Statistics with R (part 2)

Lecture 4

Joaquin Cavieres

1. Estimation, hypotesis testing and confidence intervals

In this section we will see three important statistical concepts:

- Estimation of unknown population parameters
- Hypothesis testing
- Confidence intervals

also, we will discuss them in the simple context of inference about an unknown population mean and discuss some results using R.

1.1. Estimation of the population mean

Estimators and estimates

- Estimators can be defined as functions of sample data drawn from an unknown population. Estimators are random variables because they are functions of random data
- Estimates are numeric values computed by estimators based on the sample data. Estimates are nonrandom numbers.

As example, consider the salary of geographers in Germany as a random variable, thus we can denote the salary as Y. If we are interested in to calculate the mean and μ_Y of the Y. To calculate μ_Y we should interview tp all geographers working currently in Germany, but the problem with this is the cost to do this survey. But, we can calculate the mean of a random sample $Y_i, ..., Y_n$ for a $i = 1, ..., n$ and estimate μ_Y from that random sample. This would be:

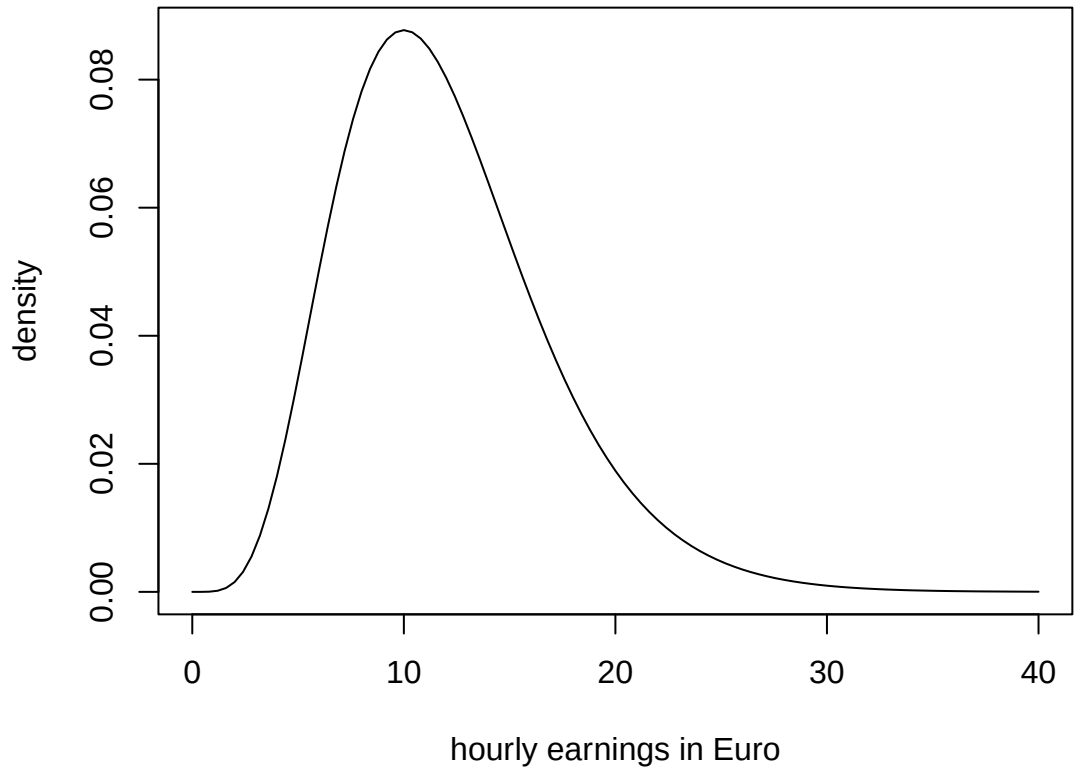
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{1}$$

which yields the mean of the sample Y. Well, if we can consider a sample to estimate the mean of the population, is Y_1 a good estimator? Let's see, consider the following:

$$Y \sim \chi^2_{12}$$

which is not too unreasonable as hourly income is non-negative and we expect a earn between 5 - 15 euros by hour. Moreover, it is common for income distributions to be skewed to the right — a property of the χ^2_{12} distribution.

```
# plot the chi_12^2 distribution
curve(dchisq(x, df=12),
      from = 0,
      to = 40,
      ylab = "density",
      xlab = "hourly earnings in Euro")
```



We now take a sample of size $n = 200$ observations and consider the first observation Y_1 as an estimate for μ_Y

```r
# set seed for reproducibility
set.seed(123)

# sample from the chi_12^2 distribution, use only the first observation
chi_samp <- rchisq(n = 200, df = 12)
chi_samp[1]
```

```
## [1] 8.528203
```

The estimate is 8.52 and its not so far away from $\mu_Y = 12$ (the mean of a $\chi^2_k = k$), however the estimator of Y_1 does not consider a lot of information and its variance is $\text{Var}(Y_1) = \text{Var}(Y) = 2 \cdot 12 = 24$ (here the variance of a $\chi^2_k = 2 * k$, where k to the "degrees of freedom" (12 in our case))

What is a good estimator of an unknown parameter?

We can answer this question in the following topic.

1.2. Bias, consistency and efficiency

An estimator must have the following characteristics:

a) Unbiasedness

Let's assume that $\hat{\mu}_Y$ is the mean estimator of a sampling distribution for the mean of the population μ_Y, that is:

$$E(\hat{\mu}_Y) = \mu_Y,$$

then the estimator $\hat{\mu}_Y$ is unbiased for μ_Y. The term "unbiased" means:

$$E(\hat{\mu}_Y) - \mu_Y = 0$$

b) Consistency

If we want to find an estimator for μ_Y be the most similar to it, an estimator $\hat{\mu}_Y$ should falls withing an acceptable interval around to the true value of μ_Y, in this way the uncertainty of the estimator must decrease as the number of observations in the sample grows. For the above we can write:

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y.$$

c) Variance and efficiency

Also, another property of the estimator is that it must be efficient. For example, if we have two estimators, let's say $\hat{\mu}_Y$ and $\tilde{\mu}_Y$, then we should expect:

$$E(\hat{\mu}_Y) = E(\widetilde{\mu}_Y) = \mu_Y,$$

for some sample of size "n". However,

$$\text{Var}(\hat{\mu}_Y) < \text{Var}(\widetilde{\mu}_Y).$$

So, we should prefer to use $\hat{\mu}_Y$ because it has a lower variance than $\tilde{\mu}_Y$.

2. Hypothesis tests concerning the population mean

The idea before of this topic is to cover the concepts of hypothesis testing (briefly) and make some experiments in R. Only we will consider the inference about an unknown population mean.

Generally, in a test of significance, we want to know if our belief (hypothesis) can be contrasted by empirical results. Our main objective is to answer a basic question by a "yes" or "no". In statistics, an hypothesis test deal with two different results:

- The *null hypothesis*, $\boldsymbol{H}_0$, is our proposal hypothesis.
- An *alternative hypothesis*, $\boldsymbol{H}_1$, that is thought to hold if the null hypothesis is rejected.

The $\boldsymbol{H}_0$ (*null hypothesis*) proposing that the mean of Y is equals to μ_Y is written as:

$$\boldsymbol{H}_0 : E(Y) = \mu_Y,$$

On the contrary, the alternative hypothesis ($\boldsymbol{H}_1$) for this example could be:

$$\boldsymbol{H}_1 : E(Y) \neq \mu_Y,$$

exposing that "$E(Y)$ is different to the value under $\boldsymbol{H}_1$"

2.1. *p*-value

Ok, if our hypothesis is true (the null hypothesis $\boldsymbol{H}_0$), we can define the p-value as the probability that a particular statistical measure, such as the mean or standard deviation, of an assumed probability distribution will be greater than or equal to (or less than or equal to in some instances) observed results.

If we are working with the population mean and the sample mean. then we can write a mathematical expression for this as:

$$p\text{-value} = P_{\boldsymbol{H}_0}\left[|\overline{Y} - \mu_Y| > |\overline{Y}_{act} - \mu_Y|\right], \tag{2}$$

where $\overline{Y}_{act}$ is the sample mean for the available data.

The p- value can be calculated knowing the sampling distribution of $\overline{Y}$ (a random variable), when $\boldsymbol{H}_1$ is true, however, almost always the sampling distribution is unknown. To deal with this problem, the uses of the central limit theorem (https://en.wikipedia.org/wiki/Central_limit_theorem), we can approximate for a large sample:

$$\overline{Y} \approx \mathcal{N}(\mu_Y, \sigma^2_{\overline{Y}}) \quad , \quad \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n},$$

so $\boldsymbol{H}_0 : E(Y) = \mu_Y$ is true. Furthermore, it follows for large "n" that:

$$\frac{\overline{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

In summary, and for a large sample (big number of observations), the p-value can be calculated without knowledge of the exact sampling distribution of $\overline{Y}$ using the normal approximation.

2.2. Calculating the p-value when the standard deviation is known

When we assume that σ_Y is kown, then we can write (2) as:

$$p\text{-value} = P_{H_0}\left[\left|\frac{\overline{Y} - \mu_Y}{\sigma_{\overline{Y}}}\right| > \left|\frac{\overline{Y}^{act} - \mu_Y}{\sigma_{\overline{Y}}}\right|\right] \tag{3}$$

$$= 2 \cdot \Phi\left[-\left|\frac{\overline{Y}^{act} - \mu_Y}{\sigma_{\overline{Y}}}\right|\right]. \tag{3.3}$$

where the p-value is the area in the tails of the normal distribution:

$$\pm\left|\frac{\overline{Y}^{act} - \mu_Y}{\sigma_{\overline{Y}}}\right| \tag{3.4}$$

For example, let see this in R:

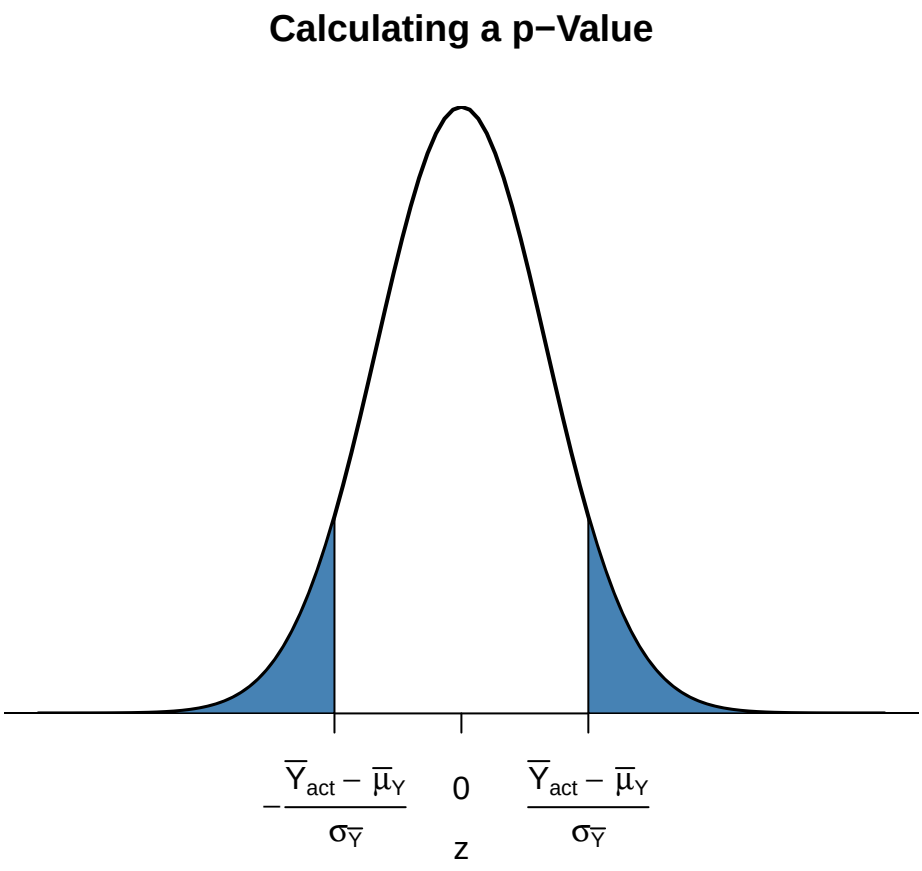
```
# plot the standard normal density on the interval [-4,4]
curve(dnorm(x),
      xlim = c(-5, 5),
      main = "Calculating a p-Value",
      yaxs = "i",
      xlab = "z",
      ylab = "",
      lwd = 2,
      axes = "F")

# add x-axis
```

```
axis(1,
     at = c(-1.5, 0, 1.5),
     padj = 0.75,
     labels = c(expression(-frac(bar(Y)["act"]~-~bar(mu)[Y], sigma[bar(Y)])),
                0,
                expression(frac(bar(Y)["act"]~-~bar(mu)[Y], sigma[bar(Y)]))))

# shade p-value/2 region in left tail
polygon(x = c(-6, seq(-6, -1.5, 0.01), -1.5),
        y = c(0, dnorm(seq(-6, -1.5, 0.01)),0),
        col = "steelblue")

# shade p-value/2 region in right tail
polygon(x = c(1.5, seq(1.5, 6, 0.01), 6),
        y = c(0, dnorm(seq(1.5, 6, 0.01)), 0),
        col = "steelblue")
```



2.2. Sample variance, sample standard deviation and standard error

As generally we don't have information about the σ_Y^2, we have to estimate it. So, we can do it using the sample variance as following:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{4}$$

and the standard deviation as:

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{5}$$

.

In R we can show as s_Y is a consistent estimator for σ_Y when:

$$s_Y \xrightarrow{p} \sigma_Y.$$

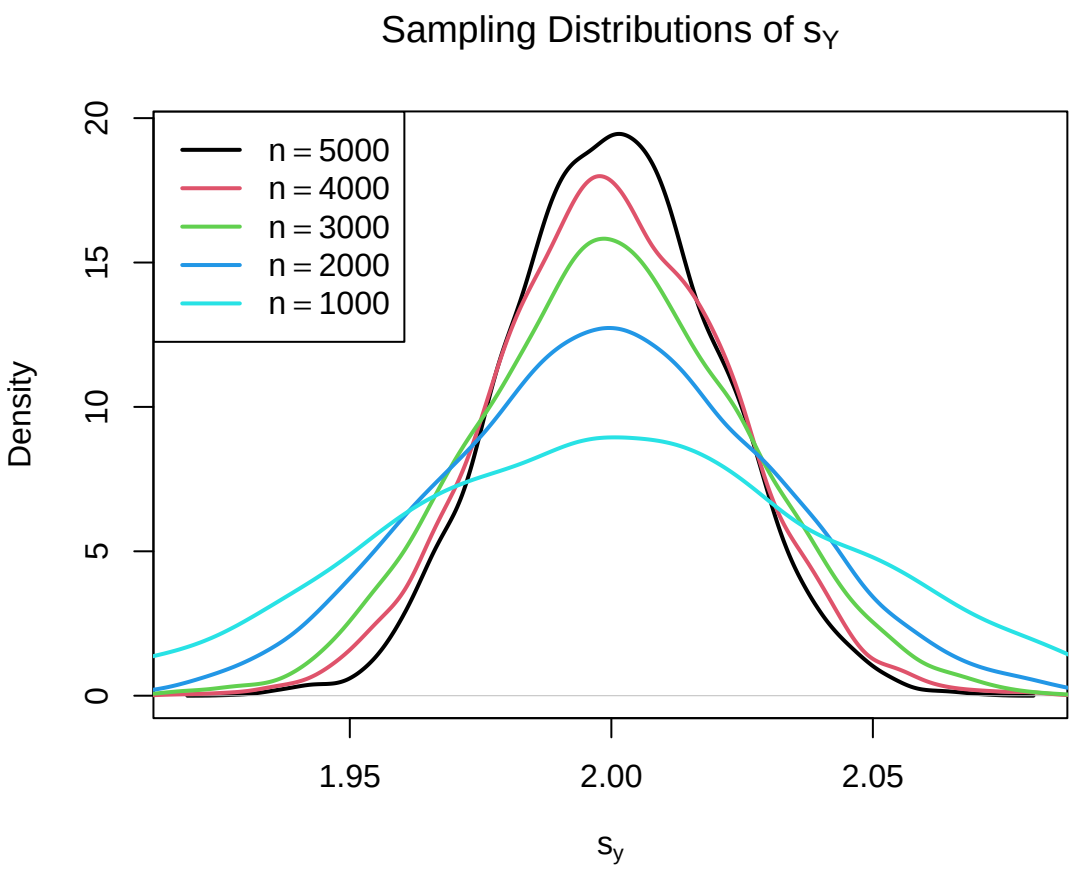
For this, we have to generate a large number of samples $s_Y \xrightarrow{p} \sigma_Y$. assuming $Y \sim \mathcal{N}(5, 4)$, in this way we can estimate σ_Y using s_Y.

```r
# vector of sample sizes
n <- c(5000, 4000, 3000, 2000, 1000)

# sample observations, estimate using 'sd()' and plot the estimated distributions
sq_y <- replicate(n = 5000, expr = sd(rnorm(n[1], 5, 2)))
plot(density(sq_y),
     main = expression("Sampling Distributions of" ~ s[Y]),
     xlab = expression(s[y]),
     lwd = 2)

for (i in 2:length(n)) {
  sq_y <- replicate(n = 5000, expr = sd(rnorm(n[i], 5, 2)))
  lines(density(sq_y),
        col = i,
        lwd = 2)
}

# add a legend
legend("topleft",
       legend = c(expression(n == 5000),
                  expression(n == 4000),
                  expression(n == 3000),
                  expression(n == 2000),
                  expression(n == 1000)),
       col = 1:5,
       lwd = 2)
```



The previous plot shows that the distribution of
s_Y narrows around the true value $\sigma_Y = 4$ as the number of observations increase (n).

The function that estimates the standard deviation of an estimator is called the standard error of the estimator. So, we can define the standard error of $\overline{Y}$ as:

Standard error of $\overline{Y}$

Let consider an *i.i.d* sample of Y. If the mean of Y is estimated consistently by $\overline{Y}$, and its a random variable, then is has a sampling distribution with variance σ^2_Y/n. Therefore, the standard deviation of $\overline{Y}$, expressed as $SE(\overline{Y})$, is an estimator of the standard deviation of $\overline{Y}$ such as:

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = \frac{s_Y}{\sqrt{n}}$$

Let's consider the following example. Wwe have a sample $n = 200$ for *i.i.d* observations from a Bernoulli distributed variable Y. The success probability is $p = 0{,}1$, thus, $E(Y) = p = 0{,}1$ and

$\text{Var}(Y) = p(1-p)$. Considering this we can estimate $E(Y)$ from $\overline{Y}$, that in turn has variance:

$$\sigma^2_{\overline{Y}} = p(1-p)/n = 0{,}1 * (1-0{,}1)/200 = 0{,}00045$$

and standard deviation

$$\sigma_{\overline{Y}} = \sqrt{p(1-p)/n} = 0{,}1 * (1-0{,}1)/200 = 0{,}02.$$

So, we can calculate the standard error of $\overline{Y}$ as:

$$SE(\overline{Y}) = \sqrt{\overline{Y}(1-\overline{Y})/n}.$$

In R we can compute this as following:

```
# draw 1000 samples of size 100 and estimate the mean of Y and
# estimate the standard error of the sample mean

mean_estimates <- numeric(1000)
se_estimates <- numeric(1000)

for (i in 1:1000) {

  s <- sample(0:1,
              size = 200,
              prob = c(0.9, 0.1),
              replace = T)

  mean_estimates[i] <- mean(s)
  se_estimates[i] <- sqrt(mean(s) * (1 - mean(s)) / 200)

}

mean(mean_estimates)
```

```
## [1] 0.099695
```

```
mean(se_estimates)
```

```
## [1] 0.02102655
```

2.3 Confidence intervals for the population mean

It is important to say here that we will never estimate the exact value of the mean population of Y using a sample. For the same, we can calculate intervals where our estimated "lives" value for the population mean.

A simple definition could be: "an interval confidence produces intervals, for repeated samples, which contains the real parameter with a specified probability, the called *confidence level*". Those confidence intervals are calculated using the information containing the sample.

As the samples come from a stochastic process, confidence intervals are random variables as well. Wen can calculate the confidence intervals for the unknown pupulation mean$E(Y)$ as follow:

Confidence intervals for the population mean

A 95 % confidence interval for the parameter μ_Y is a random variable that contains the true paramerer μ_Y in a 95 % of all the random samples. If n is large, then we can use the normal approximation.

$$95\,\%\text{ confidence interval for } \mu_Y = \left[\overline{Y} \pm 1{,}96 \times SE(\overline{Y})\right]$$

So, for example, for a repeated sample, the interval:

$$\left[\overline{Y} \pm 1{,}96 \times SE(\overline{Y})\right]$$

contains the true value of μ_Y with probability of 95 %.

In R is easy to compute the test hypothesis about the mean of a population using the function `t.test()` from the stats package.

```
# set seed
set.seed(123)

# Genererate sample (random) data
sam_data <- rnorm(250, 5, 4)

# What is type of data produced using the t.test() function?
typeof(t.test(sam_data))
```

```
## [1] "list"
```

```
# display the list elements produced by t.test
ls(t.test(sam_data))
```

```
##  [1] "alternative" "conf.int"    "data.name"   "estimate"    "method"
##  [6] "null.value"  "p.value"     "parameter"   "statistic"   "stderr"
```

For this example we only are interested in to compute the 95 % confidence interval for the mean, thus:

```
t.test(sam_data)$"conf.int"
```

```
## [1] 4.496349 5.435167
## attr(,"conf.level")
## [1] 0.95
```

This yields in that the mean is the interval $[4{,}496349 - 5{,}435167]$. However, we know that the true value (mu_Y) is 5.

References

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019). Introduction to Econometrics with R. University of Duisburg-Essen, 1-9.