Descriptive statistics

Lecture 2

Louis SIRUGUE

CPES 2 - Fall 2022

Quick reminder

1. Import data

```
fb <- read.csv("C:/User/Documents/ligue1.csv", encoding = "UTF-8")
```

2. Class

```
is.numeric("1.6180339") # What would be the output?
```

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

3. Subsetting

```
fb$Home[3]
```

```
## [1] "Troyes"
```

Today we learn how to describe data

1. Distributions
- 1.1. Definition
- 1.2. Graphical representation
- 1.3. Common distributions

2. Central tendency
- 2.1. Mean
- 2.2. Median
- 2.3. Mean vs. median

3. Spread
- 3.1. Range, quantiles, and the IQR
- 3.2. Variance and standard deviation
- 3.3. Standard deviation vs. IQR

4. Inference
- 4.1. Data generating process
- 4.2. Empirical vs. theoretical moments
- 4.3. Confidence interval

5. Wrap up!

Today we learn how to describe data

1. Distributions

1.1. Definition

- The point of descriptive statistics is to **summarize a big table** of values with a small set of **tractable statistics**
- The most comprehensive way to characterize a variable/vector is to compute its **distribution**:
  - **What** are the **values** the variable takes?
  - **How frequently** does each of these values appear?

→ Consider for instance the following variable:

Variable 1
3 5 4 6 5 4 5 7 7 6 1 7 6 7 6 4 7 7 6 6 5 6 6 3 4 5 2 6 8 8

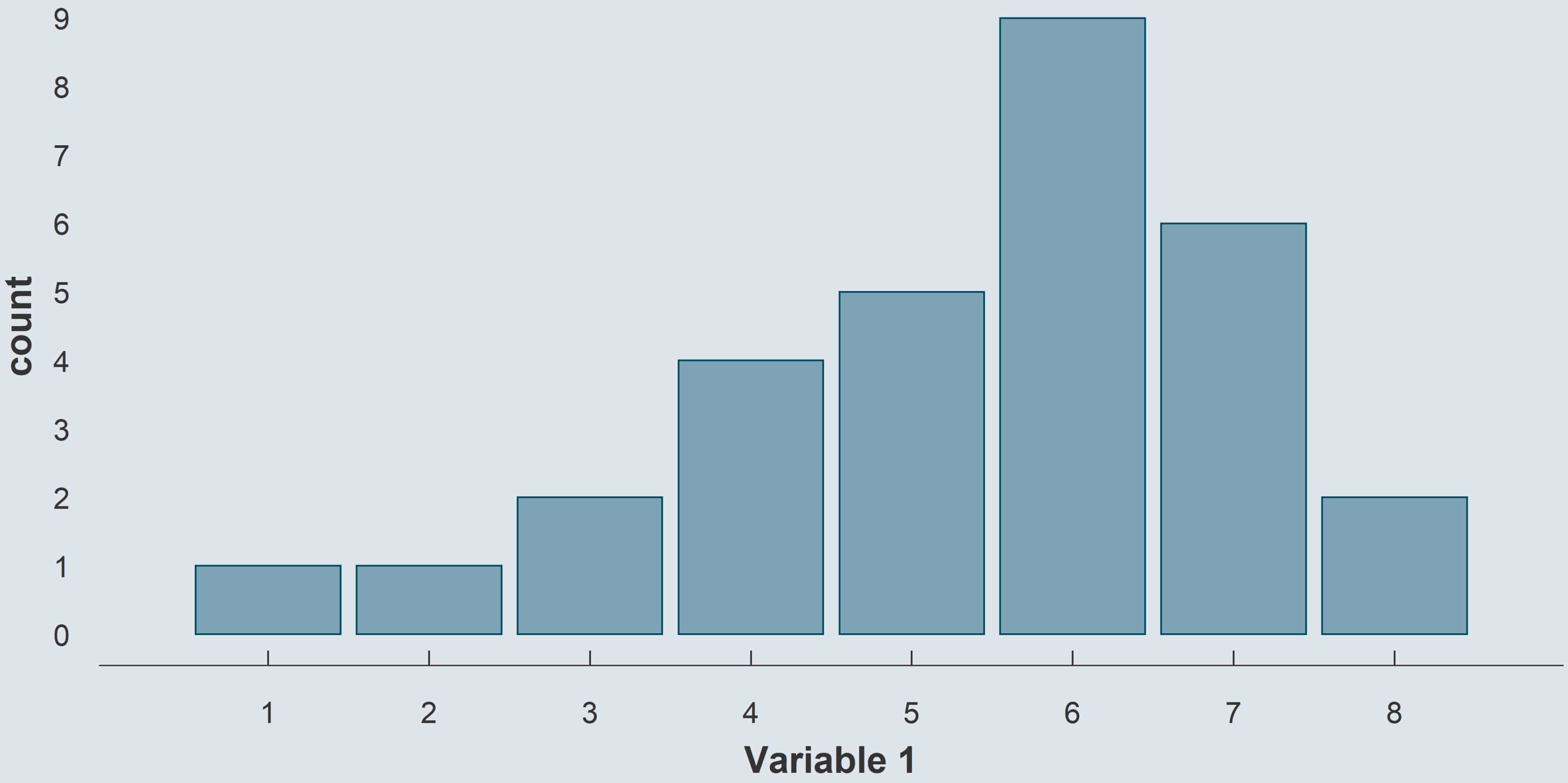
- We can count how many times each value appears

Variable 1	1	2	3	4	5	6	7	8
n	1	1	2	4	5	9	6	2

- And we can represent this distribution graphically with a bar plot
  - Each possible value on the x-axis
  - Their number of occurrences on the y-axis

1. Distributions

1.2. Graphical representation



1. Distributions

1.2. Graphical representation

- But what if we would like to do the same thing for the following variable?

Variable 2

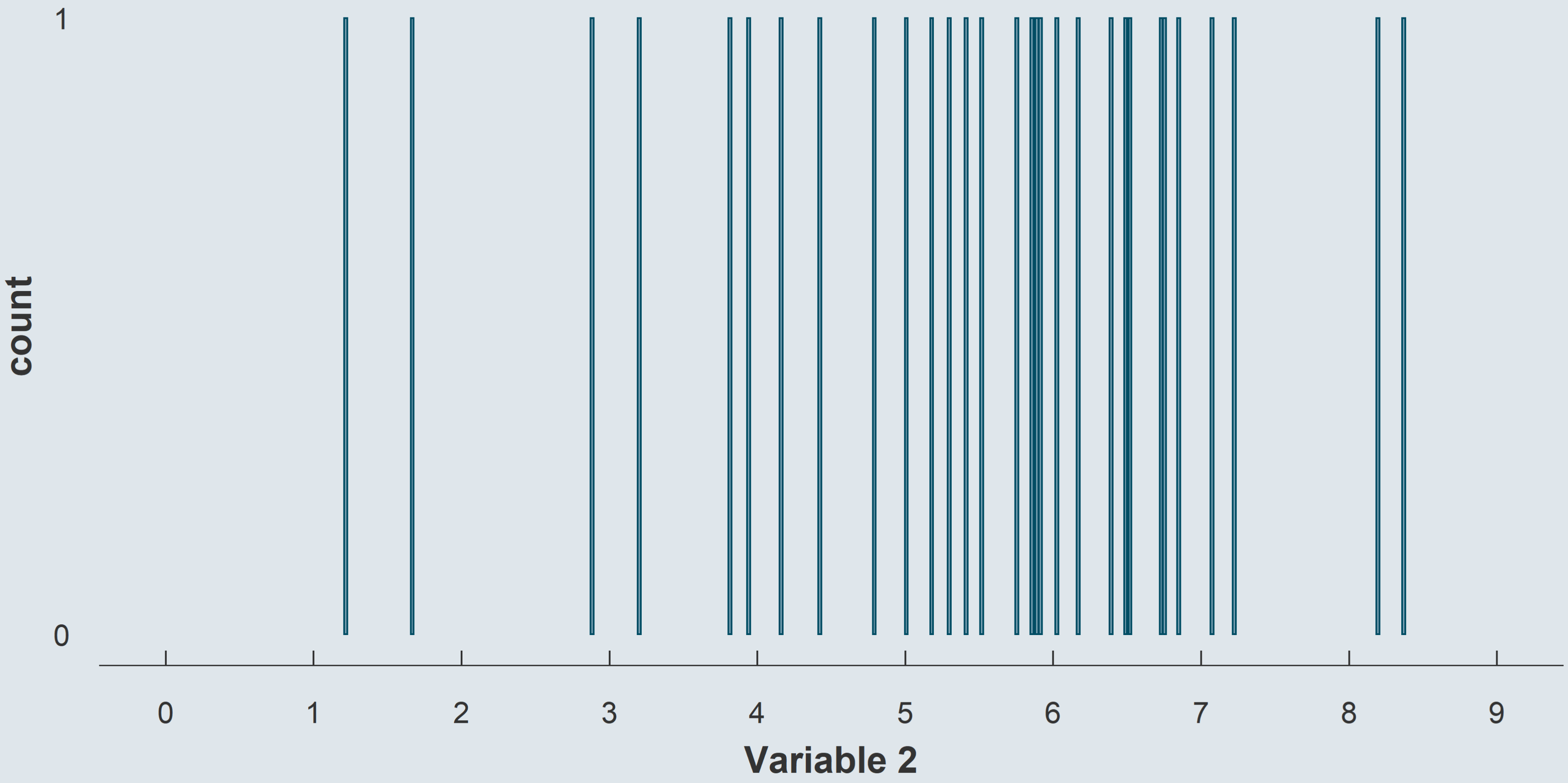
5.912877	5.006781	5.517149	5.854849	5.177872	3.815240
1.666582	4.422721	6.025062	5.411020	5.889811	6.729103
4.160800	6.519049	6.849172	8.368158	6.167404	2.882974
6.751888	3.202183	6.390224	3.942039	6.488909	8.195647
7.073922	4.790039	5.297919	1.218109	5.754213	7.225030

- Each value appears only once
  - So the count of each variable does not help summarizing the variable

➜ ***Let's have a look at the corresponding bar plot***

1. Distributions

1.2. Graphical representation



1. Distributions

1.2. Graphical representation

- It does not look good for this variable because it is continuous, while the first one was discrete
  - **Discrete variables:** variables that can take a finite (or, in practice, a sufficiently small) number of values, e.g., number of siblings, eye color, ...
  - **Continuous variables:** variables that can take an infinite (or, in practice, a sufficiently large) number of values, e.g., annual income, height in centimeters, ...

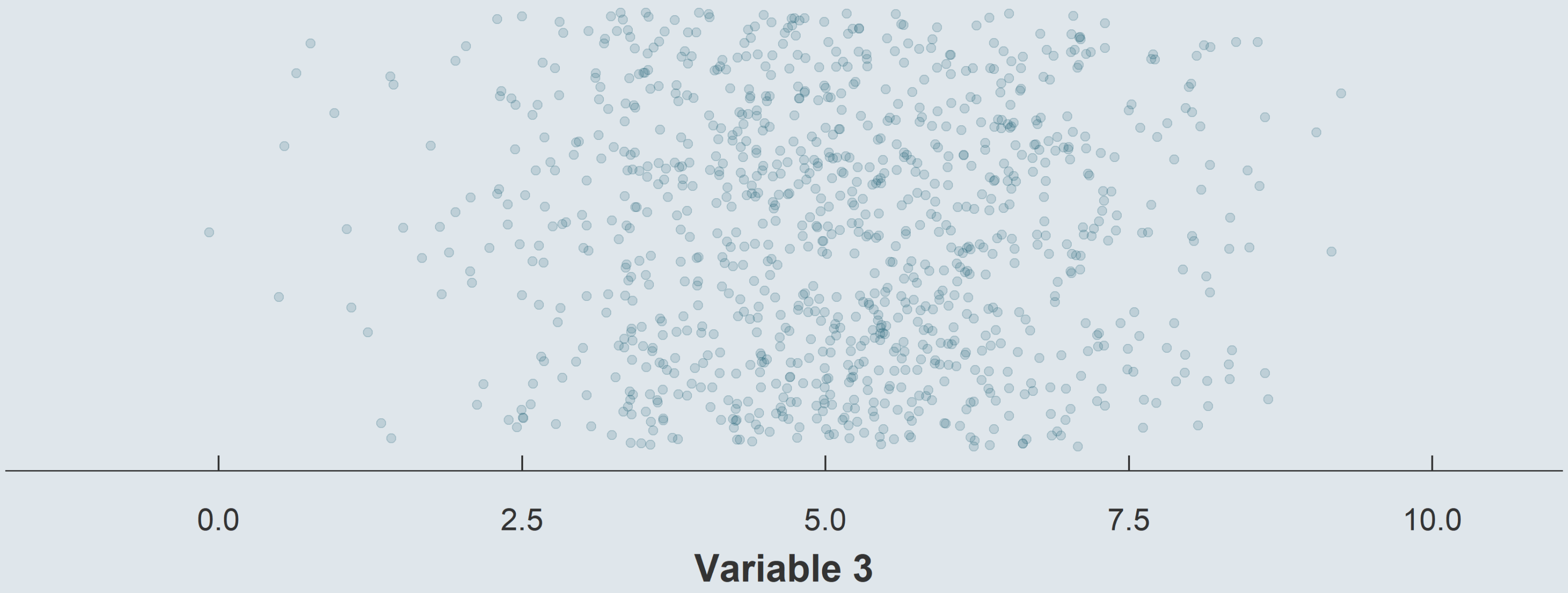
→ *In practice some variables can be difficult to classify. For instance, **age (in years)** can be viewed **as a discrete** variable because it can take a finite set of values, but this set being possibly quite wide, one could also view it **as a continuous variable**. It often depends on the context.*

- One solution to get a sense of the **distribution** of a **continuous variable** is to do a **histogram**
  - Instead of taking each value separately, group them into *bins* and show how many values fall into each bin
  - The bar plots we've seen so far are basically histograms with the number of bins being equal to the number of possible values

1. Distributions

1.2. Graphical representation

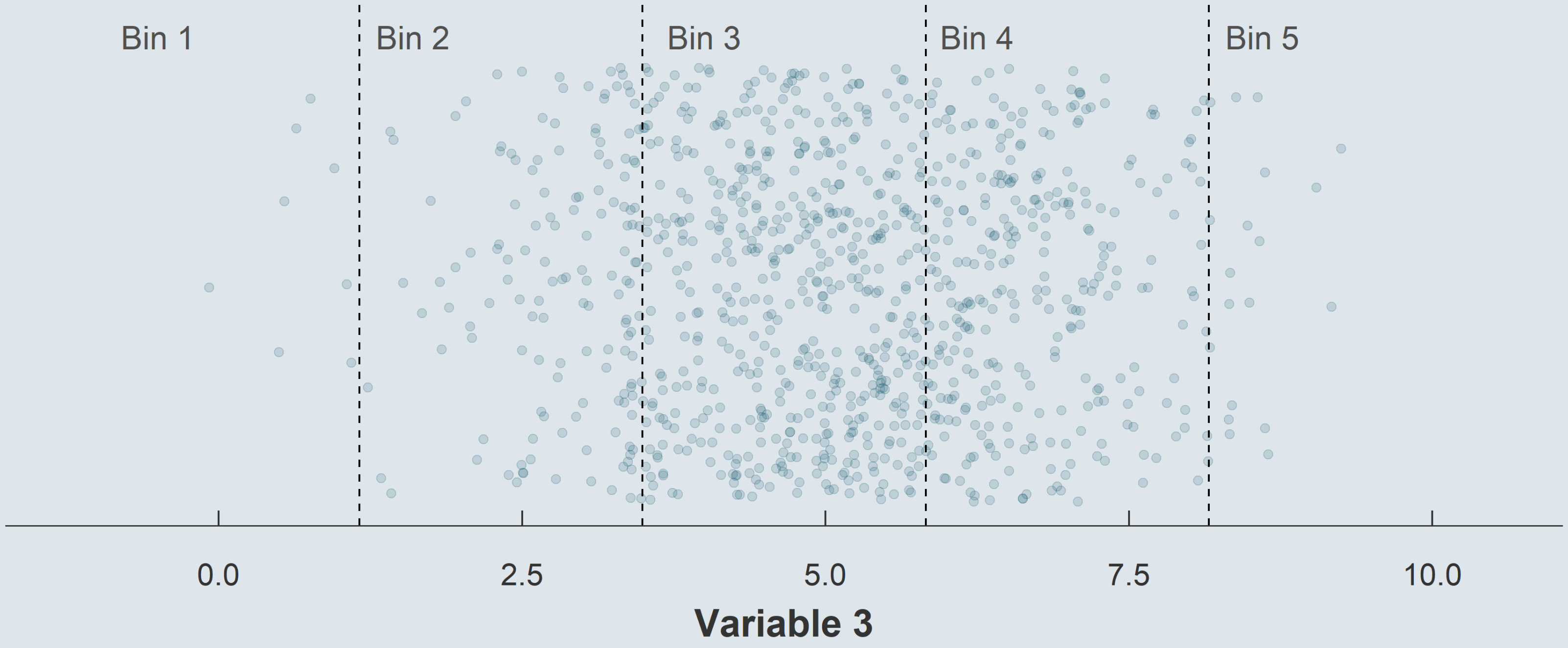
- Consider for instance the following variable. For clarity each point is shifted vertically by a random amount



1. Distributions

1.2. Graphical representation

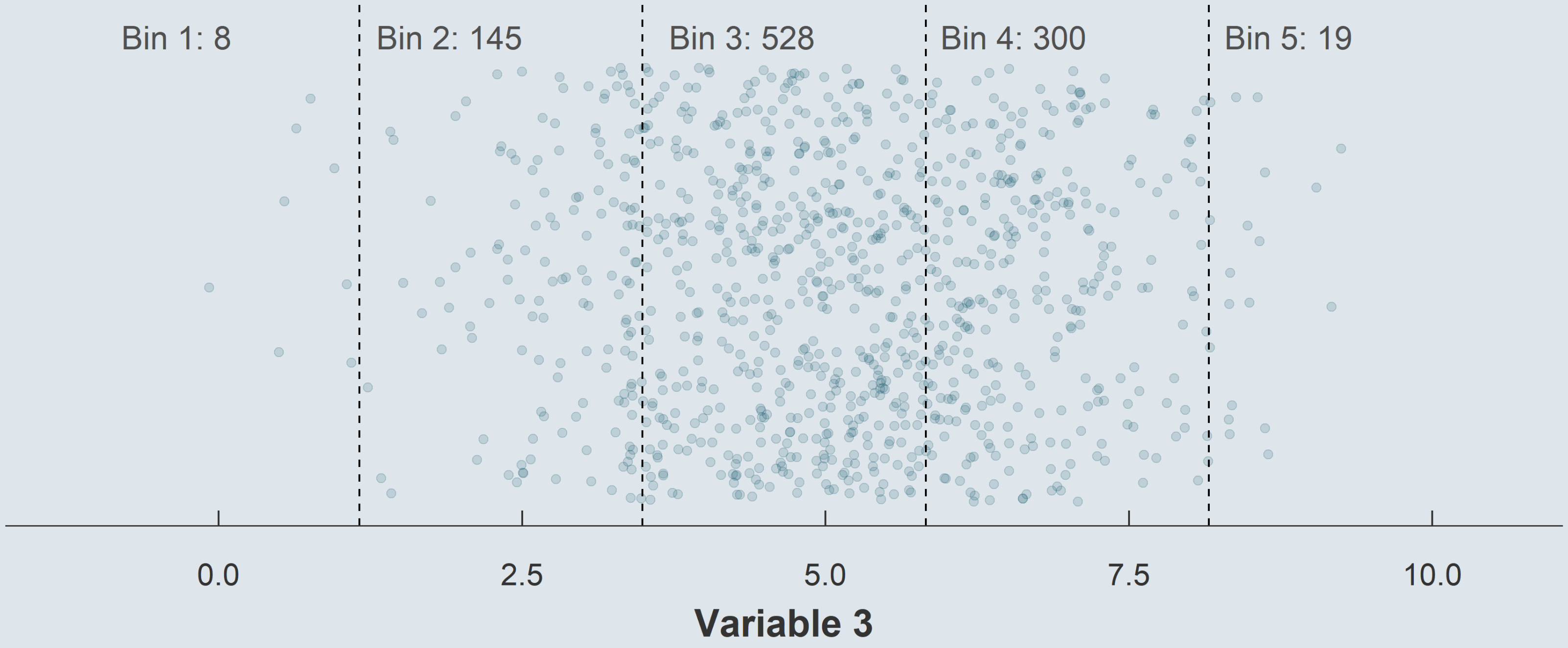
- Consider for instance the following variable. For clarity each point is shifted vertically by a random amount
  - We can divide the domain of this variable into 5 bins



1. Distributions

1.2. Graphical representation

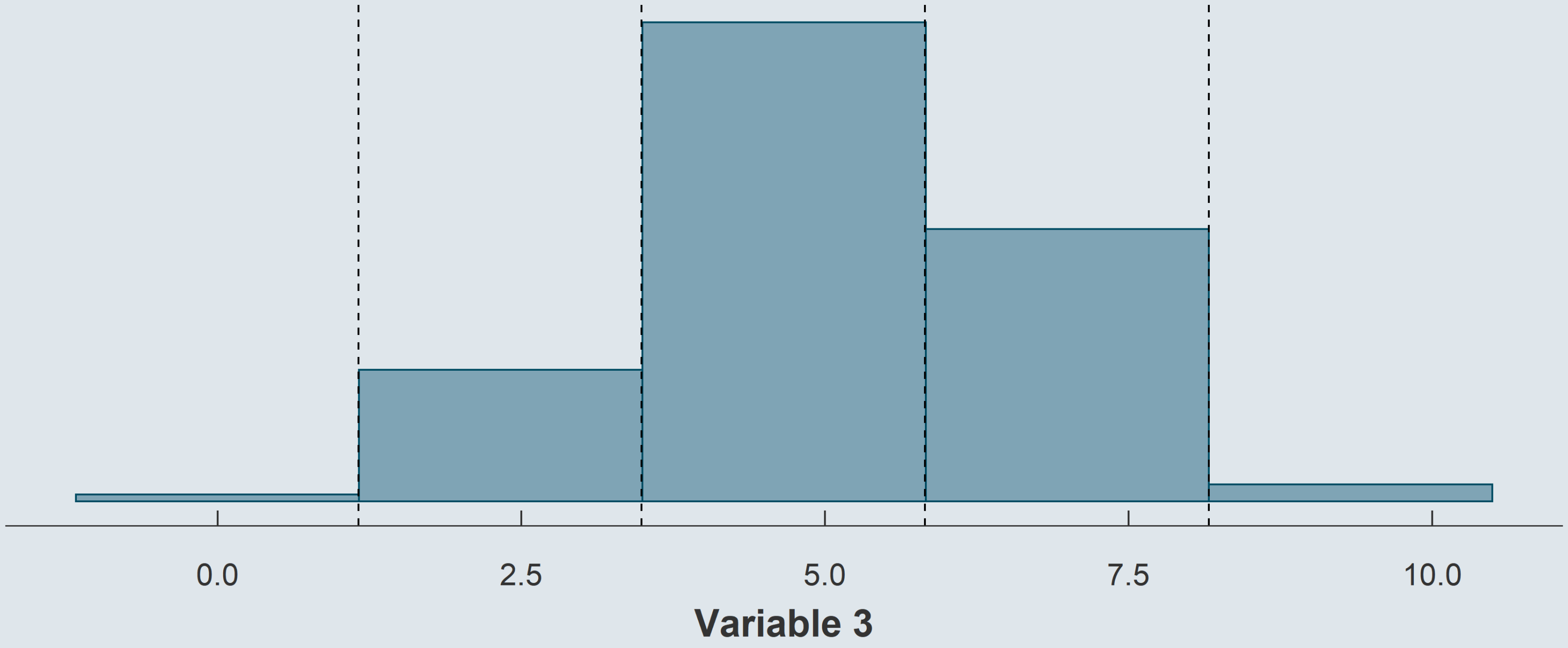
- Consider for instance the following variable. For clarity each point is shifted vertically by a random amount
  - We can divide the domain of this variable into 5 bins
  - And count the number of observations within each bin



1. Distributions

1.2. Graphical representation

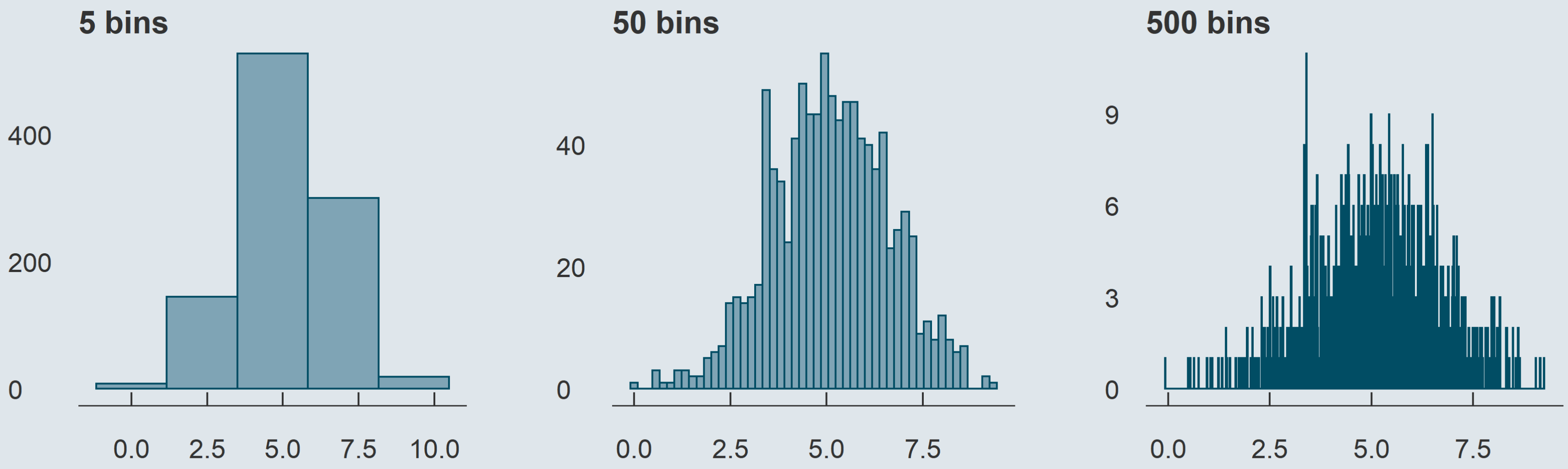
- Consider for instance the following variable. For clarity each point is shifted vertically by a random amount
  - We can divide the domain of this variable into 5 bins
  - And count the number of observations within each bin



1. Distributions

1.2. Graphical representation

- There's no definitive rule to choose the number of bins
  - But too many or too few can yield misleading histograms



➜ Note that choosing the number of bins is equivalent to choosing the width of each bin

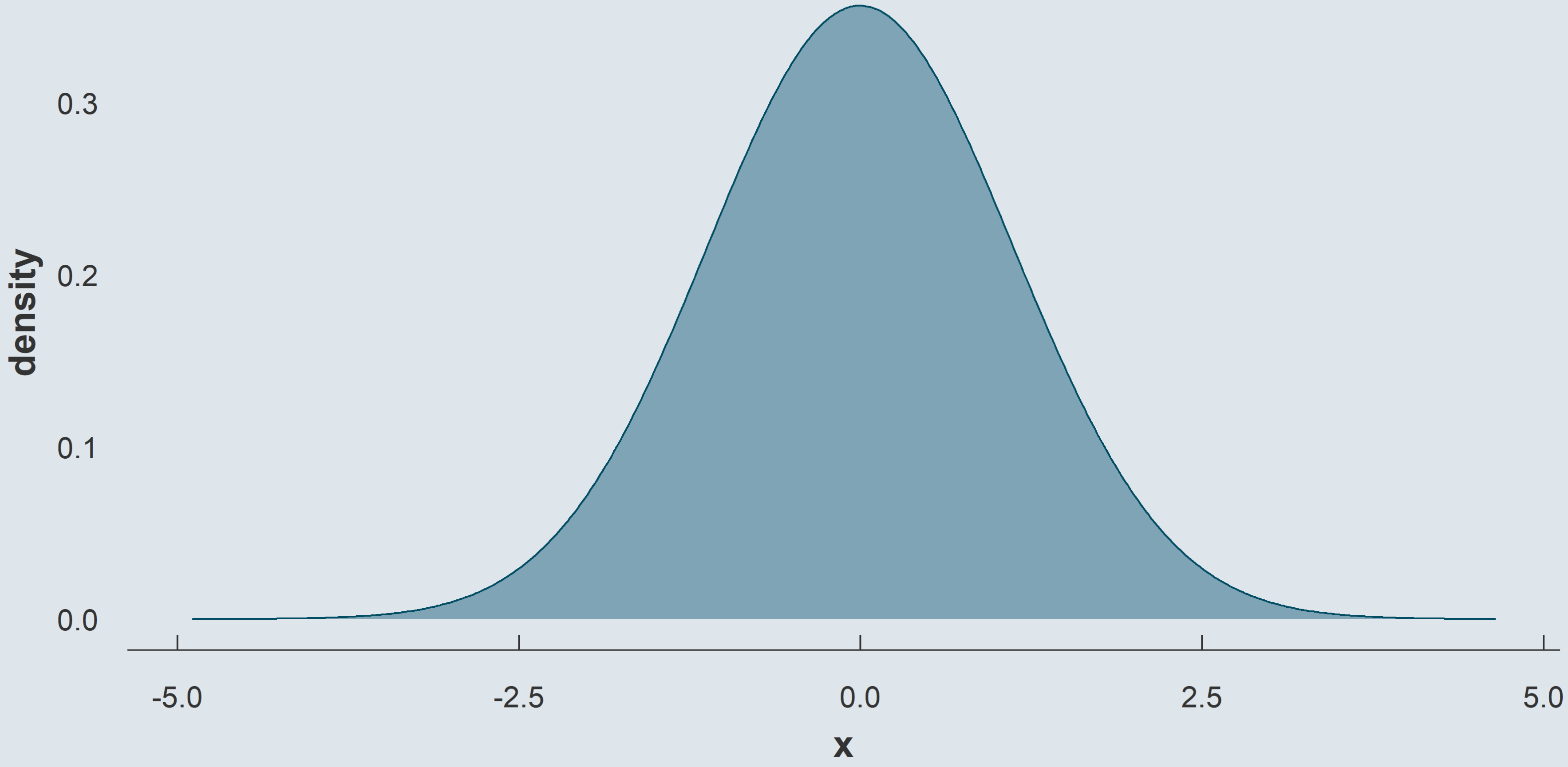
1. Distributions

1.2. Graphical representation

- **Densities** are often used instead of **histograms**
  - Both are based on the **same principle**, but densities are **continuous**
- We won't learn how to derive it in this course but the idea is the same
  - The **higher the value** on the y-axis, the **more observations** there are around the corresponding x location
- The **smoothness** of the density can be tuned with the **bandwidth**
  - The larger the smoother

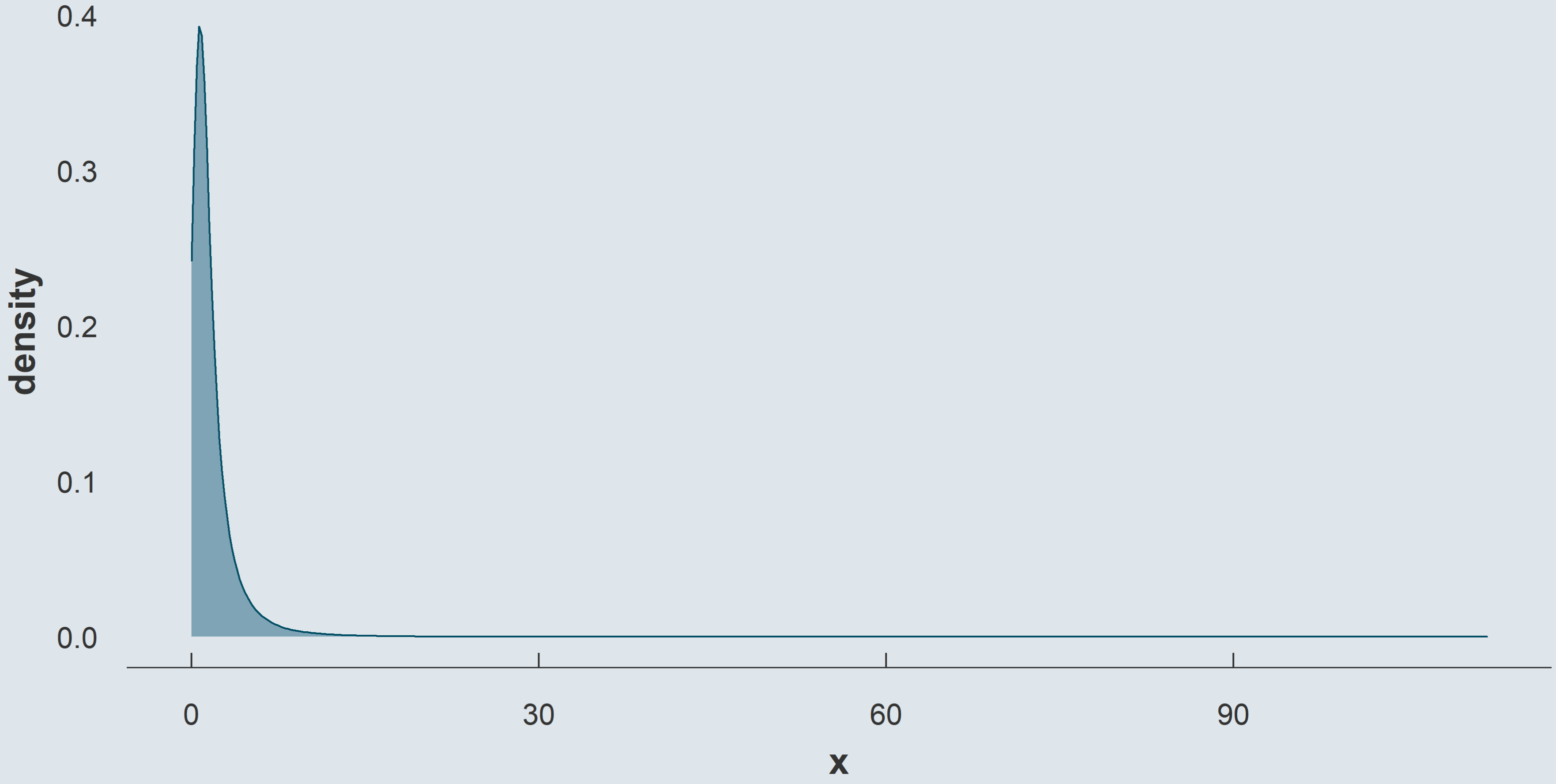
1. Distributions

1.3. Common distributions: Normal distribution



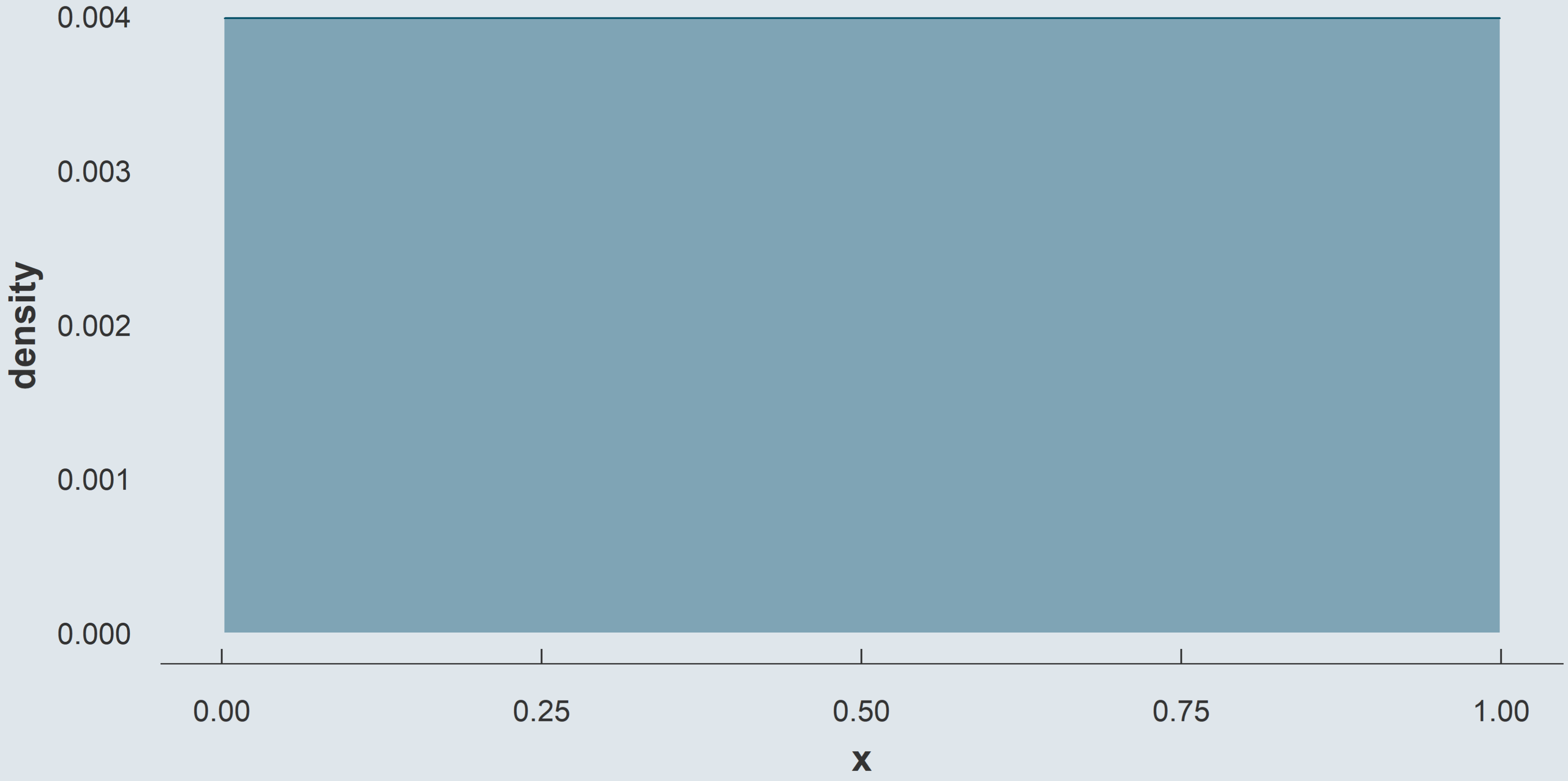
1. Distributions

1.3. Common distributions: Log-normal distribution



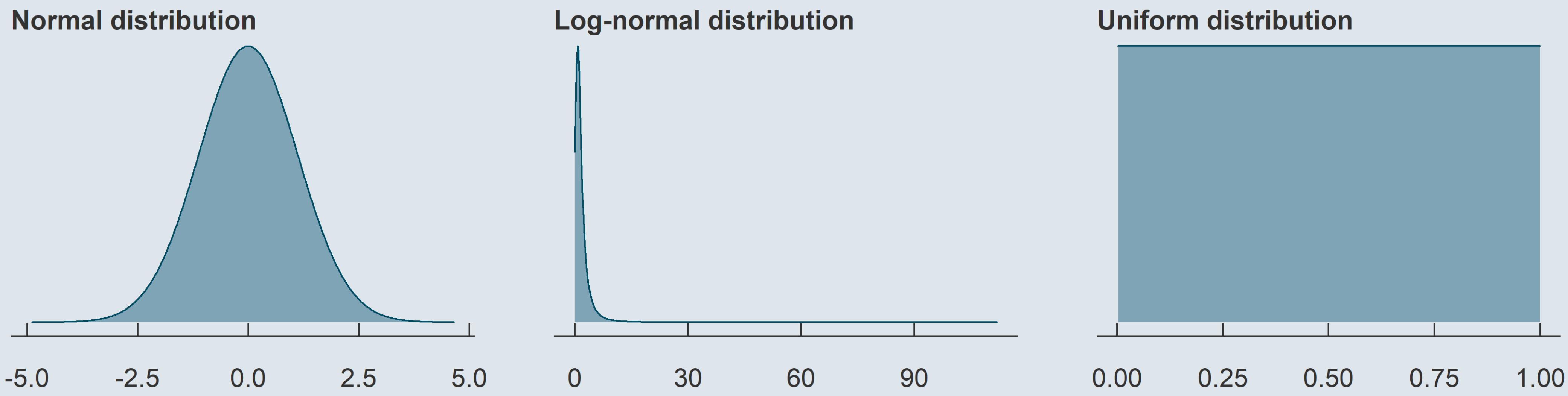
1. Distributions

1.3. Common distributions: Uniform distribution



1. Distributions

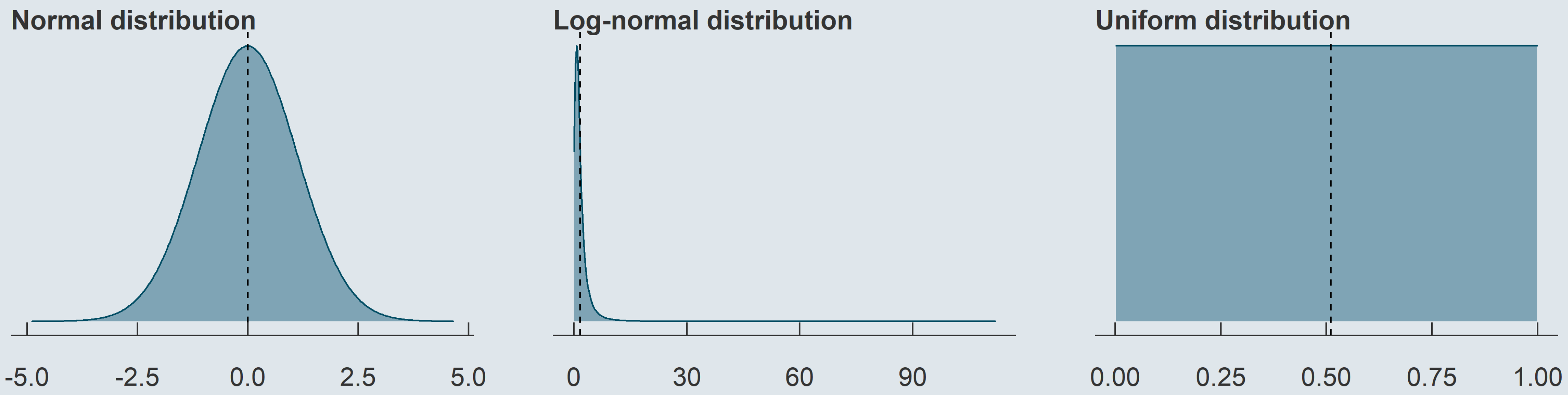
1.3. Common distributions: Summarizing distributions



- How to **summarize** these distributions with simple statistics?

1. Distributions

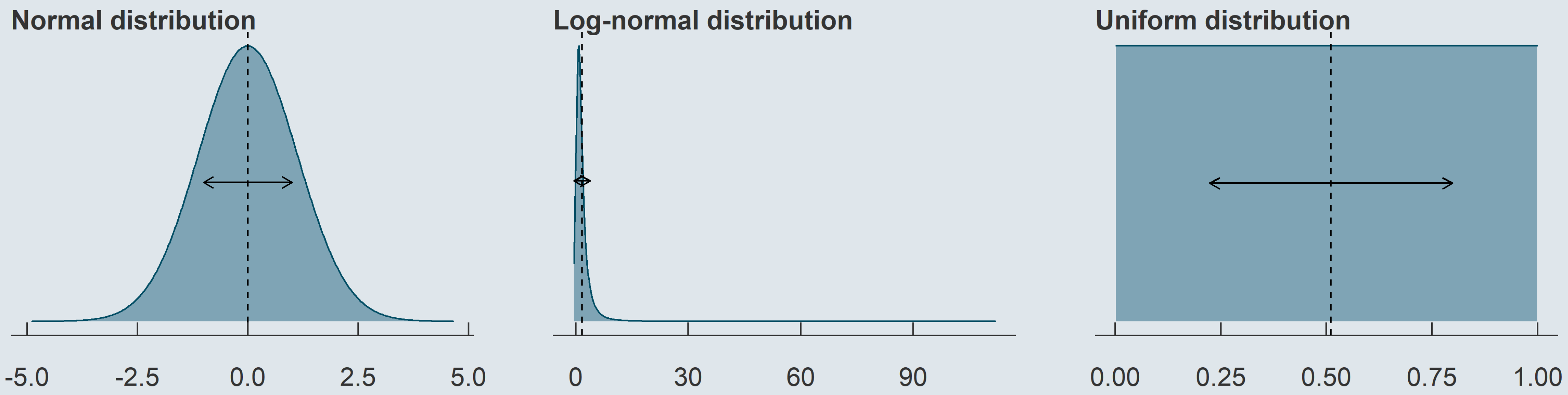
1.3. Common distributions: Summarizing distributions



- How to **summarize** these distributions with simple statistics?
  - By describing their **central tendency** (e.g., mean, median)

1. Distributions

1.3. Common distributions: Summarizing distributions



- How to **summarize** these distributions with simple statistics?
  - By describing their **central tendency** (e.g., mean, median)
  - And their **spread** (e.g., standard deviation, inter-quartile range)

Overview

Overview

2. Central tendency

2.1. Mean

- The mean is the most common statistic to describe central tendencies
  - Take for instance the grades I gave to the final projects in spring 2021:

Grades I gave in spring 2021

20	17.5	16	16.0	14.5	19.5	18.5
20	17.5	16	14.5	19.5	18.5	18.5

- The mean is simply the sum of all the grades divided by the number of grades:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{20+20+17.5+17.5+16+16+16+14.5+14.5+19.5+19.5+18.5+18.5+18.5}{14} = 17.61$$

2. Central tendency

2.1. Mean

- The mean is the most common statistic to describe central tendencies
  - Take for instance the grades I gave to the final projects in spring 2021:

Grades I gave in spring 2021						
20	17.5	16	16.0	14.5	19.5	18.5
20	17.5	16	14.5	19.5	18.5	18.5

- Note that it can also be expressed as the sum of each value weighted by its proportion in the distribution

$$\bar{x} = \frac{2}{14} \times 20 + \frac{2}{14} \times 17.5 + \frac{3}{14} \times 16 + \frac{2}{14} \times 14.5 + \frac{2}{14} \times 19.5 + \frac{3}{14} \times 18.5 = 17.61$$

2. Central tendency

2.2. Median

- To obtain the median you first need to **sort the values**:

Grades I gave in spring 2021

1	2	3	4	5	6	7	8	9	10	11	12	13	14
14.5	14.5	16	16	16	17.5	17.5	18.5	18.5	18.5	19.5	19.5	20	20

- The median is the value that **divides** the distribution into **two halves**
- When there is an even number of observations, the median is the average of the last value of the first half and the first value of the second half

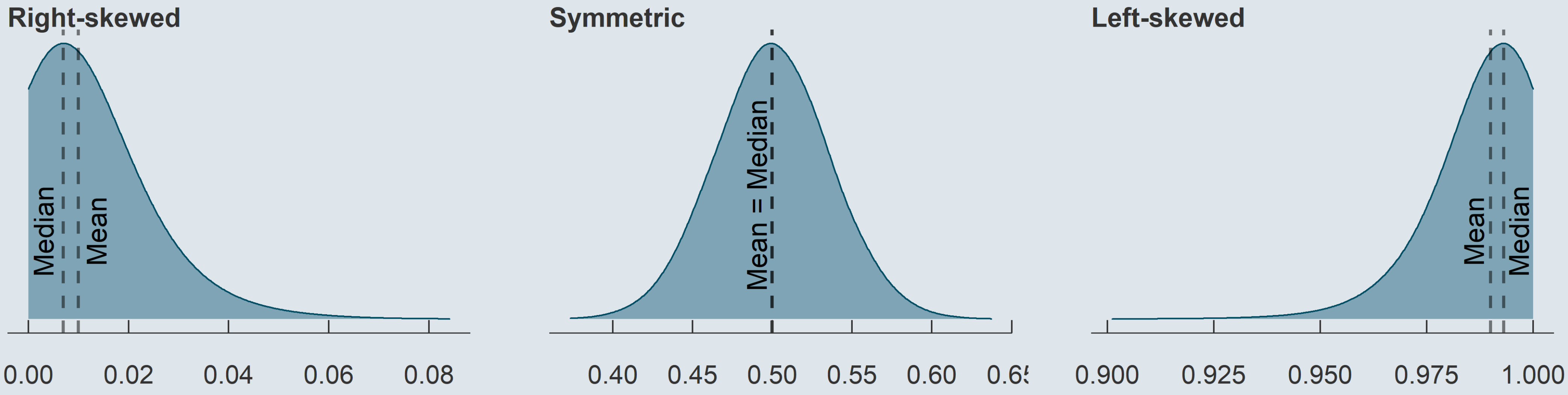
As we have 14 observations, the median is the average of the 7$^{\text{th}}$ and the 8$^{\text{th}}$ observations:

$$\text{Med}(x) = \begin{cases} x[\frac{N+1}{2}] & \text{if } N \text{ is odd} \\ \frac{x[\frac{N}{2}]+x[\frac{N}{2}+1]}{2} & \text{if } N \text{ is even} \end{cases} = \frac{17.5+18.5}{2} = 18$$

2. Central tendency

2.3. Mean vs. median: relative magnitude

- The **relative magnitude** of the mean and the median depends on the **symmetry of the distribution**:
  - The **mean is larger** than the median if the distribution is **right-skewed**
  - The mean and the median are **equal** if the distribution is **symmetric**
  - The **mean is lower** than the median if the distribution is **left-skewed**



2. Central tendency

2.3. Mean vs. median: robustness

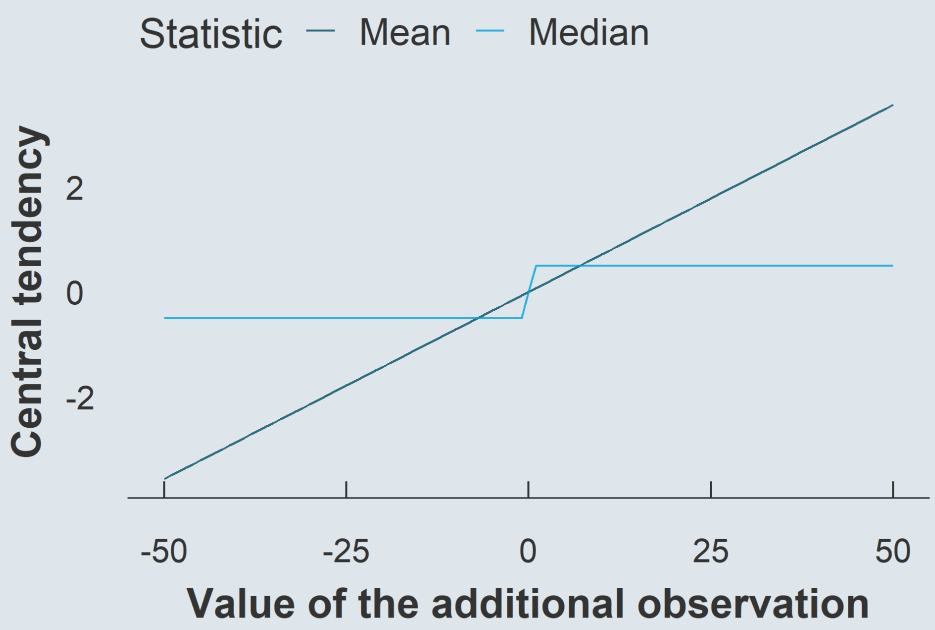
- The **median** is indeed **less sensitive** than the mean to thick tails and outliers
- For this reason we say that the median is a ***robust statistic***

Let's illustrate that with a small example!

- Consider the following variable:

-3	-2	-2	-1	-1	-1	0	1	1	1	2	2	3

- How would the mean and the median **react** if we were to **add one single observation**?
  - We can plot the value of the additional observation on the x axis and the value of the mean and the median on the y axis



2. Central tendency

2.3. Mean vs. median: in R

- Both statistics have **dedicated R functions**

```
variable <- c(1, 2, 4, 8, 12)
c(mean(variable), median(variable))
```

```
## [1] 5.4 4.0
```

- As always, you should **pay attention to NAs** when using these functions

```
mean(c(1, 2, 3, 4, NA))
```

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

```
mean(c(1, 2, 3, 4, NA), na.rm = T)
```

```
## [1] 2.5
```

2. Central tendency

2.3. Mean vs. median: with binary variable

- A **binary variable** is a variable that can take only **two values** *(e.g., male/female, accepted/rejected)*
    - Any binary variable can be expressed as a sequence of **0s and 1s**

- Consider the following binary variable of length 4

0	1	1	1

- The **mean** of a binary variable is equal the the **percentage of 1s**:

$$\frac{0+1+1+1}{4} = \frac{3}{4} = 75\%$$

- The **median** of a binary variable is equal to the **mode** *(mode = most frequent value of a variable)*

$$\frac{1+1}{2} = 1$$

Overview

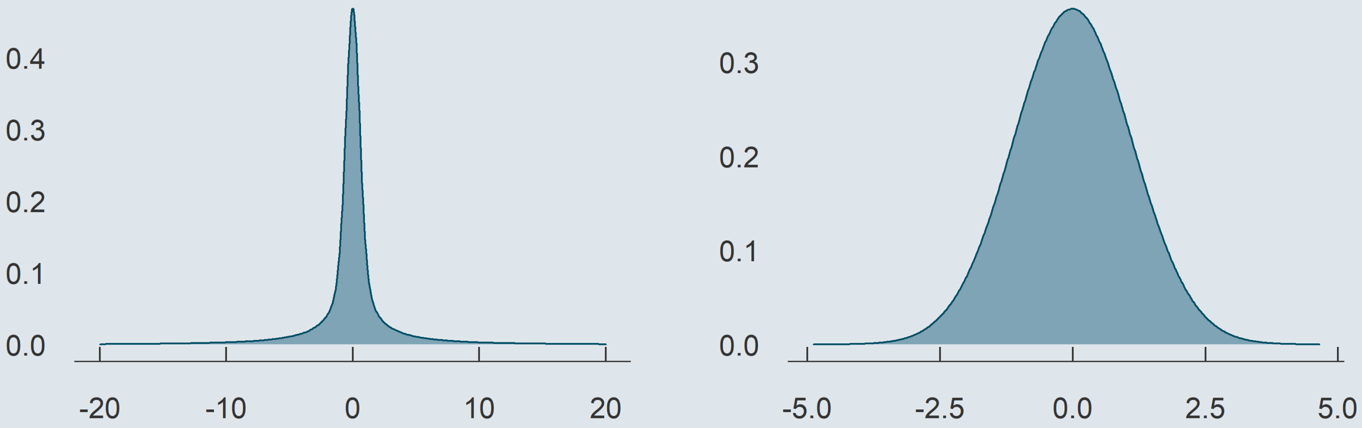
Overview

3. Spread

3.1. Range, quantiles, and the IQR

- The **most intuitive** statistic to describe the spread of a variable is probably
  - **The range: the minimum and maximum value it can take**

- But consider the following two distributions:



- In the presence of outliers or very skewed distributions, the **full range** of a variable **may not be representative** of what we mean by *'spread'*

- That's why we tend to prefer **inter-quantile** ranges

3. Spread

3.1. Range, quantiles, and the IQR

- **Quantiles** are observations that **divide** the population into **groups of equal size**
  - The **median** divides the population into **2 groups** of equal size
  - **Quartiles** divide the population into **4 groups** of equal size
  - There are also **terciles**, **quintiles**, **deciles**, and so on

- One way to **compute quartiles**: divide the ordered variable according to the median
  - The lower quartile value is the median of the lower half of the data
  - The upper quartile value is the median of the upper half of the data
  - *If there is an odd number of data points in the original ordered data set, don't include the median in either half*

-3	-2	-1	0	1	2	3

$$Q_1 = -2,\ \ Q_2 = 0,\ \ Q_3 = 2$$

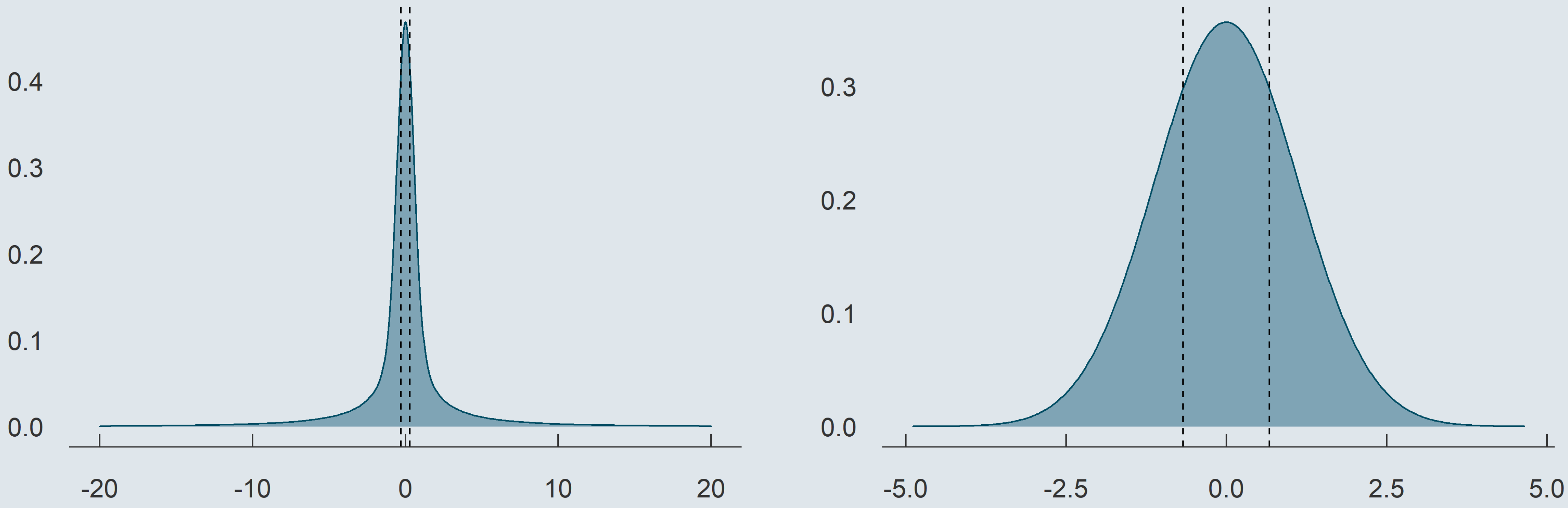
-3	-2	-1	0	0	1	2	3

$$Q_1 = -1.5,\ \ Q_2 = 0,\ \ Q_3 = 1.5$$

3. Spread

3.1. Range, quantiles, and the IQR

- The **interquartile range** is the difference between the third and the first quartile: $\mathrm{IQR} = Q_3 - Q_1$
- Put differently, it corresponds to the **bounds** of the set which contains the **middle half** of the distribution



3. Spread

3.2. Variance and standard deviation

- The **variance** is a way to quantify how the values of a variable tend to **deviate** from their **mean**
  - If values tend to be **close to the mean**, then the **spread is low**
  - If values tend to be far **from the mean**, then the **spread is large**

- Can we just take the **average deviation** from the mean?

x	mean(x)	x - mean(x)
1	2.5	-1.5
4	2.5	1.5
-3	2.5	-5.5
8	2.5	5.5

- By construction it would **always be 0**: values above and under the mean compensate
  - But we can use the **absolute value** of each deviation: $|x_i - \bar{x}|$
  - Or their **square**: $(x_i - \bar{x})^2$

3. Spread

3.2. Variance and standard deviation

- This is how the **variance** is computed: by **averaging the squared deviations from the mean**

$$\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Because the **variance** is a **sum of squares**, it can get **quite big** compared to the other statistics like the mean, the median or the interquartile range.
    - To express the spread in the **same unit** as the data, we can take the **square root** of the variance, which is called the **standard deviation**
    - In a way, *the standard deviation is to the mean what the IQR is to the median*

$$\mathrm{SD}(x) = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

3. Spread

3.3. Standard deviation vs. interquartile range

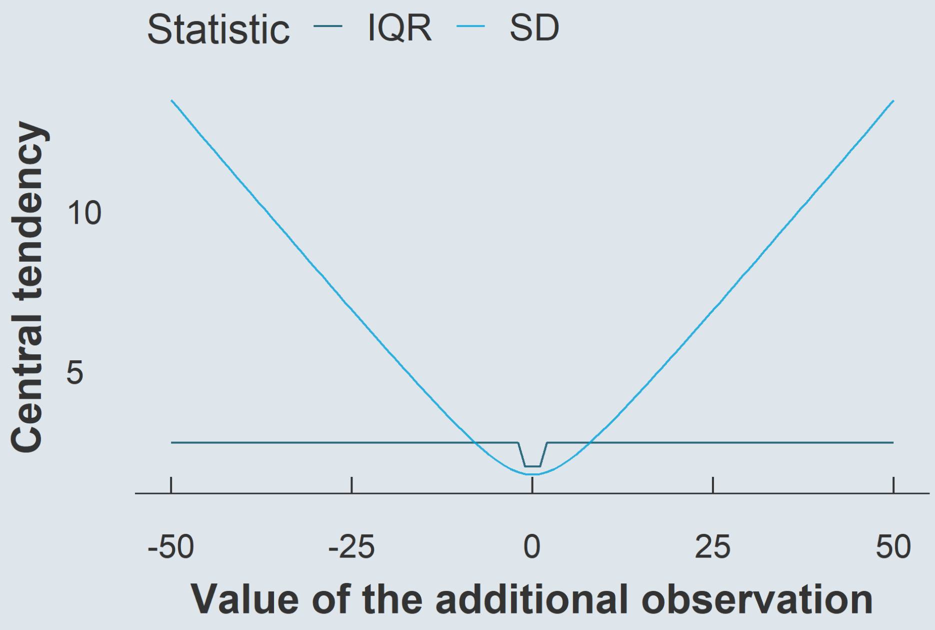
- Remember that the median is **less sensitive** than the mean to thick tails and outliers
- This is also the case for the **IQR** relative to the **standard deviation**

Let's go back to our previous example!

- Consider the following variable:

-3	-2	-2	-1	-1	-1	0	1	1	1	2	2	3

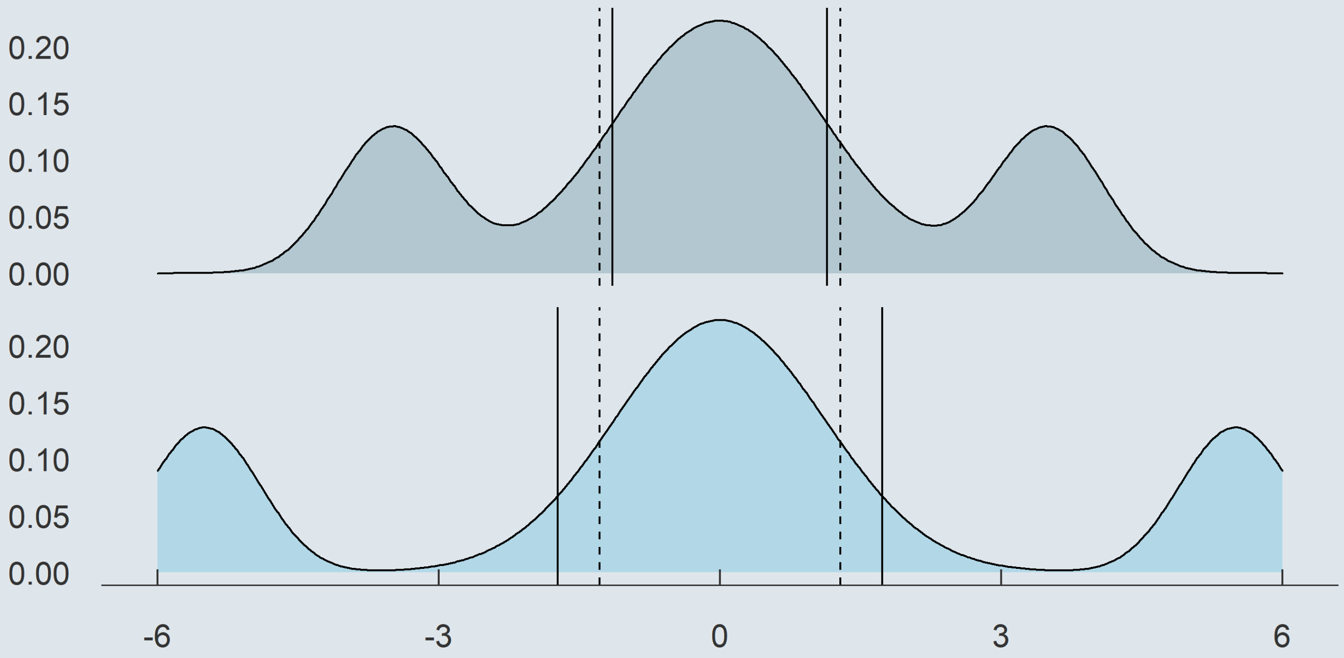
- How would the standard deviation and the IQR **react** if we were to **add one single observation**?
    - We can plot the value of the additional observation on the x axis and the value of the mean and the median on the y axis



3. Spread

3.3. Standard deviation vs. interquartile range

- But like for the median vs. the mean, it does **not** mean that one is **better than the other**
  - They just **capture different things**



- These two distributions
  - Have the **same interquartile range**
  - Have **different standard deviations**

3. Spread

3.3. Standard deviation vs. interquartile range: in R

- Both statistics have **dedicated R functions**

```
variable <- c(0, 1, 3, 4, 6, 7, 8, 10, 11)
c(sd(variable), IQR(variable))
```

```
## [1] 3.844188 5.000000
```

- You can obtain the **quantiles** of a variable using the `quantile()` function

```
quantile(variable)
```

```
##   0%  25%  50%  75% 100% 
##    0    3    6    8   11
```

→ See the help file ?quantile() for more info on quantile computation

Practice

05:00

➜ **Consider the following variable**

```
variable <- c(1, 3, 8, 4, 9, 5, 3, 8, 8, 7, 4, 9,
              6, 5, 1, 999, 1, 2, 4, 5, 6, 9, 7, NA)
```

1) Copy/paste the line above into an .R script and run it

2) Compute the mean of this distribution

3) Compute the three quartiles of this distribution

4) Compute the interquartile range of this distribution

You've got 5 minutes!

Solution

1) Compute the mean of this distribution

```
mean(variable, na.rm = T)
```

```
## [1] 48.43478
```

2) Compute the three quartiles

```
quartiles <- quantile(variable, 1:3/4, na.rm = T, names = F)
quartiles
```

```
## [1] 3.5 5.0 8.0
```

3) Compute the inter quartile range

```
quartiles[3] - quartiles[1]
```

```
## [1] 4.5
```

→ The outlier 999 pulls the mean outside of the IQR! Descriptive statistics is a good tool to make sure the data is clean

Overview

Overview

4. Inference

4.1. Data generating process

- In **practice**, we manipulate **concrete** variables such as age, sex, earnings, etc.
  - But on the **theoretical** side, we denote such variables with an **abstract** letter like x

- In Statistics and Econometrics, we indeed use letters like x to denote what we call **random variables**
  - These variables can take values according to a **data generating process** (DGP)
  - The data generating process is the *mechanism that causes the data to be the way we observe it*

- For instance your grades can be seen as a random variable
  - Which takes given values according to an unknown data generating process
  - The DGP probably depends on your effort, your background, many environmental factors, ...

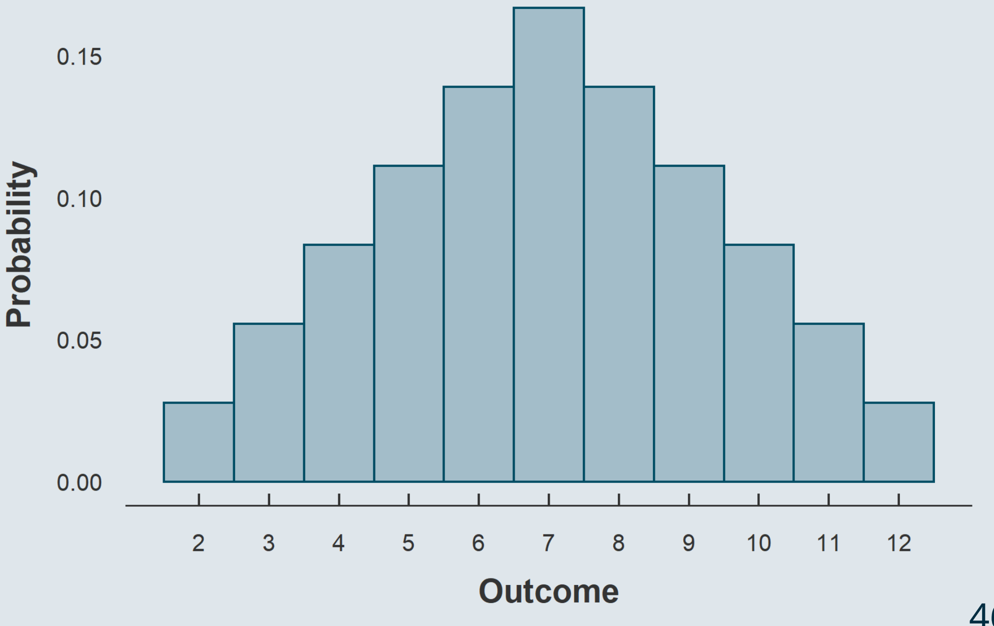
- With descriptive statistics, we actually **infer** properties of the DGP **given the outcomes** we observe
  - **Like backward engineering**, from the output we try to understand the process
  - One **crucial implication** is that the mean we compute is just an **estimation** of the parameter of the DGP we're interested in

4. Inference

4.1. Data generating process

- Consider for instance the **outcome of two dice** as a random variable
  - Contrarily to the variables we usually study, **we know the DGP** of this one
- The DGP causes our random variable to take the following values with the following probabilities:

2 - 1/36 (⚀⚀)
3 - 2/36 (⚀⚁ - ⚁⚀)
4 - 3/36 (⚀⚂ - ⚂⚀ - ⚁⚁)
5 - 4/36 (⚁⚂ - ⚂⚁ - ⚁⚃ - ⚃⚁)
6 - 5/36 (⚂⚂ - ⚁⚃ - ⚃⚁ - ⚄⚀ - ⚀⚄)
7 - 6/36 (⚂⚃ - ⚃⚂ - ⚄⚁ - ⚁⚄ - ⚀⚅ - ⚅⚀)
8 - 5/36 (⚃⚃ - ⚄⚂ - ⚂⚄ - ⚅⚁ - ⚁⚅)
9 - 4/36 (⚃⚄ - ⚄⚃ - ⚅⚂ - ⚂⚅)
10 - 3/36 (⚅⚃ - ⚃⚅ - ⚄⚄)
11 - 2/36 (⚅⚄ - ⚄⚅)
12 - 1/36 (⚅⚅)



4. Inference

4.2. Empirical vs. theoretical moments

- Because we know the data generating process of our random variable, we can compute its **expected value**:

$$\mathrm{E}(x) = \frac{(2 \times 1) + (3 \times 2) + (4 \times 3) + (5 \times 4) + (6 \times 5) + (7 \times 6)}{36} + \frac{(8 \times 5) + (9 \times 4) + (10 \times 3) + (11 \times 2) + (12 \times 1)}{36} = \frac{252}{36} = 7$$

- This is the parameter we are actually interested in
  - The **expected value** is what we call a **theoretical moment** *(the first one)*
  - While the **mean** is the corresponding **empirical moment**

- **How confident** to be in our estimate of the expected value (i.e., *the mean*) depends on the **sample size**
  - For a given number of draws the mean won't necessarily be exactly 7
  - But if we were to do **infinitely many draws**, the mean would **converge** towards 7 *(Law of Large Numbers)*

4. Inference

4.2. Empirical vs. theoretical moments

- Just like the **mean** that we compute empirically is an estimate of the **first moment** of the distribution,
  - the **variance** that we compute empirically is an estimate of the **second moment** of the distribution

	Theoretical moment	**Empirical moment**
First moment:	$\mathrm{E}(x_{\text{discrete}}) = \sum_{i=1}^{k} x_i p_i$ $\mathrm{E}(x_{\text{continuous}}) = \int_{\mathrm{R}} x f(x) dx$	$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Second moment:	$\mathrm{Var}(x) = \mathrm{E}\left[(x - \mathrm{E}(x))^2\right] \equiv \sigma^2$	$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$

4. Inference

4.2. Empirical vs. theoretical moments

Expected value operations

$$\begin{aligned} \mathrm{E}[X+Y] =& \mathrm{E}[X] + \mathrm{E}[Y] \\ \mathrm{E}[aX] =& a\mathrm{E}[X] \\ \mathrm{E}[a] =& a \\ \mathrm{E}[\mathrm{E}[X]] =& \mathrm{E}[X] \\ \mathrm{E}[XY] \neq& \mathrm{E}[X]\mathrm{E}[Y] \text{ unless } X \perp Y \end{aligned}$$

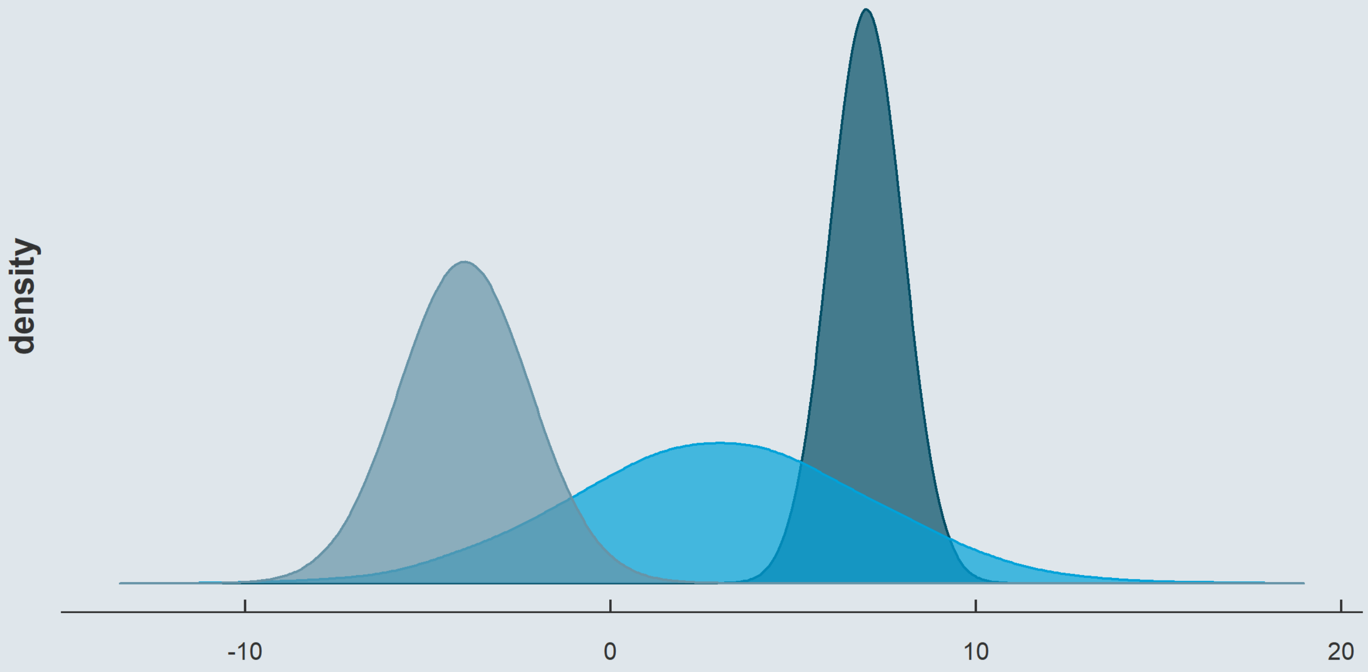
Variance operations

$$\begin{aligned} \mathrm{Var}(X) >& 0 \quad \mathrm{Var}(a) = 0 \\ \mathrm{Var}(X+a) =& \mathrm{Var}(X) \\ \mathrm{Var}(aX) =& a^2\mathrm{Var}(X) \\ \mathrm{Var}(aX+bY) =& a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + \\ & 2ab\mathrm{Cov}(X,Y) \\ \mathrm{Var}(aX-bY) =& a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) - \\ & 2ab\mathrm{Cov}(X,Y) \end{aligned}$$

4. Inference

4.3. Confidence interval

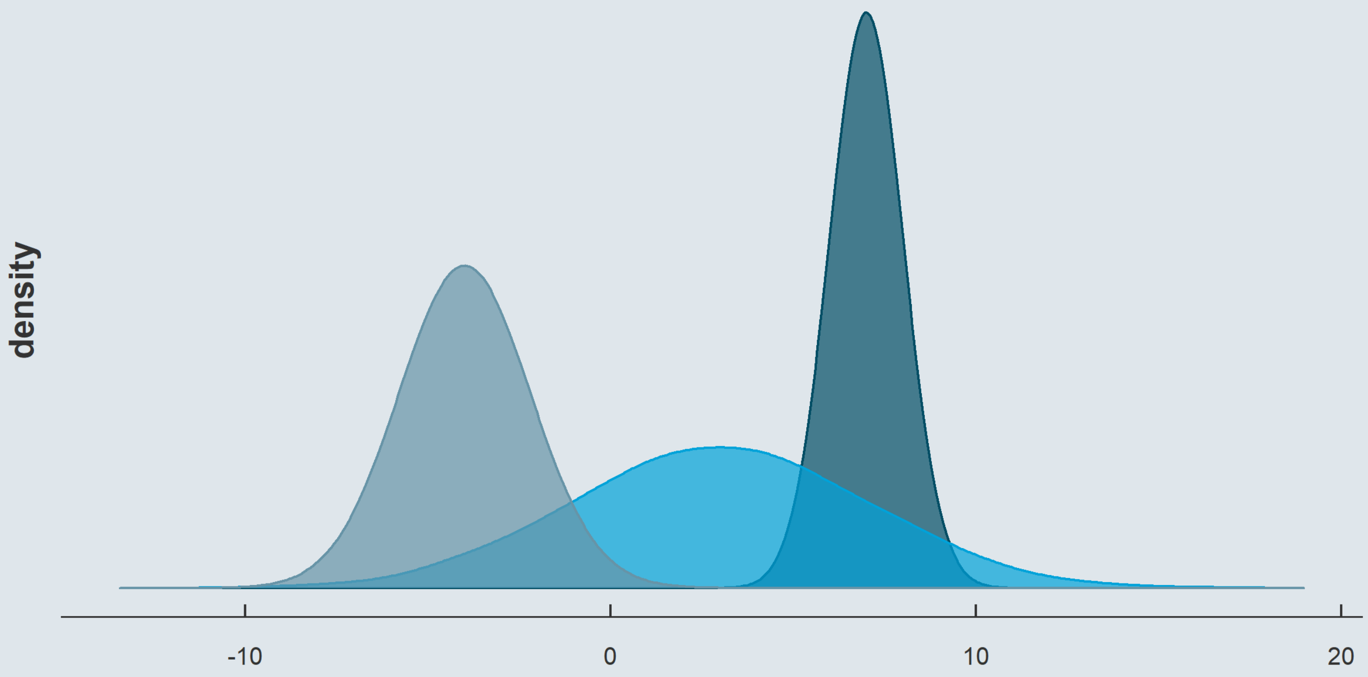
- Because the mean is an empirical **estimation** of the theoretical expected value
  - We need a measure of the **confidence** we can have in this estimations
  - This is something we can do as long as our variable is **normally distributed** *(bell-shaped)*



4. Inference

4.3. Confidence interval

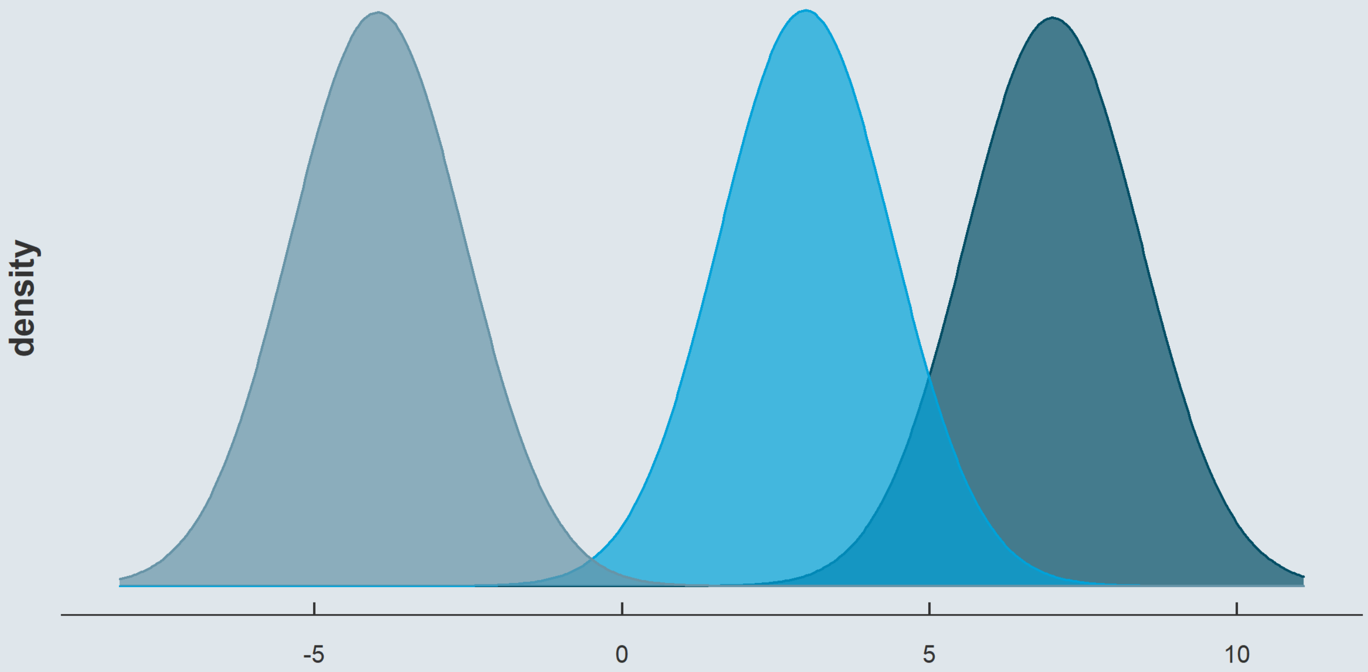
- Indeed, with such distributions we can recover **something we know**



4. Inference

4.3. Confidence interval

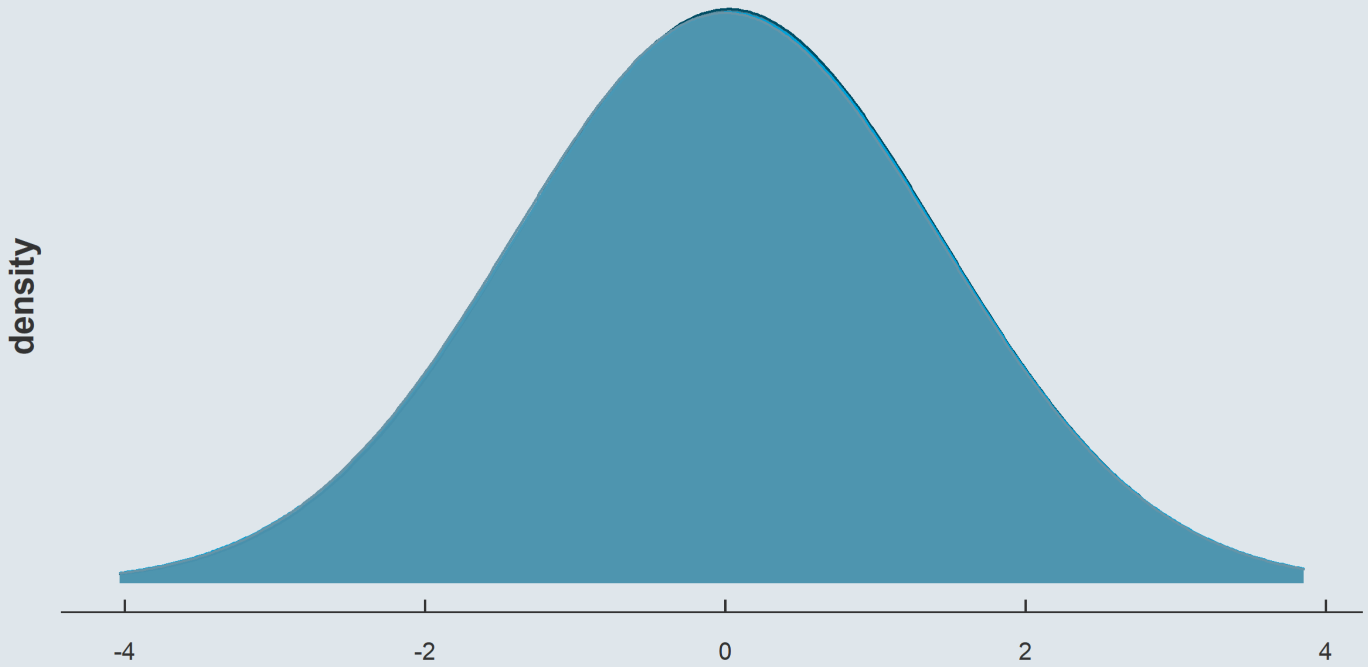
- Indeed, with such distributions we can recover **something we know**
    - If we **divide** all the values of the variable by its **standard deviation**, the **variance becomes 1**



4. Inference

4.3. Confidence interval

- Indeed, with such distributions we can recover **something we know**
  - If we **divide** all the values of the variable by its **standard deviation**, the **variance becomes 1**
  - If we **subtract the mean** from all the values of the variable, the **mean becomes 0**



4. Inference

4.3. Confidence interval

- In mathematical notation, what we just saw writes:

$$\frac{x - \text{E}(x)}{\text{SD}(x)} \sim \mathcal{N}(0, 1)$$

- And if we compute means on random draws of x
  - These means would behave the same way:

$$\frac{\bar{x} - \text{E}(x)}{\text{SD}(x)} \sim \mathcal{N}(0, 1)$$

  - This is actually **true with** a theoretical **infinite sample** of x (i.e., the DGP)
  - But **in practice**, we work with **finite samples** so things work slightly differently

- When we have a limited number n of observations:
  - We standardize using the **standard *error*** of the mean $\text{SE}(x) = \text{SD}(x)/\sqrt{n}$
  - And we know that:

$$\frac{\bar{x} - \text{E}(x)}{\text{SE}(x)} \equiv t \sim t_{n-1}$$

  - Where t reads *"t-stat"* and t_{n-1} denotes a Student's t distribution with $n-1$ degrees of freedom

4. Inference

4.3. Confidence interval

- The Student's t distribution is **very similar to the normal** distribution
  - It is just **a bit flatter** when n is low
  - But it **converges quickly** to a normal distribution as $n \to \infty$

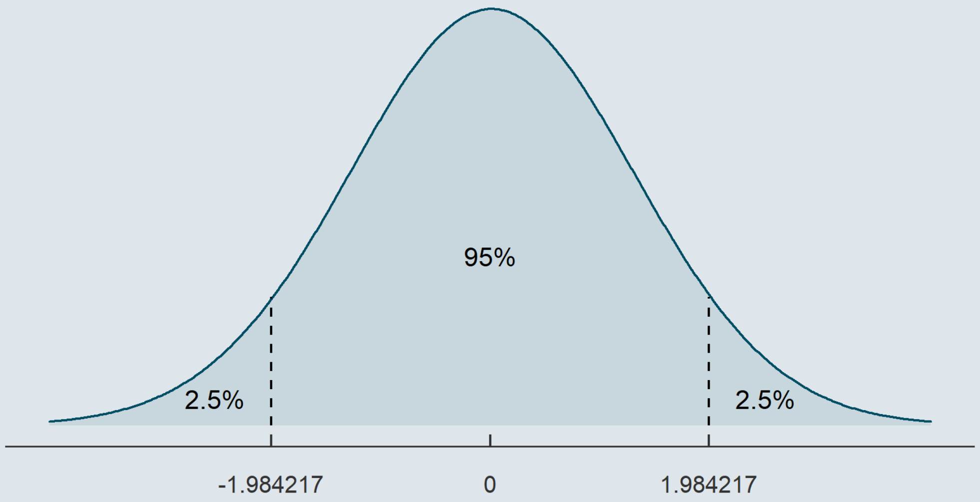
4. Inference

4.3. Confidence interval

- The good news is that:
  - Because **we know** how $t \equiv \frac{\bar{x} - \text{E}(x)}{\text{SE}(x)}$ is distributed ($\sim t_{n-1}$)
  - We also know what are **the chances** that $\frac{\bar{x} - \text{E}(x)}{\text{SE}(x)}$ takes **certain values**

- Consider a variable $x \sim \mathcal{N}(\text{E}(x), \text{SD}(x)^2)$
  - We know that with $n = 100$, $\frac{\bar{x} - \text{E}(x)}{\text{SD}(x)/\sqrt{n}} \sim t_{99}$
  - And we know between which values lies a given share of the t_{99} distribution
  - For instance, 95% of the distribution lie in

$$[-t_{99,97.5\%};\ t_{99,97.5\%}] \approx [-1.98;\ 1.98]$$



4. Inference

4.3. Confidence interval

- In mathematical notation, what the previous graph shows writes:

$$\Pr\left[-t_{99,97.5\%} \leq \frac{\bar{x} - \mathrm{E}(x)}{\mathrm{SE}(x)} \leq t_{99,97.5\%}\right] = 95\%$$

- Rearranging the terms yields:

$$\Pr\left[\bar{x} - t_{99,97.5\%} \times \mathrm{SE}(x) \leq \mathrm{E}(x) \leq \bar{x} + t_{99,97.5\%} \times \mathrm{SE}(x)\right] = 95\%$$

- Thus, we can say that there's 95% chance for $\mathrm{E}(x)$ to be within:

$$\bar{x} \pm t_{99,97.5\%} \times \mathrm{SE}(x)$$

➜ This is our 95% confidence interval of the mean!

4. Inference

4.3. Confidence interval

- We can apply this calculations **in R** to get a **95% CI of the mean** of the grade distribution

```r
grades <- c(20, 20, 17.5, 17.5, 16, 16, 16, 14.5, 14.5, 19.5, 19.5, 18.5, 18.5, 18.5)

# Mean, standard deviation, and n
mean <- mean(grades)
sd <- sd(grades)
n <- length(grades)

# Standard error
se <- sd / sqrt(n)

# t-stat
t <- qt(.975, n - 1) # qt returns t-stat from 1 - ((1 - CL) / 2) and degrees of freedom

# Confidene interval
c(mean - t*se, mean + t*se)
```

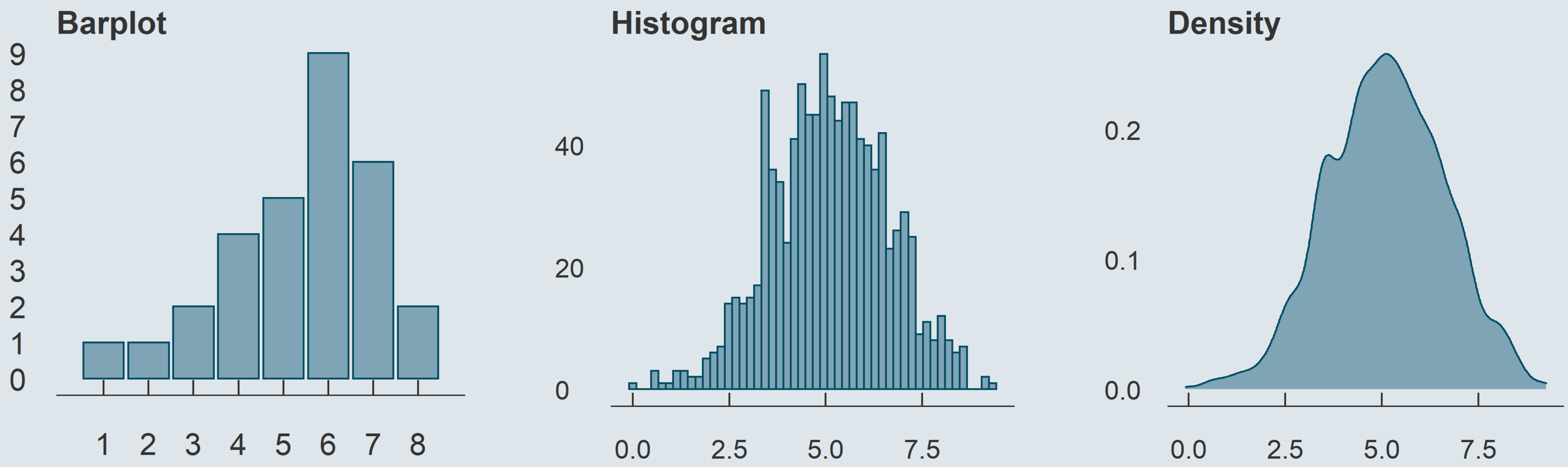
```
## [1] 16.49665 18.71764
```

Overview

5. Wrap up!

1. Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are

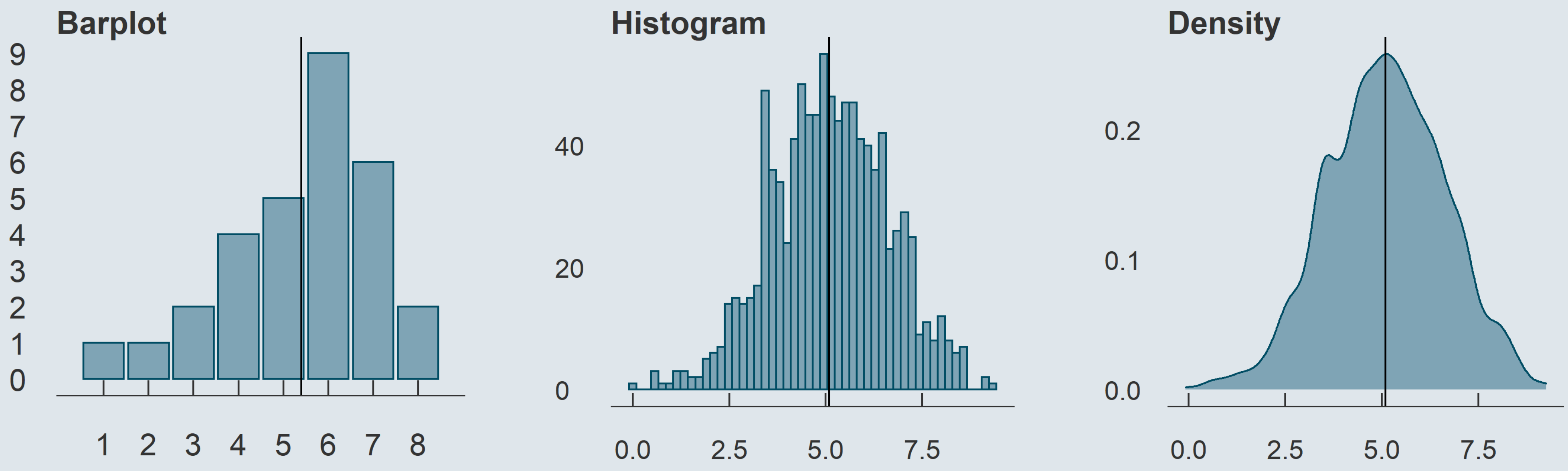


- We can describe a distribution with:

5. Wrap up!

1. Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are

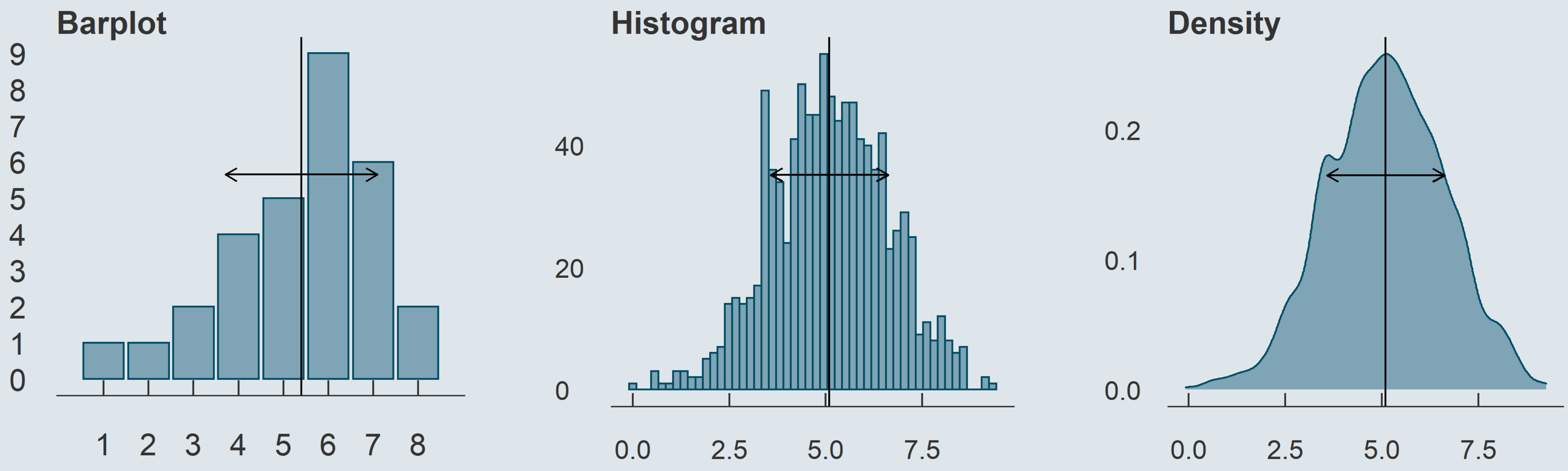


- We can describe a distribution with:
  - Its **central tendency**

5. Wrap up!

1. Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are



- We can describe a distribution with:
  - Its **central tendency**
  - And its **spread**

5. Wrap up!

2. Central tendency

- The **mean** is the sum of all values divided by the number of observations

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The **median** is the value that divides the (sorted) distribution into two groups of equal size

$$\text{Med}(x) = \begin{cases} x[\frac{N+1}{2}] & \text{if } N \text{ is odd} \\ \frac{x[\frac{N}{2}]+x[\frac{N}{2}+1]}{2} & \text{if } N \text{ is even} \end{cases}$$

3. Spread

- The **standard deviation** is square root of the average squared deviation from the mean

$$\text{SD}(x) = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- The **interquartile range** is the difference between the maximum and the minimum value from the middle half of the distribution

$$\text{IQR} = Q_3 - Q_1$$

5. Wrap up!

4. Inference

- In Statistics, we view variables as a given realization of a **data generating process**
  - Hence, the **mean** is what we call an **empirical moment**, which is an **estimation**...
  - ... of the **expected value**, the **theoretical moment** of the DGP we're interested in

- To know how confident we can be in this estimation, we need to compute a **confidence interval**

$$[\bar{x} - t_{n-1,\,97.5\%} \times \frac{\text{SD}(x)}{\sqrt{n}}\,;\; \bar{x} + t_{n-1,\,97.5\%} \times \frac{\text{SD}(x)}{\sqrt{n}}]$$

  - It gets **larger** as the **variance** of the distribution of x increases
  - And gets **smaller** as the **sample size** n increases