Review of Statistics with R

Lecture 3

Joaquin Cavieres

1. Probability theory

In this lecture we will see basics concepts of probability theory and how it can be used in R. Almost all the functionalities of statistics can be done by the `stats`, which provides simple functions to make descriptive data analysis and facilitate computations involving a variety of probability distributions.

1.2. Random variables and probability distributions

- The mutually exclusive results of a random process are called the outcomes. 'Mutually exclusive' means that only one of the possible outcomes can be observed.
- We refer to the probability of an outcome as the proportion that the outcome occurs in the long run, that is, if the experiment is repeated many times.
- The set of all possible outcomes of a random variable is called the sample space.
- An event is a subset of the sample space and consists of one or more outcomes.

All the previous definitions summaries in the mean of "random variable". In summary, a random variable is a numerical summary of random outcomes and can be discrete or continuous:

- Discrete random variables have discrete outcomes, e.g., 0 and 1
- A continuous random variable may take on a continuum of possible values.

1.2. Probability distributions for discrete random variables

A common example of a discrete random variable T is the result of a dice roll, where we can get random samples of size 1 from a set of numbers which are mutually independent outcomes. In this example, the space of the sample is $\{1, 2, 3, 4, 5, 6\}$.

To reproduce this example in R we can use the function `sample` as follow:

```r
sample(1:6, 1)
```

```
## [1] 3
```

Here the probability distribution of the random variable is the result of all possible values of the variable and their probabilities which sum 1. We also can summaries this as:

Outcome	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
CDF	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

Expected value and variance

A simple the definition for the expected value can be: the long-run average value of its outcomes when the number of repeated trials is large. For a discrete random variable, the expected value is computed as a weighted average of its possible outcomes whereby the weights are the related probabilities.

Next we can see a mathematical expression for the expected value in a discrete random variable.

Definition

If a random variable Y takes i possible values, $y_1, \ldots, y_n$, the expected value of Y, let´s say E(Y), is defined as:

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_n p_n = \sum_{i=1}^{n} y_i p_i \tag{1}$$

where p_i is the probability that Y takes on y_i, $\sum_{i=1}^{n} y_i p_i$ is the sum of all the $y_i p_i$ from i to n. The expected value is also called the **mean** of Y or the expectation of Y. The symbol to represent the mean commonly is μ_Y.

Considering the dice roll example, the discrete random variable T take 6 possibles values (results), $t_1 = 1, t_2 = 2, t_3 = 3, ..., t_6 = 6$. If we dice is correctly calibrated, all the possible results occurs with a probability of 1/6, thus, we can calculate the exact value of the expected value for T as:

$$E(T) = 1/6 \sum_{i=1}^{6} t_i = 3{,}5, \tag{2}$$

where $E(T)$ is the average of the numbers from 1 to 6. This same result can be obtained in R using the function `mean`

```r
# mean for the numbers from 1 to 6
mean(1:6)
```

```
## [1] 3.5
```

An example of sampling with replacement is rolling a dice three times in a row.

```r
set.seed(123)

# rolling a dice four times in a row
sample(1:6, 4, replace = TRUE)
```

```
## [1] 3 6 3 2
```

Putting `replace = TRUE` gives a different outcome since we draw with replacement at random.

If we consider a bigger number of trials, for example 500, and calculate the mean, then the result should be the same as if we apply the mathematical expression of the expected value for $T(D) = 3{,}5$

```r
set.seed(123)

# compute the sample mean of 10000 dice rolls
mean(sample(1:6, 500, replace = T))
```

```
## [1] 3.506
```

Variance and standard deviation

Others important statistic are the variance and the standard deviation. They help us to measure the dispersion of a random variable.

Definition

The variance of a discrete random variable can be expressed by the symbol σ_Y^2 and its mathematical expression is:

$$\sigma_Y^2 = \text{Var}(Y) = E\left[(Y - \mu_Y)^2\right] = \sum_{i=1}^{n}(y_i - \mu_y)^2 p_i \tag{3}$$

where $\sum_{i=1}^{k} y_i p_i$ is the sum of all the $y_i p_i$ from i to n. The expected value is also called the mean of Y or the expectation of Y. The standard deviation is denoted by σ_Y and it is simply the root square of σ_Y^2.

We have to do a clarification here. The previous definition of the variance is for the total population and it is not implemented in R, but we have the function `var` which calculate the sample variance:

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

This “sample” variance is different to the “population” variance for a discrete random variable Y, since the last is expressed as:

$$\text{Var}(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_Y)^2, \tag{5}$$

note the difference between the size of the population “N” and the size of the sample “n” in the mathematical expressions for the variances.

The sample variance measures how “n” observations are dispersed around the sample average $\bar{y}$, the $\text{Var}(Y)$ measures the dispersion of the complete population “N” around the mean μ_Y. So, if we use the same example for the dice roll, for the discrete random variable T we have:

$$\text{Var}(T) = \frac{1}{6}\sum_{i=1}^{6}(d_i - 3{,}5)^2 = 2{,}92 \tag{6}$$

which is different of the result obtained in R

```
var(1:6)
```

```
## [1] 3.5
```

1.3. Probability distributions of continuous random variables

A continuous random variable takes only contentious possible values, we can not use the concept of a probability distribution as used for discrete random variables. Instead, the probability distribution of a continuous random variable is summarized by its probability density function (PDF).

Definition: Probabilities

If $f_Y(y)$ is the probability density function of Y, then the probability that Y falls between the interval c and d, where $c < d$ is:

$$P(c \leq Y \leq d) = \int_c^d f_Y(y)\mathrm{d}y. \tag{7}$$

besides, $P(-\infty \leq Y \leq \infty) = 1$, therefore $\int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y = 1$. So, the expected value of Y can be calculated as:

$$E(Y) = \mu_Y = \int y f_Y(y)\mathrm{d}y, \tag{8}$$

and the variance as:

$$\text{Var}(Y) = \sigma_Y^2 = \int (y - \mu_Y)^2 f_Y(y)\mathrm{d}y. \tag{9}$$

Consider a continuous random variable X with PDF:

$$f_X(x) = \frac{3}{x^4}, x > 1. \tag{10}$$

- First we can show analytically that the integral of $f_X(x)$ is 1:

$$
\begin{aligned}
\int f_X(x)\mathrm{d}x =& \int_1^\infty \frac{3}{x^4}\mathrm{d}x \qquad (11)\\
=& \operatorname*{lím}_{t\to\infty}\int_1^t \frac{3}{x^4}\mathrm{d}x \qquad (12)\\
=& \operatorname*{lím}_{t\to\infty} -x^{-3}|_{x=1}^{t} \qquad (13)\\
=& -\left(\operatorname*{lím}_{t\to\infty}\frac{1}{t^3}-1\right) \qquad (14)\\
=& 1 \qquad (15)
\end{aligned}
$$

- So, The expectation of X can be computed as follows:

$$
\begin{aligned}
E(X) = \int x\cdot f_X(x)\mathrm{d}x =& \int_1^\infty x\cdot\frac{3}{x^4}\mathrm{d}x \qquad (16)\\
=& -\frac{3}{2}x^{-2}|_{x=1}^{\infty} \qquad (17)\\
=& -\frac{3}{2}\left(\operatorname*{lím}_{t\to\infty}\frac{1}{t^2}-1\right) \qquad (18)\\
=& \frac{3}{2} \qquad (19)
\end{aligned}
$$

- The variance of X can be expressed as $\mathrm{Var}(X) = E(X^2) - E(X)^2$. As we already computed $E(X)$, we have to calculate $E(X^2)$:

$$
\begin{aligned}
E(X^2) = \int x^2\cdot f_X(x)\mathrm{d}x =& \int_1^\infty x^2\cdot\frac{3}{x^4}\mathrm{d}x \qquad (20)\\
=& -3x^{-1}|_{x=1}^{\infty} \qquad (21)\\
=& -3\left(\operatorname*{lím}_{t\to\infty}\frac{1}{t}-1\right) \qquad (22)\\
=& 3 \qquad (23)
\end{aligned}
$$

In this example we saw that the area under the curve is 1, $E(X) = \frac{3}{2}$ and $\mathrm{Var}(X) = \frac{3}{4}$. The problem here is that this calculus must be done "by hand" and result tedious. Thus, we can do the same but in R using the function `integrate`:

```
# define functions
f <- function(x) 3 / x^4
g <- function(x) x * f(x)
h <- function(x) x^2 * f(x)
```

Now we can use the `integrate` function:

```
# compute area under the density curve
area <- integrate(f,
                 lower = 1,
                 upper = Inf)$value
area
```

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

```
# compute E(X)
EX <- integrate(g,
                lower = 1,
                upper = Inf)$value
EX
```

```
## [1] 1.5
```

```
# compute Var(X)
VarX <- integrate(h,
                  lower = 1,
                  upper = Inf)$value - EX^2
VarX
```

```
## [1] 0.75
```

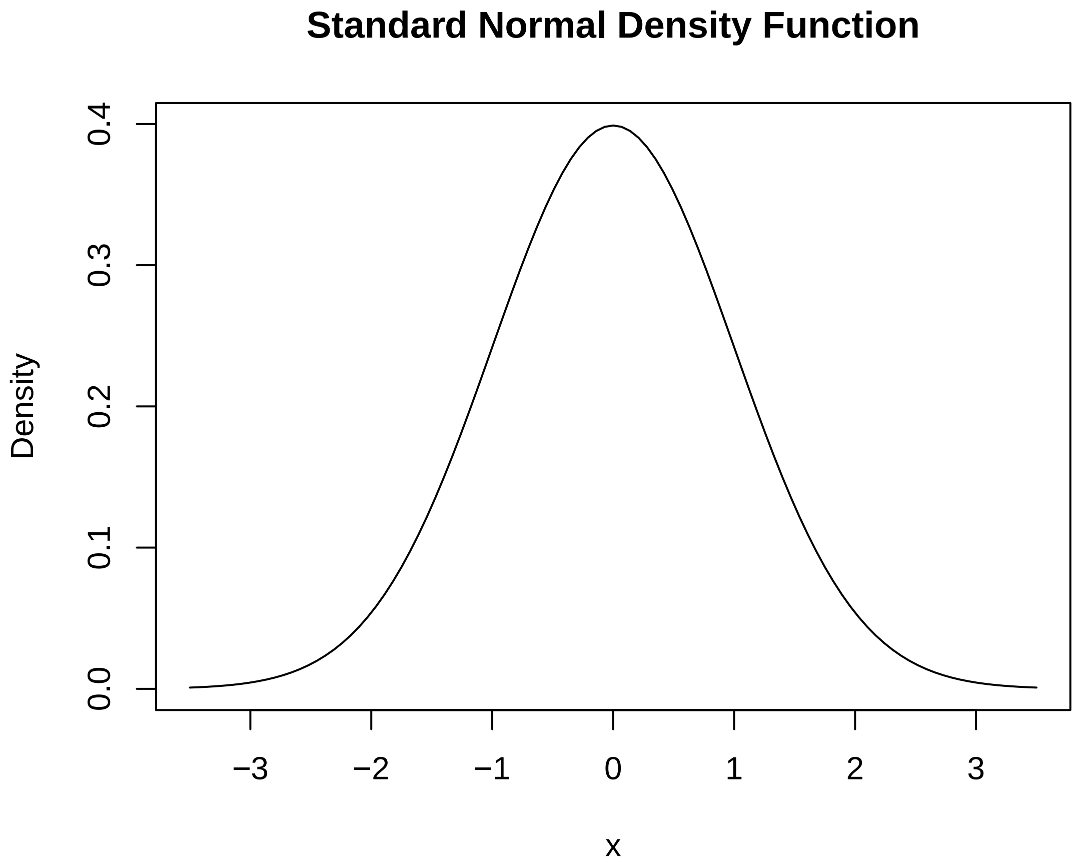
1.3.1. Normal distribution

The normal (Gaussian) distribution is maybe the most important probability functions distribution. It is symmetric and bell-shaped, also it is characterized by its mean μ and its standard deviation σ, and if a random variable has this distribution, then is denoted as $Y \sim \mathcal{N}(\mu, \sigma^2)$. The PDF of the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp -(x-\mu)^2/(2\sigma^2). \tag{24}$$

The **standard normal distribution** is paremeterized by $\mu = 0$ and $\sigma = 1$. In R we can see the shape of the normal distribution as following:

```
# draw a plot of the N(0,1) PDF
curve(dnorm(x),
      xlim = c(-3.5, 3.5),
      ylab = "Density",
      main = "Standard Normal Density Function")
```



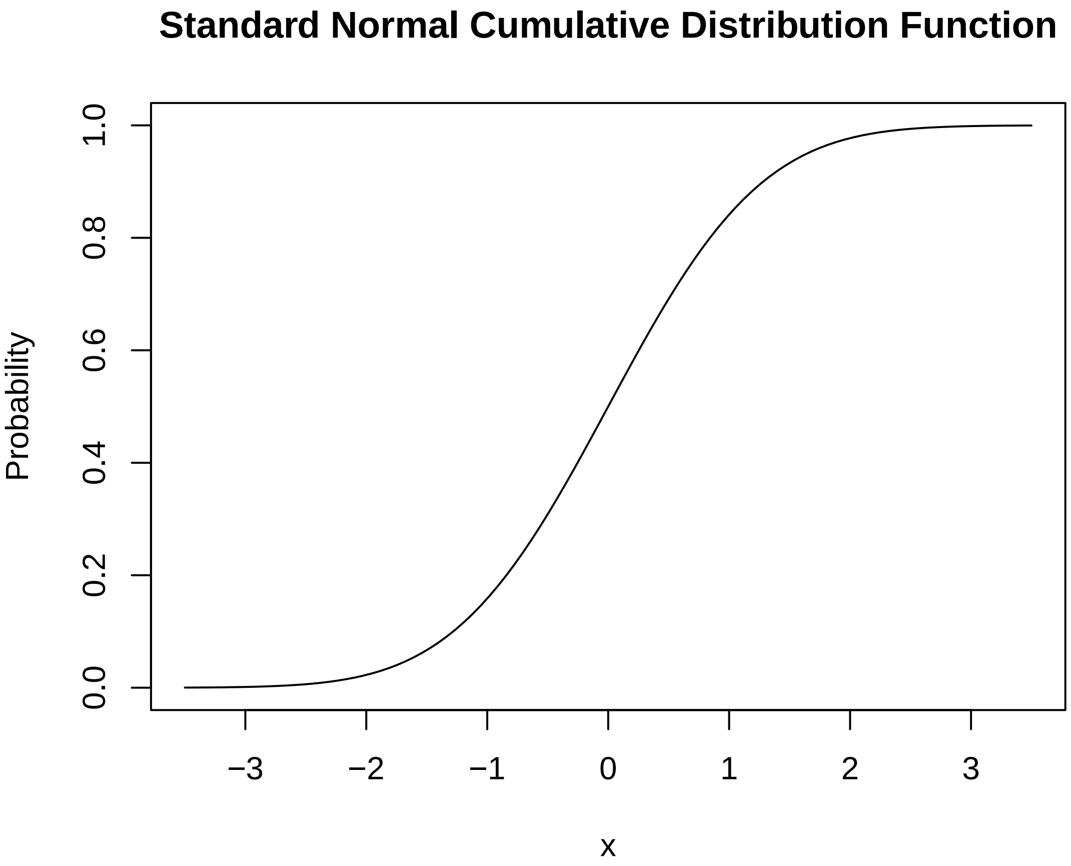
We also can obtain the density at different positions using the function in R `dnorm()`, for example:

```r
# compute density at x=-1.96, x=0 and x=1.96
dnorm(x = c(-1.96, 0, 1.96))
```

```
## [1] 0.05844094 0.39894228 0.05844094
```

In R we also can compute the CDF using the function `curve()`, but its more convenient use the function `pnorm()`:

```r
# plot the standard normal CDF
curve(pnorm(x),
      xlim = c(-3.5, 3.5),
      ylab = "Probability",
      main = "Standard Normal Cumulative Distribution Function")
```



2. 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.

References

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