Random Variables, pt. II

ECON 3640–001

Marcio Santetti

Spring 2022

Motivation

Housekeeping

Notes based on `Keller (2009)`, ch. 8

From discrete to continuous

Last week, we were introduced to **random variables**.

The starting point was to study **discrete** outcomes

- That is, events from experiments that can be **listed**.

However, in many cases one is **not able** to count all possible outcomes from an experiment.

- For instance, how much money would you like to make five years from now?

This answer is probably best given through an **interval**, and not an exact amount.

That is where **continuous random variables** come in.

Continuous random variables

Continuous random variables

A **continuous random variable** can take on **any** real value in an interval.

Going back to the salary example, there is an **infinite** number of possible values one can think of.

- This way, *the probability of each individual value is virtually 0*.

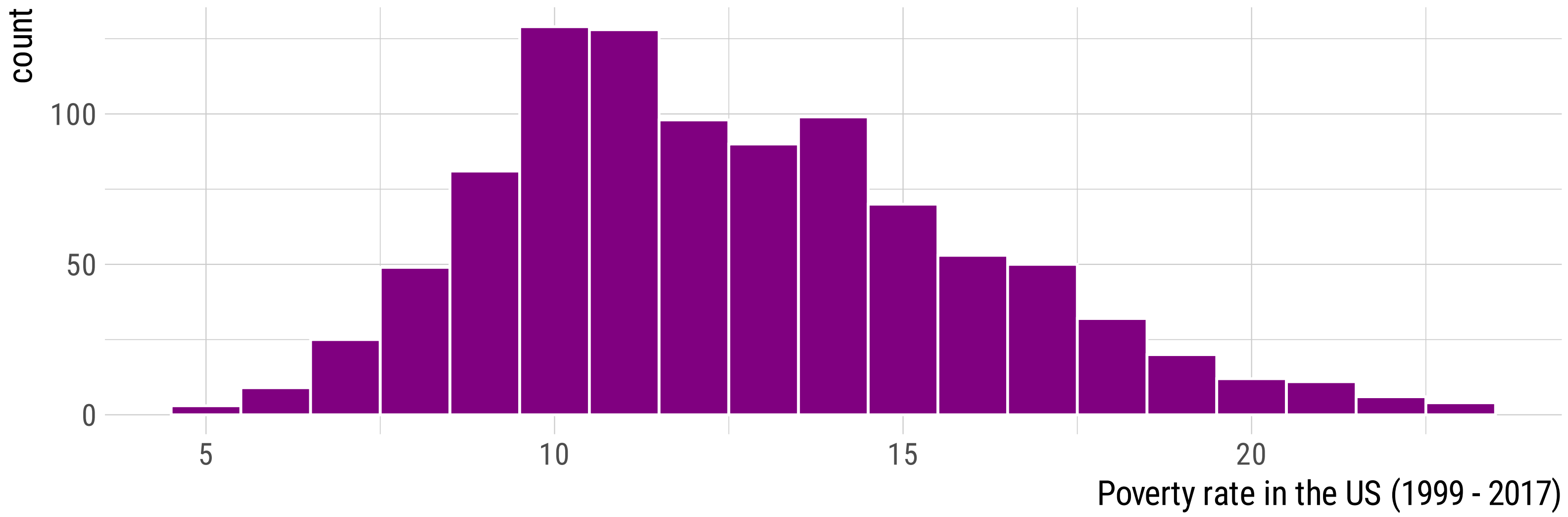
Thus, the probability is best given through a **range** of values.

Visually, this can be represented through a **histogram**.

Continuous random variables

```
data ← read_csv("cdc_data.csv")

data %>%
  ggplot(aes(poverty_rate)) +
  geom_histogram(color = "white", fill = "#800080", binwidth = 1) +
  labs(x = "Poverty rate in the US (1999 - 2017)") +
  easy_x_axis_title_size(13) +
  easy_y_axis_title_size(13)
```



Continuous random variables

Recall that, in a **histogram**, we count the number of occurrences of a range values of a variable in a predetermined **interval**.

These intervals configure the histogram's *bin size* (or bin width).

By dividing the number of counts within each bin by the sample size, we obtain the **relative frequencies** of each range of values.

```
data %>% nrow()
```

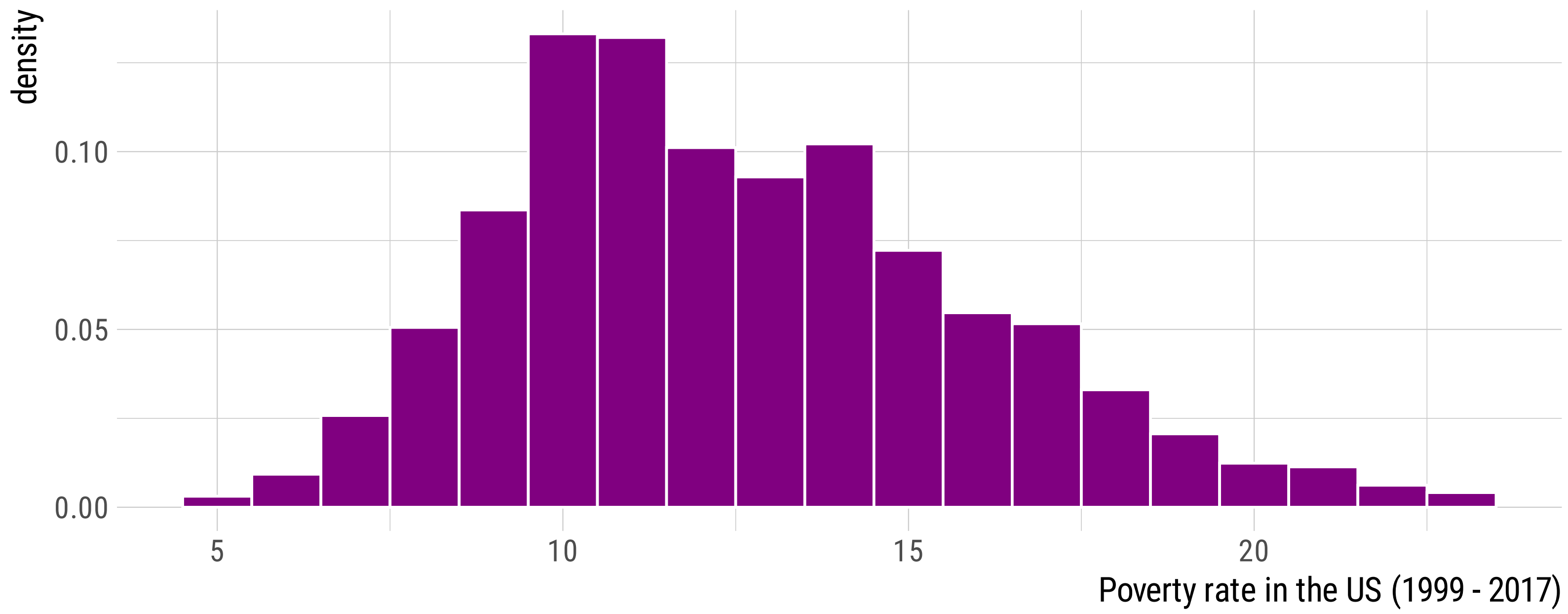
```
#> [1] 969
```

For instance, there are 129 values that fall within the [10 —11) interval, and 90 values within the [13 —14) interval.

- 129/969 = 0.13 is the **probability** that a randomly chosen poverty rate will lie *between* 10% and 10.99%.
- 90/969 = 0.092 is the **probability** that a randomly chosen poverty rate will lie *between* 13% and 13.99%.

Continuous random variables

```
data %>%
  ggplot(aes(poverty_rate)) +
  geom_histogram(aes(y = ..density..), color = "white", fill = "#800080", binwidth = 1) +
  labs(x = "Poverty rate in the US (1999 - 2017)") +
  easy_x_axis_title_size(13) +
  easy_y_axis_title_size(13)
```



Continuous random variables

The **sum** of all relative frequencies must add up to 1.

In case we draw the histogram with a large number of **small** bins, it is possible to **smooth** its edges.

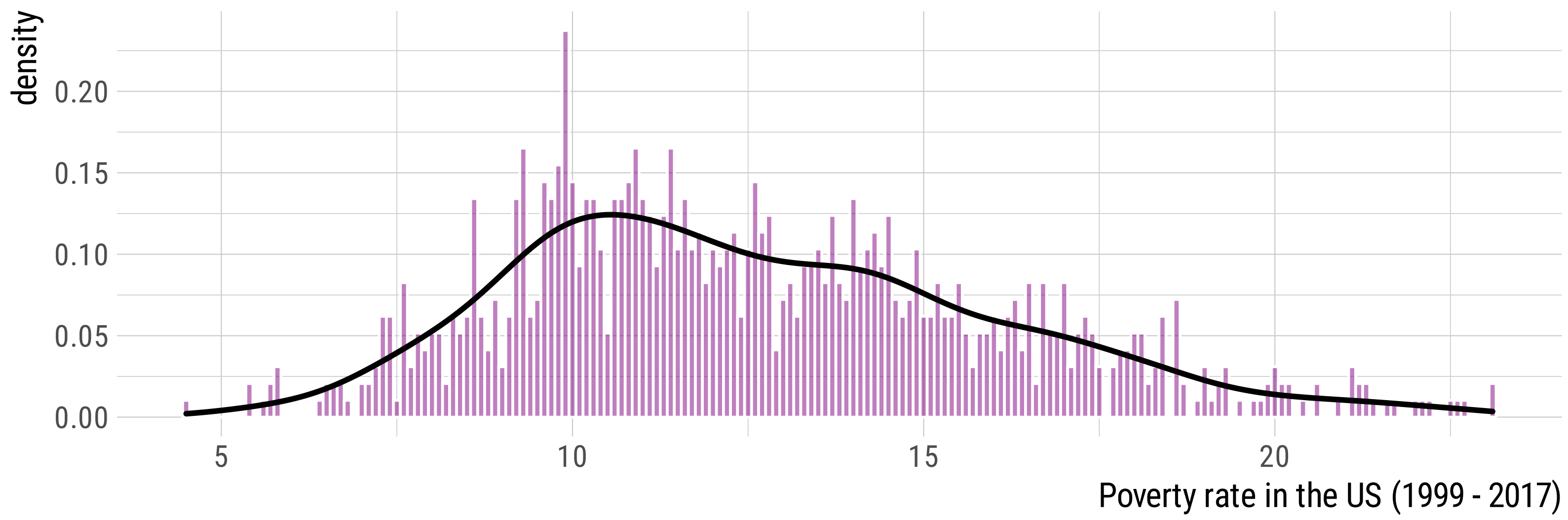
This generates a **density curve**.

It is possible to approximate this density curve through calculus, obtaining a function $f(x)$.

$f(x)$ is called a **probability density function**.

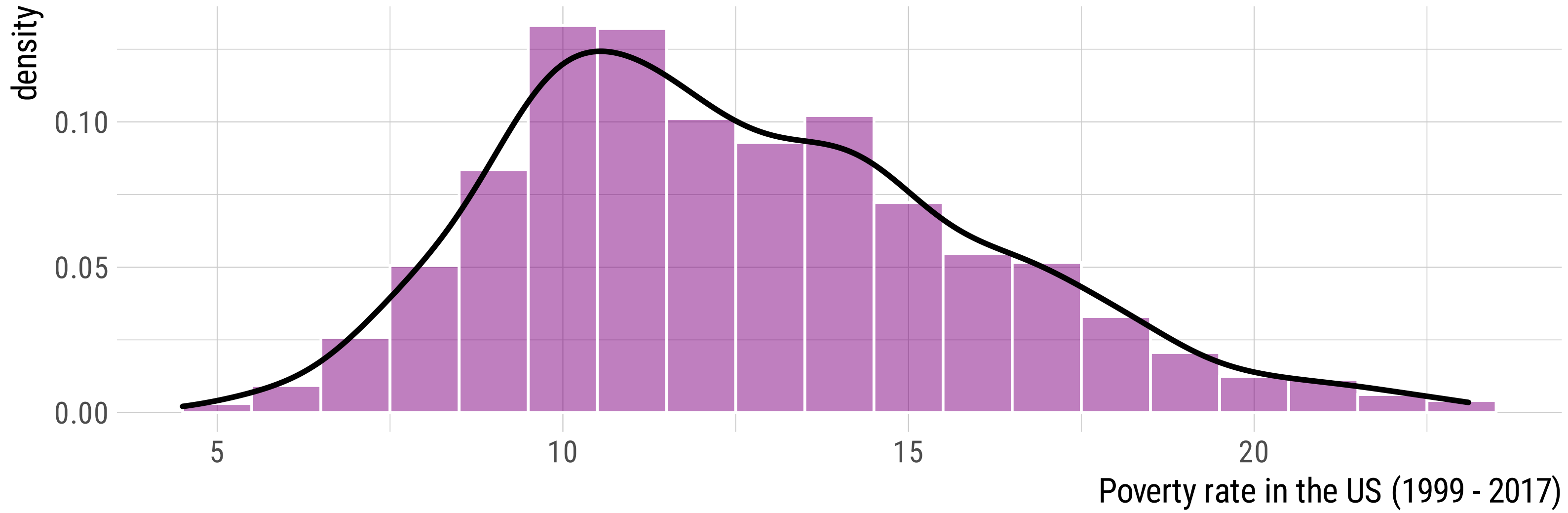
Continuous random variables

```
data %>%
  ggplot(aes(poverty_rate)) +
  geom_histogram(aes(y = ..density..), color = "white", fill = "#800080", binwidth = 0.1, alpha = 0.5) +
  geom_density(alpha = .3, size = 1) +
  labs(x = "Poverty rate in the US (1999 - 2017)") +
  easy_x_axis_title_size(13) +
  easy_y_axis_title_size(13)
```



Continuous random variables

```
data %>%
  ggplot(aes(poverty_rate)) +
  geom_histogram(aes(y = ..density..), color = "white", fill = "#800080", binwidth = 1, alpha = 0.5) +
  geom_density(alpha = .3, size = 1) +
  labs(x = "Poverty rate in the US (1999 - 2017)") +
  easy_x_axis_title_size(13) +
  easy_y_axis_title_size(13)
```



Continuous random variables

A *probability density function* (**PDF**) whose range is $a \le x \le b$ must fulfill the following two **requirements**:

1. $f(x) \ge 0$ for all x between a and b;
2. The total *area* under the curve between a and b is 1.

Just as with discrete RVs, some random variables show such *specific behaviors* that they can be put into certain categories of PDFs.

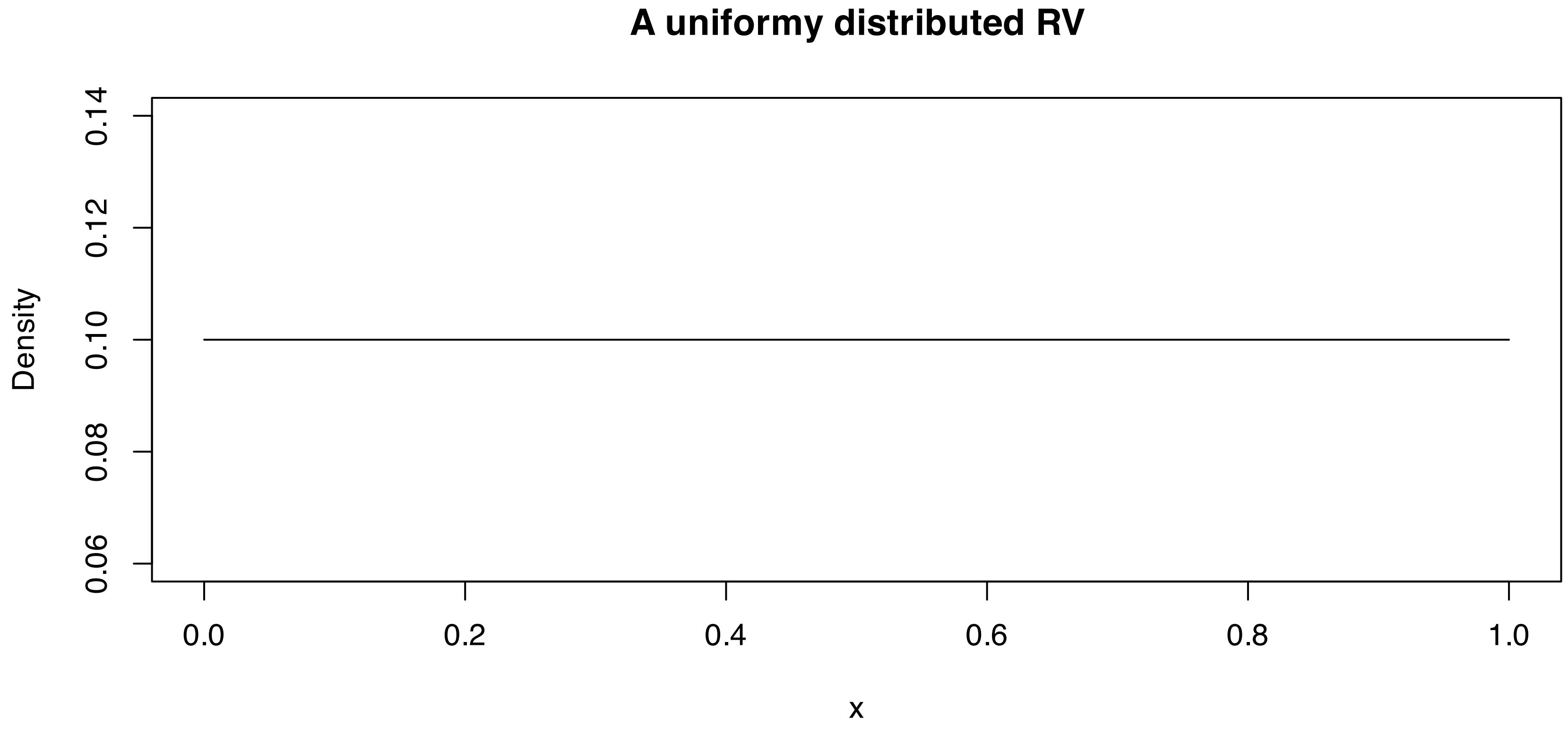
We will study **two** of the most popular *continuous probability distributions*:

- The **Uniform**;
- And the **Normal** (Gaussian) distribution.

The Uniform distribution

The Uniform distribution

The **Uniform** (*aka* rectangular) distribution is useful when a random variable is **uniformly** or **equally likely** to take on *any value* in a given range.



The Uniform distribution

Its **probability density function (PDF)** is given by:

$$f(x) = \frac{1}{b - a}$$

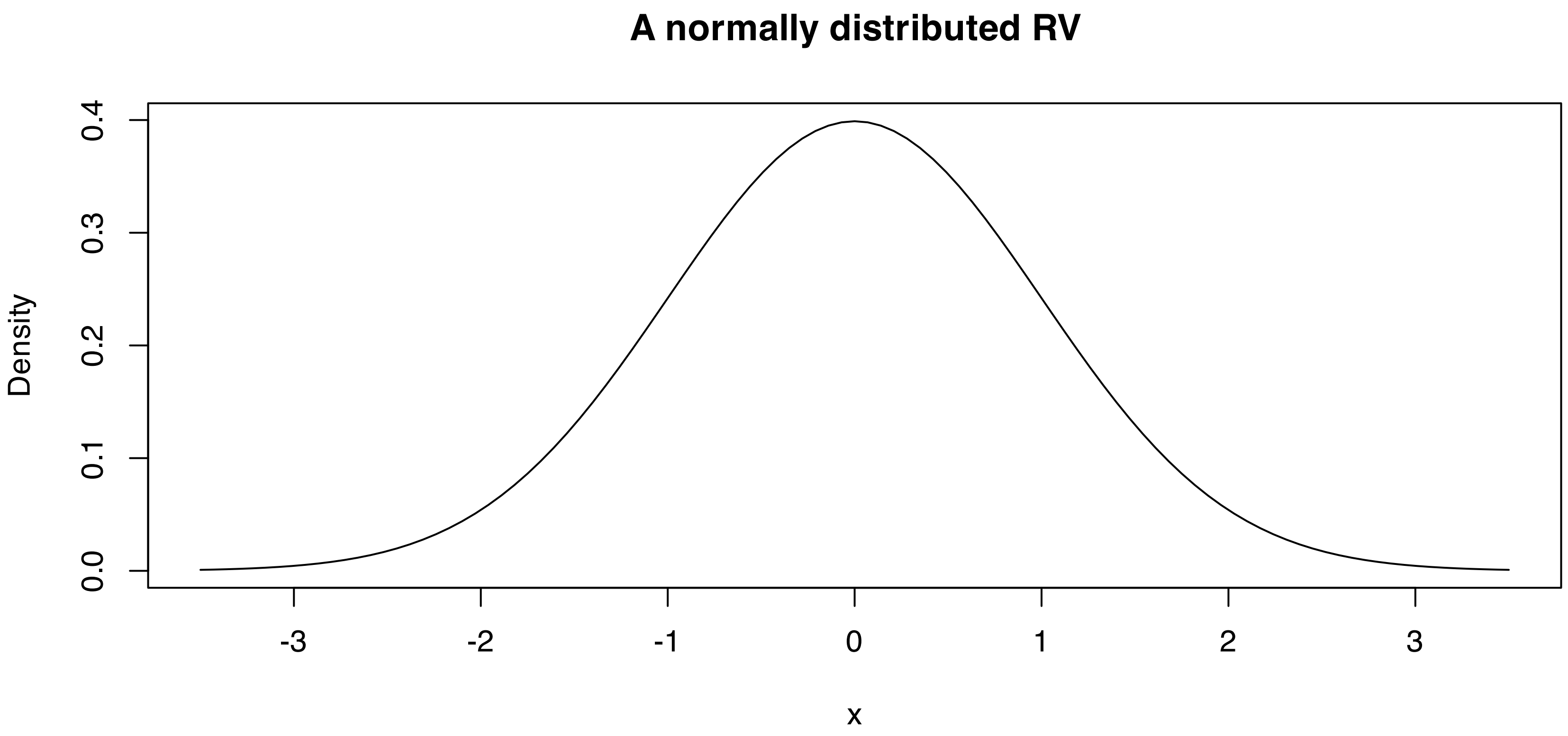
where $a \leq x \leq b$.

The Normal distribution

The Normal distribution

The **Normal** (*aka* Gaussian) distribution is the **most popular** probability distribution in Statistics.

It is called "Normal" due to its patterns being so *commonly observed* in data.



The Normal distribution

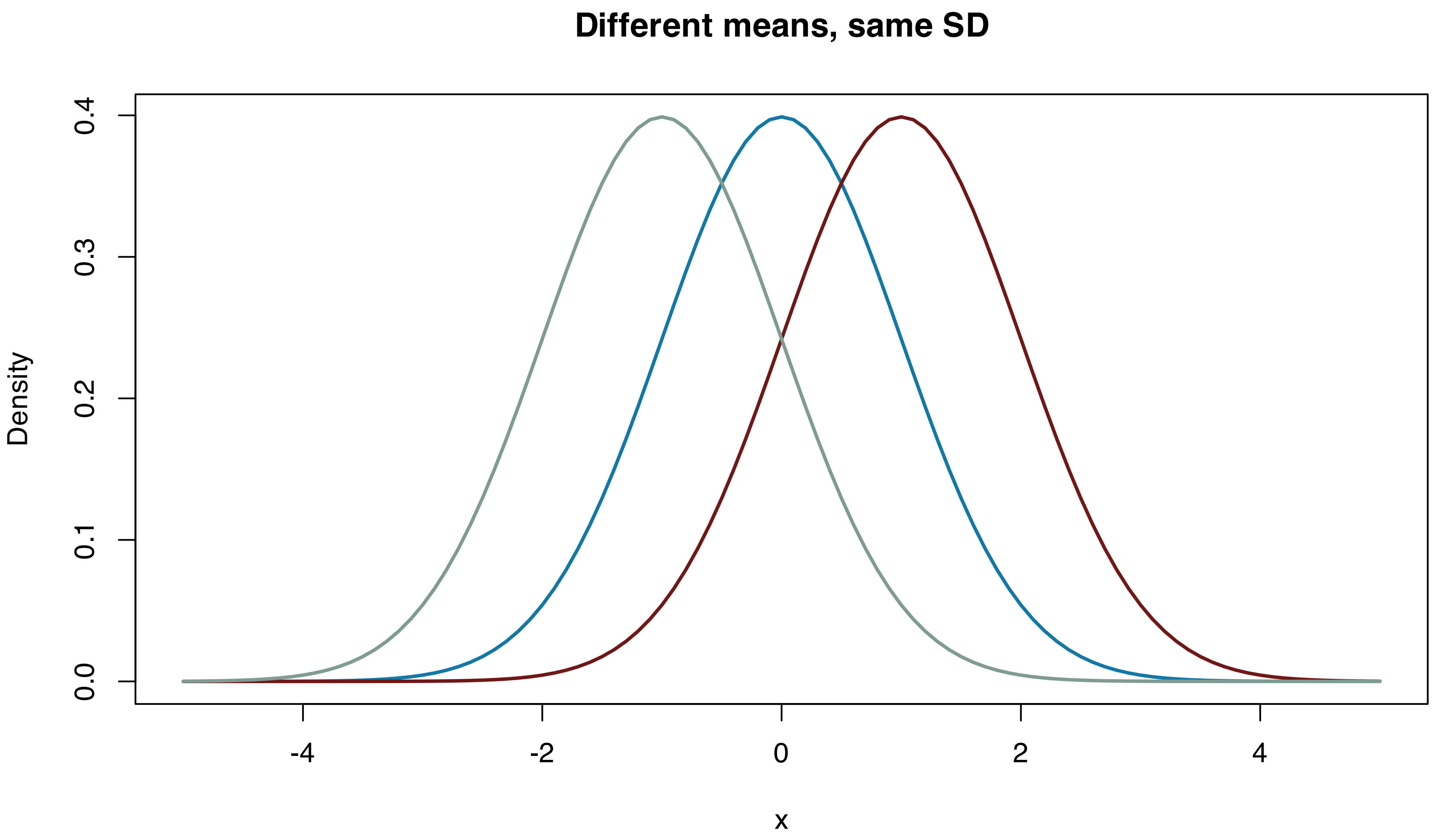
Its **probability density function (PDF)** is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \; ; \quad -\infty < x < \infty$$

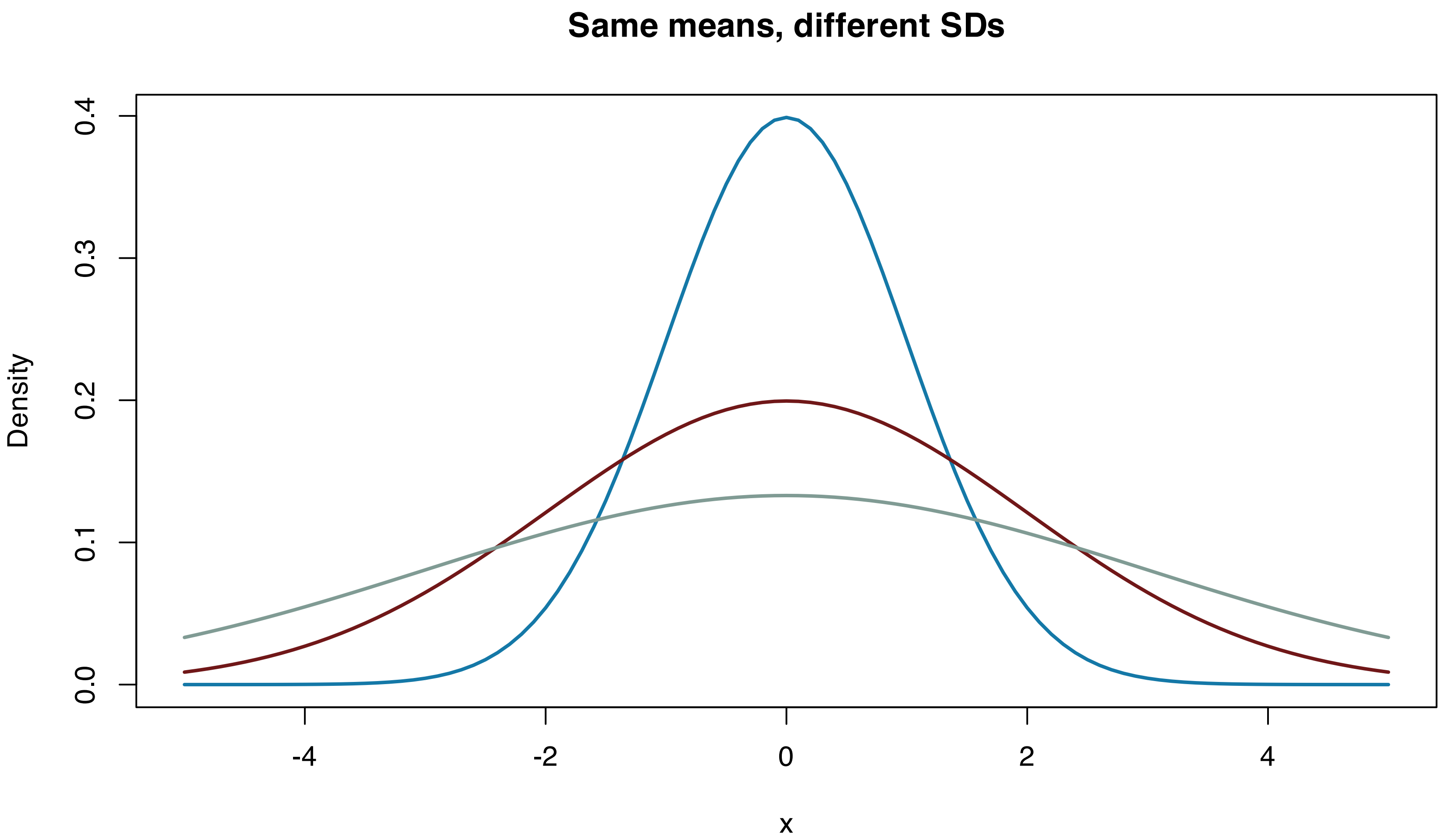
As it is possible to see from the formula, the Normal distribution is described by **two parameters**:

1. The population **mean**, μ;
2. And the population **standard deviation**, σ.

The Normal distribution



The Normal distribution



Next time: Dealing with continuous distributions