for the estimator distribution using the estimated standard error.

## Confidence intervals

Confidence intervals are a tool for quantifying the uncertainty related to the estimated value of a parameter. Let consider again $X$ follows $f(x;\theta)$, and consider a random sample $(X_1, \ldots, X_n)$ with an estimator $T_n = T(X_1, \ldots, X_n)$. Let $t_n = t(x_1, \ldots, x_n)$ be the estimate of $\theta$. In reality, no matter how accurate the estimator $T_n$ is, a single number $t_n$ (i.e., point estimate), does not provide any indication on the probabilities that the

estimate takes on a value close or equal to the *true* parameter value $\theta$. The $\alpha$ *confidence intervals* overcome this inconvenience and allow us to establish a range of plausible estimates associated with a fixed level of confidence.

In other word, it allows us to define a level of confidence $\alpha$ in our population parameter estimate gleaned from a sample.

Confidence interval estimation is often calculated based on the point estimate by adding and subtracting a value known as the *margin of error*

$$[\text{Point Estimate} - \text{Margin of Error}; \text{Point Estimate} + \text{Margin of Error}].$$

We want an interval that will bracket the true value of the parameter in $(1-\alpha)\%$ of the instances of an experiment that is repeated a large number of times.

Consider the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and assume the variance $\sigma_X^2$ is known. For this estimator, we know that the expectation and the variance are given as

$$\mathbb{E}[\bar{X}] = \mu_X, \qquad \mathbb{E}[(\bar{X} - \mu_X)^2] = \frac{\sigma_X^2}{n}.$$

Considering the sample mean, this estimator is asymptotically normal, that is

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

Using this information, we can construct the $\alpha$ confidence intervals, with $\alpha \in (0,1)$. Given a value $\alpha$, we aim to find $u_\alpha$ such that $P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$. Using the symmetry of the normal distribution we get

$$P\left(z_{\alpha/2} \leq \frac{\bar{X} - \mu_X}{\sigma_X} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

where $z_b$ denotes the quantile of level $b$ of a normal standard distribution. Rewriting the terms with respect $\mu$, we have that the confidence interval for $\mu$ at level $1-\alpha$ is

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

where $\hat{\mu} = \bar{x}$ is the computed sample mean estimate.

Considering the standard normal distribution, the displayed quantiles correspond to

```
alpha <- 0.05
1 - alpha
```

```
## [1] 0.95
```

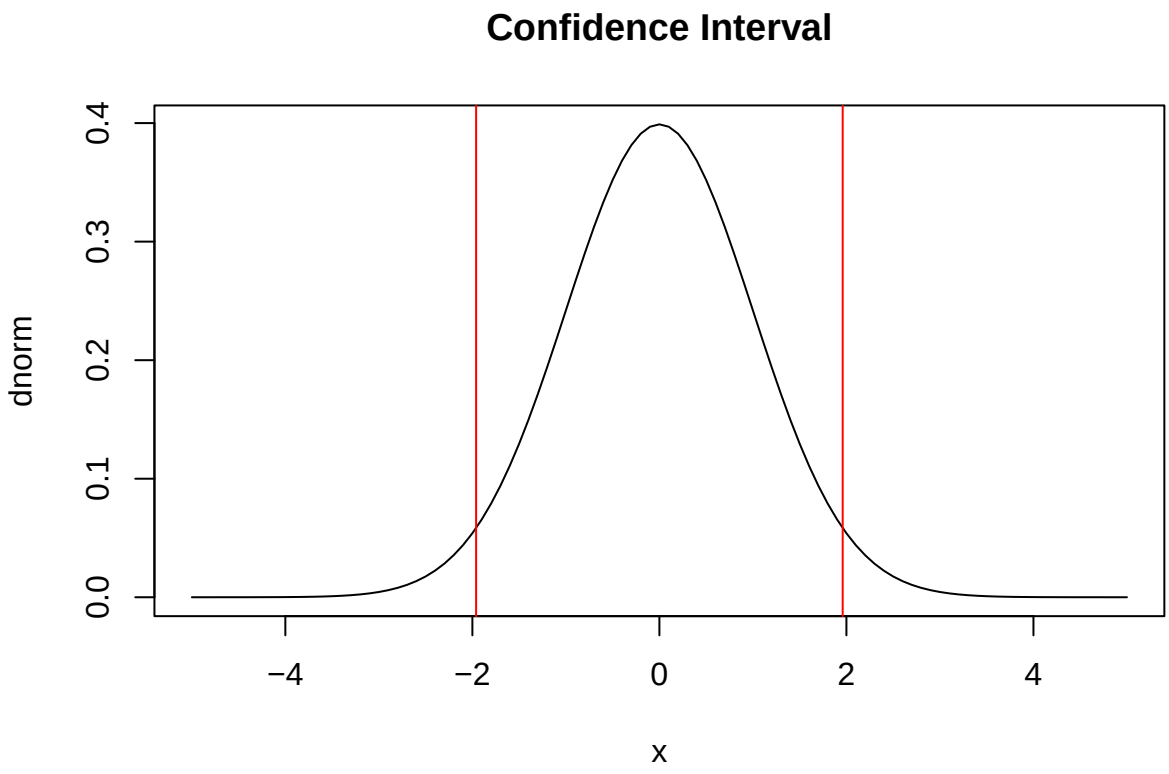
```
qnorm(alpha / 2) # -u_alpha
```

```
## [1] -1.959964
```

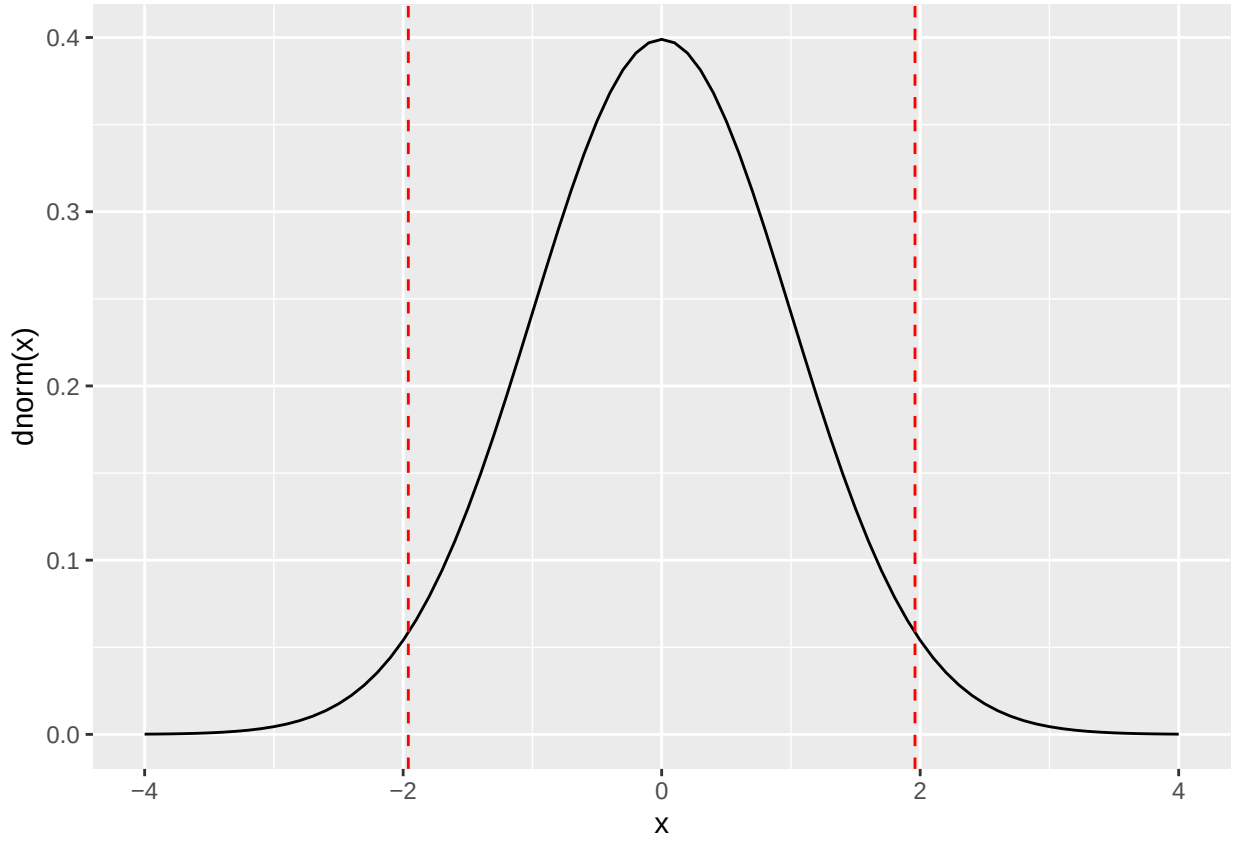
```
qnorm(1 - alpha / 2) # u_alpha
```

```
## [1] 1.959964
```

```
# Plot confidence interval
plot(dnorm, from = -5, to = 5, main = "Confidence Interval")
abline(v = qnorm(alpha / 2), col = "red")
abline(v = qnorm(1 - alpha / 2), col = "red")
```




```
# Using ggplot
x <- seq(-4, 4, 0.1)
ggplot(mapping = aes(x=x, y=dnorm(x)))+
  geom_line() +
  geom_vline(xintercept=qnorm(alpha/2), color="red", linetype="dashed") +
  geom_vline(xintercept=qnorm(1 - alpha/2), color="red", linetype="dashed")
```




In other words, if we estimate a lot of times the confidence intervals from random samples, the $(1-\alpha)$ of the times, they include the true value of the population parameter. For example

```
set.seed(1234)
n <- 100
x <- rnorm(n, mean = 70, sd = 10)
mu_x <- mean(x)
round(mu_x,2)
```

```
## [1] 68.43
```

```
alpha <- 0.05
# we assume to know that the standard deviation is equal to 10
sigma <- 10
# Then the confidence interval is
round(c(mu_x - qnorm(1 - alpha/2)*sigma/sqrt(n),
  mu_x + qnorm(1 - alpha/2)*sigma/sqrt(n)),2)
```

```
## [1] 66.47 70.39
```

With this result, we can state that *"We are confident at 95% that the true mean is inside the interval"*. But beware, people will sometimes state this as *"... there is a 95% chance that the population mean falls between such and such values ..."* which is problematic since it implies that the population mean is a random variable when in fact it's not. The confidence interval reminds us that the chances are in the sampling and not the population parameter.

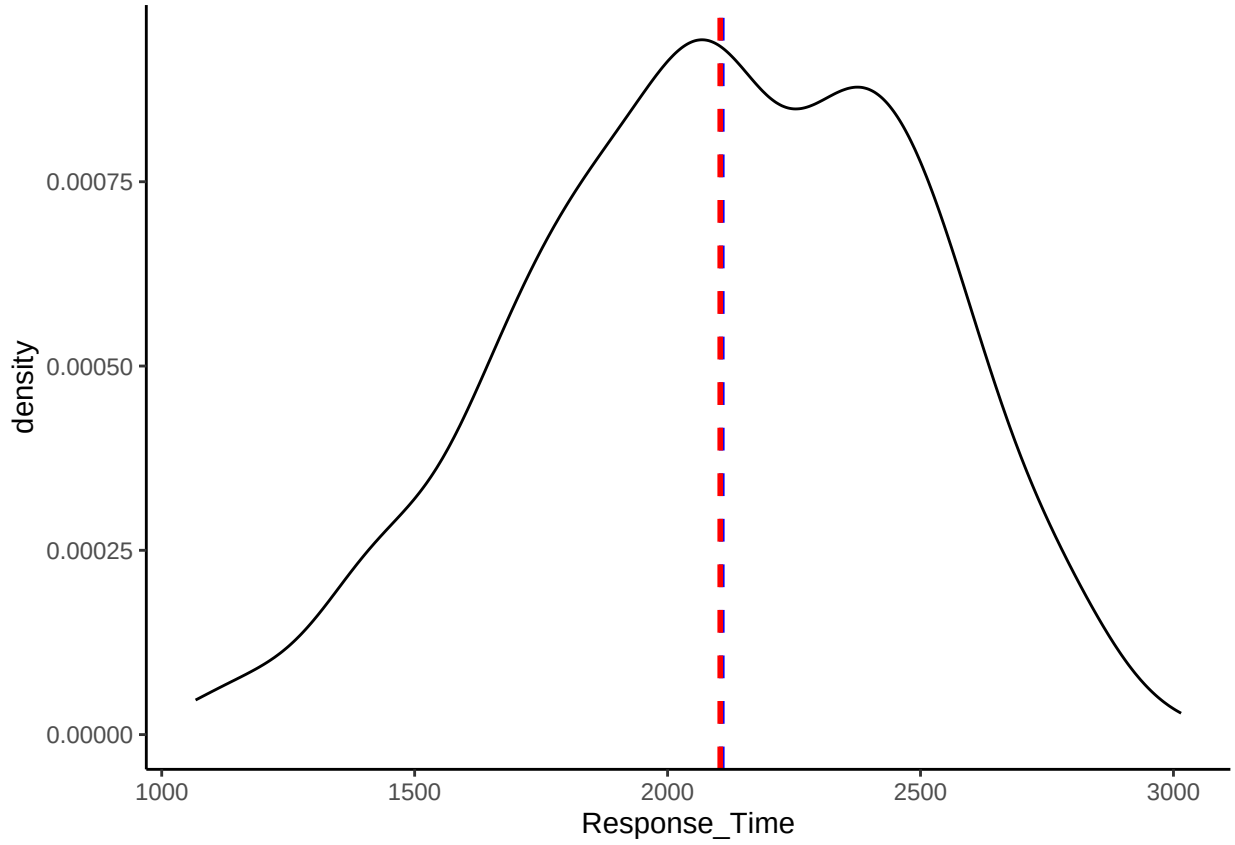
Let's consider the data set generated in the previous generated example, saved in `"first_dataframe.csv"`.

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

### Mean with known variance

We can see the population distribution of the `Response_Time` with its mean and median

```r
ggplot(dat, aes(x=Response_Time))+
  geom_density()+
  theme_classic()+
  geom_vline(xintercept = median(dat$Response_Time), color="blue", linetype="dashed", size=1) +
  geom_vline(xintercept = mean(dat$Response_Time), color="red",linetype="dashed", size=1)
```




```r
x_bar <- round(mean(dat$Response_Time),2)
s_x <- round(sd(dat$Response_Time),2)
print(paste0(x_bar, " (", s_x, ")"))
```

```
## [1] "2104.12 (382.06)"
```

```r
print(paste0(round(median(dat$Response_Time),2), " (", round(IQR(dat$Response_Time),2), ")"))
```

```
## [1] "2107.06 (566.5)"
```

Let's compute the 95% confidence interval, considering the variance known.

```r
mean_R <- mean(dat$Response_Time)
# We round the standard deviation just for supposing that the variance is known.
```

```r
sd_R <- round(sd(dat$Response_Time))
n <- nrow(dat)
alpha <- 0.05
lb <- round(mean_R - qnorm(1-alpha/2)*sd_R/sqrt(n), digits = 3)
rb <- round(mean_R + qnorm(1-alpha/2)*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 [2070.637;2137.603]"
```

With this result we can state: *We are confident at 95% that the true mean is inside this interval.*

**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.

```r
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
x_bar <- mean(thickness)
n <- length(thickness)
u_alpha <- qnorm(1 - alpha / 2)
conf_interval <- c(x_bar - u_alpha * sqrt(Var / n), x_bar + u_alpha * sqrt(Var / n))
conf_interval
```

```
## [1] 13.90002 14.02398
```

By changing the value of $\alpha$, intervals with different confidence levels can be obtained.

```r
alpha <- 0.01
u_alpha <- qnorm(1 - alpha / 2)
conf_interval <- c(x_bar - u_alpha * sqrt(Var / n), x_bar + u_alpha * sqrt(Var / n))
conf_interval
```

```
## [1] 13.88055 14.04345
```

**Remark:** If the variance is unknown, but the sample size is large, the required variance for constructing the confidence interval can be approximated using the sample variance.

## Mean with Unknown Variance

Now, let us consider the case where the variance $\sigma^2$ is unknown. In this case, we estimate $\sigma^2$ by using a proper estimator, that could be

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

where $s$ is the sample standard deviation. The variance of the mean $\sigma_X^2 = \sigma^2/n$ then it can be estimated as $s_X^2 = s^2/n$. This introduces an additional level of uncertainty and the complication that, for small samples the sample estimate of standard deviation tend to be biased compared to the true population standard deviation.
In the case of small sample sizes and unknown standard deviation, instead of using the normal distribution, it is appropriate to use the quantiles of the *Student's t-distribution*, characterized by a single parameter of degrees of freedom $df = n - 1$. For increasing $df$, the t-distributions becomes more and more similar to the standard normal distribution.

The confidence interval is computed as in the previous case using the sample standard deviation

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - t_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right].$$

**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)
fatture
```

```
##  [1] 108.98 152.22 111.45 110.59 127.46 107.26  93.32  91.97 111.56  75.71
## [11] 128.58 135.11
```

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

Note: In practice, the assumption that data follows a normal distribution must be verified. There are graphical and analytical methods to assess the "normality" of the data distribution.

We compute the sample mean $\bar{X}$ and the sample variance $S^2$:

```
mx <- mean(fatture)
s2 <- var(fatture)
mx
```

```
## [1] 112.8508
```

```
s2
```

```
## [1] 432.5565
```

In this case, we do not know the true value of $\sigma_X^2$. Given the sample size $n$ and sample variance $S^2$, 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. To construct the confidence interval for the sample mean, we calculate the quantile of a $t$-distribution with $n-1$ degrees of freedom. In our example:

```
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
```

In R, there are functions to calculate confidence intervals directly, while performing hypothesis tests. For example, the `t.test` function allows you to input the sample and the confidence level with conf.level. For instance:

```
t.test(fatture, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  fatture
## t = 18.796, df = 11, p-value = 1.039e-09
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##   94.2040 131.4977
## sample estimates:
```

```
## mean of x
##  112.8508
```

The displayed confidence interval matches with the previously computed interval.

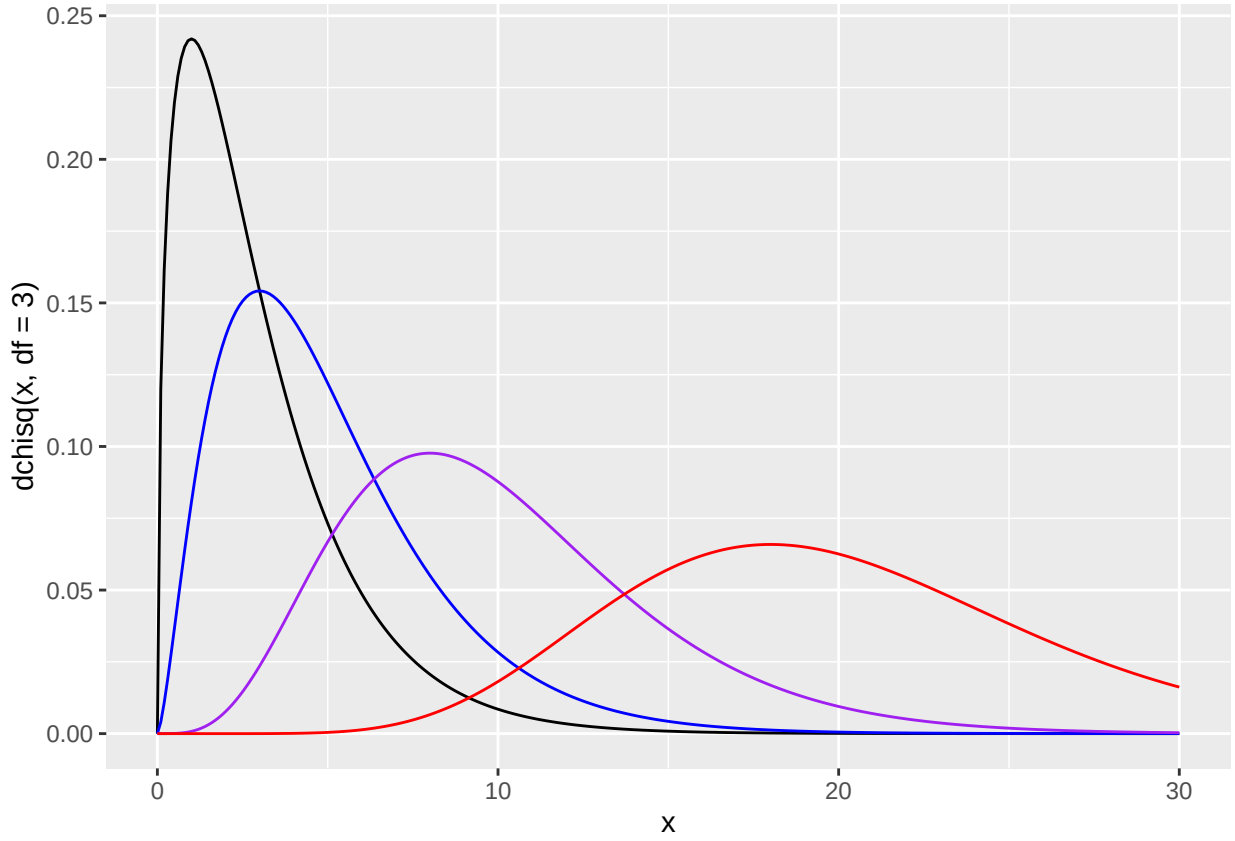
### Variance parameter

For the sample variance, we know that:

$$\frac{(n-1)s^2}{\sigma_X^2} \sim \chi^2_{n-1}$$

where $\chi^2_{n-1}$ denotes the chi-squared distribution with $n-1$ degrees of freedom.

```
# The chi-squared distribution
x=seq(0,30,0.1)
ggplot(mapping = aes(x=x)) +
  geom_line(mapping = aes(y=dchisq(x, df = 3)))+
  geom_line(mapping = aes(y=dchisq(x, df = 5)), color = "blue")+
  geom_line(mapping = aes(y=dchisq(x, df = 10)), color = "purple")+
  geom_line(mapping = aes(y=dchisq(x, df = 20)), color = "red")
```




To construct a confidence interval for the sample variance, we use the quantiles of a chi-squared distribution

$$IC_{1-\alpha}(\sigma^2) = \left[ s^2 \frac{df}{\chi^2_{\alpha/2,df}}, s^2 \frac{df}{\chi^2_{1-\alpha/2,df}} \right].$$

For example:

```r
x <- c(0.39, 0.68, 0.82, 1.35, 1.38, 1.62, 1.70, 1.71, 1.85, 2.14, 2.89, 3.69)
s2 <- var(x)
s2
```

```
## [1] 0.8501727
```

Now, we compute the confidence interval for the variance:

```r
conf_level <- 0.95
alpha <- 1 - conf_level
n <- length(x)
interval <- c((n-1) * s2 / qchisq(1 - alpha / 2, n - 1),
              (n-1) * s2 / qchisq(alpha / 2, n - 1))
interval
```

```
## [1] 0.4266368 2.4508692
```

## Proportions

So far, we have considered confidence intervals and hypothesis tests for the mean and variance of Gaussian random variables. Now, let us consider the case of a Bernoulli and binomial distributions.

The proportion is the most important descriptive statistic for a categorical variable.
Given the random variable $X \sim Ber(\pi)$ and the sample $X_1, \dots, X_n$ from $X$, the sample mean or *sample proportion* represents the proportion of successes in the random sample. It is computed as

$$\hat{p} = \frac{\text{number of observations in the category}}{n}.$$

Note that $\hat{p}$ represents the proportion of success *on average*. The standard error of $\hat{p}$ is given as

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

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. The confidence interval for the proportion of a population is

$$IC_{1-\alpha}(p) = \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

For example

```r
group <- c("Case", "Control", "Control", "Case", "Control", "Case", "Control", "Case", "Control", "Cont
n <- length(group)
p_control <- length(group[which(group == "Case")])/n
p_control
```

```
## [1] 0.4545455
```

We computed the proportion of subject in the Control group and now we compute the related confidence interval

```r
alpha <- 0.05
se <- sqrt(p_control * (1 - p_control) / (n-1))
u_alpha <- qnorm(1 - alpha/2)
interval <- c(p_control - u_alpha*se,
              p_control + u_alpha*se)
interval
```

```
## [1] 0.2415814 0.6675095
# This corresponds to the Wald method
```

We can do this using the `prop.test` function. This function requires a table with the counts.

```
prop.test(table(group), conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  table(group), null probability 0.5
## X-squared = 0.045455, df = 1, p-value = 0.8312
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2507068 0.6732606
## sample estimates:
##         p
## 0.4545455
```

**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.

## Comparison of Means of Two Normal Populations

Until now, we have considered the case of a single population. Once the measures of central tendency and variability have been calculated, we have tested the reliability of these estimates using confidence intervals. Now, we will consider cases where there are two or more populations, and the goal is to perform inference by comparing them.

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 $X_2 \sim N(\mu_2, \sigma_2)$. We can consider the difference $\mu_1 - \mu_2$ and the results seen for testing the mean estimate of a normally distributed random variable apply.

As in the case of a single mean $\mu$, if we use an estimate of $\sigma_1^2$ and $\sigma_2^2$ 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 second case, while the in the first 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)
```

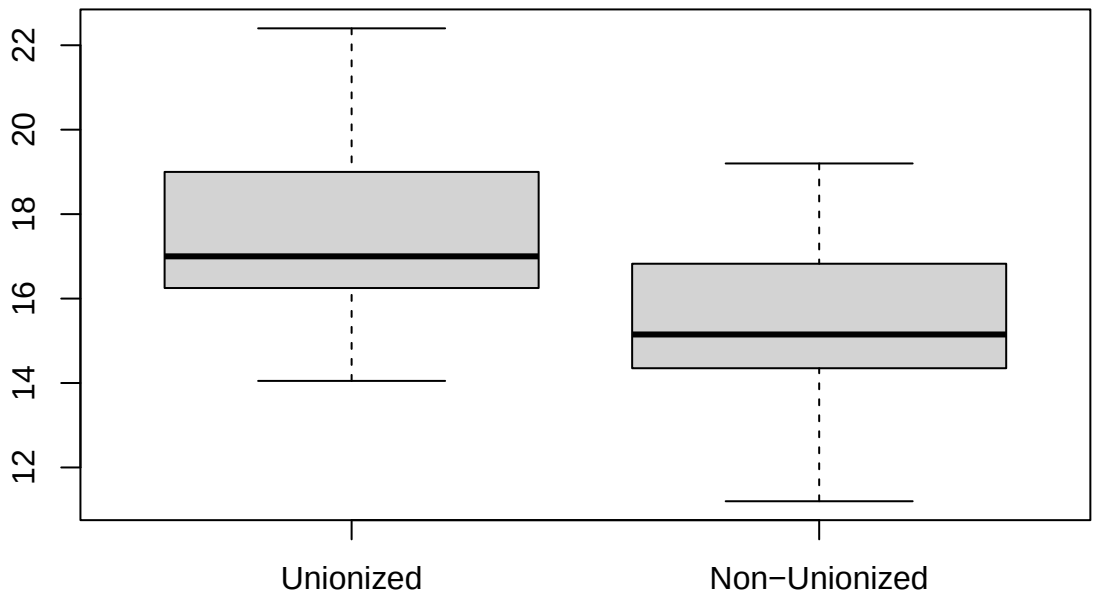
```
## [1] 15
```

```
length(non_iscritte)
```

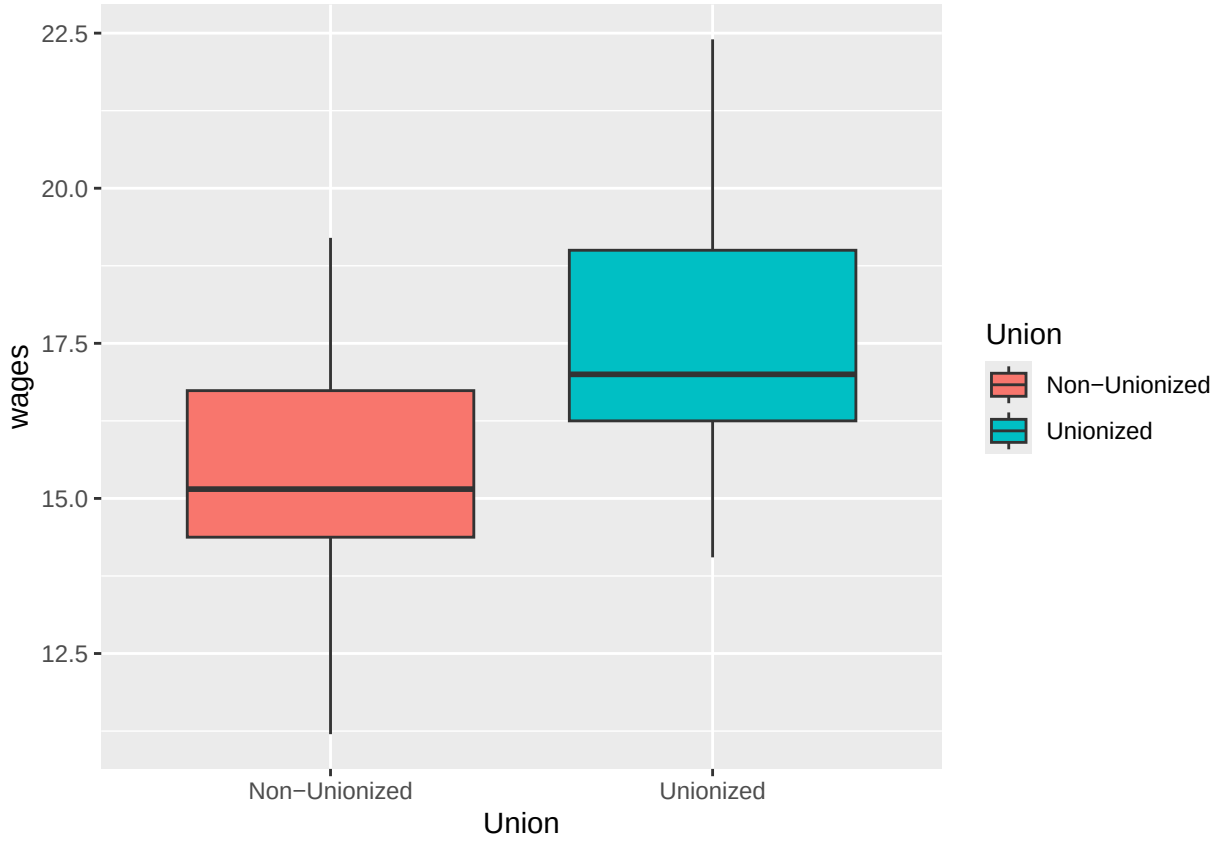
```
## [1] 20
```

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 $\mu_1$ and $\mu_2$, 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
```

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)
```

We now compute a confidence interval for the difference in means

```
interval <- c(
  mu1 - mu2 - qt(1 - alpha / 2, n + m - 2) * sqrt(s * (1 / n + 1 / m)),
  mu1 - mu2 + qt(1 - alpha / 2, n + m - 2) * sqrt(s * (1 / n + 1 / m))
)
interval
```

```
## [1] 0.963342 3.389991
```

**Paired samples** Let's now see the last case, i.e., we have a paired sampling design. A random sample is selected from one population, and each statistical unit provides two observations. Each pair of observations shares a common element: the individual on whom the measurements were taken. The two observations on the same subject are not independent because they are influenced by common individual factors.

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 consider the random variable "difference" as:

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

.

The parameter of interest in this case becomes the difference of means $\mu_D = \mu_1 - \mu_2$. We then consider the estimator $\bar{D} = \sum_{i=1}^{n} \frac{X_{1i} - X_{2i}}{n}$. Since also here we do not know $\sigma_D^2$, we plug-in the estimator

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2.$$

Finally the confidence intervals at level $1 - \alpha$ is computed as

$$IC_{1-\alpha}(\mu_d) = \left[ \bar{D} - t_{\alpha/2, n-1} \frac{\sigma_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2, n-1} \frac{\sigma_D}{\sqrt{n}} \right]$$

As you can note, this is equal to the confidence interval of the simple t-test for the mean $\mu$ but considering the new random variable $D$ instead of $X$.

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
```