Week 2: Descriptive Statistics

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

So Far

- Quantitative research involves collecting *data* - a *sample* of observations selected from a larger *population*, in which one or more *variables* are measured for each *observation*.
- The goal of collecting data is usually to calculate *statistics* which can be used to infer *parameters* of a population.
- Steps of quantitative study.
- Hypothesised relationship.

Topics for Today

- Measuring variables
- Scales
- Descriptive statistics
- Visualising data
- Measures of central tendency
- Measures of variability

Review:
Steps of Quantitative Study

- Identify your problem (formulate a research question);
- Specify your dependent variable (Y);
- Explain why it is a significant problem (i.e., why should anybody care);
- Explain how much we already know about the problem (the literature review);
- Formulate one or more hypotheses;
- Design a model to test your hypotheses - to explain why or how your dependent variable varies the way it does;
- Identify a dataset suitable to testing your model/hypotheses;
- Measure your variables;
- Perform statistical tests on your data.

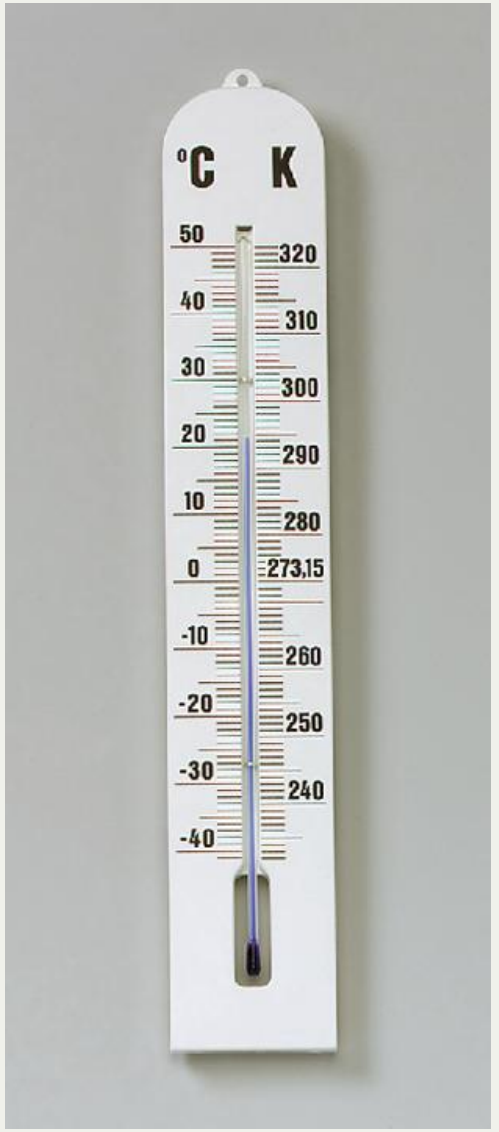
Review: Hypothesised relationship

- A hypothesis can often be described like this: $X \rightarrow Y : Z$
- Or "Y depends on X in the presence of Z"
- Or "Y is associated with X conditional on Z"
- Re-writing as an equation gives us: $Y = X + Z + \epsilon$
- ϵ means that the relationship is not perfect (one-to-one)!
- There is always some error involved.
- More on Z in future lectures.

Values of Y and X

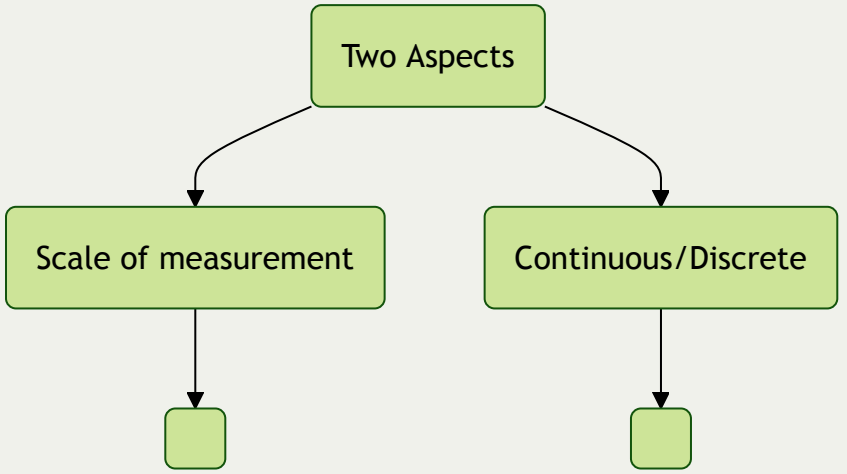
- Our **dependent** (Y) and **independent** (X) **variables** take different **values**.
- What are these values?
- It depends on how they were measured!

Our ideal 🌡️



Wikipedia

Types of Variables



Measurement Scales

Scale	Descriptive Statistic	Examples
Nominal	Mode	vote choice, religion
Ordinal	Mode/Median	trust in government
Interval	Mode/Median/Mean	age, income

Nominal/Categorical Scale

- Unordered categorical variables are said to be measured on **nominal scale**.
- E.g. regime type, gender, occupation.
- Values taken by such variable are called **levels** of the scale.
- However, no level is greater or smaller than the other level.
- In practice, models rely on numeric values even for categorical variables.
- E.g. 0 for men, 1 for women.
- Variables that have only 2 levels are called **binary** (dichotomous).

Ordinal Scale

- Variables with ordered categories are said to be measured on **ordinal scale**.
- E.g. social class, political preferences (on left-right scale).
- The levels of such a variable have a natural **ordering**.
- Such variables are often treated as nominal/categorical.
- However, they also closely resemble a quantitative variable.

Interval Scale

- Quantitative variables are said to be measured on **interval scale**.
- E.g. age, income, election turnout.
- There is a specific numeric distance or **interval** between values.
- Values of such variables can be directly compared in terms of magnitude.
- Differences (intervals) have the same meaning across the scale.

Discrete and Continuous Variables

- **Discrete** variables take values from a limited set of possibilities.
- E.g. vote choice, education, number of children (any 'number of …' variable).
- **Continuous** variables can take any value, which is a real number.
- E.g. election turnout, income, GDP.
- All categorical and ordinal variables are discrete.
- Interval variables can be discrete or continuous.

Why Does It Matter?

- Type of variable determines what statistical test applicable to your data.
- The borderline for some variables might be fuzzy (e.g. ordinal scale).
- But all tests rely on some **assumptions**.
- It is important to keep in mind how heroic these assumptions are.

Why Statistical Methods? 📈

- Statistical methods allow to *describe* collected data (sample).
- They also provide tools to *infer* properties of the population (from which the sample was drawn).
- Hence, we often talk about
  - **descriptive statistics** and
  - **inferential statistics**.
- Let's first look at descriptive statistics.
- But we will spend a good portion of this module discussing inference!

Descriptive Statistics

- First step after acquiring any data 💾.
- Check whether data makes sense 🤔.
- E.g:
  - Summarise (describe) key variables of interest.
  - Make frequency tables for categorical variables.
  - Check distributions of quantitative variables 📊.

Descriptive Statistics for Categorical Data

- Tabulate the variable (count the number of cases falling into each category).
- **Frequency distribution** - number of observations at each level of the variable.
- **Relative frequency distribution** - their respective proportions/percentages.
- *Proportions* must sum up to 1.
- *Percentage* must sum up to 100%.

Example: Democracy in 2020 

```r
democracy_2020 <- read.csv("../data/democracy_2020.csv")
```

```r
# Tabulate data
table(democracy_2020$democracy)
```

```
 0   1
77 118
```

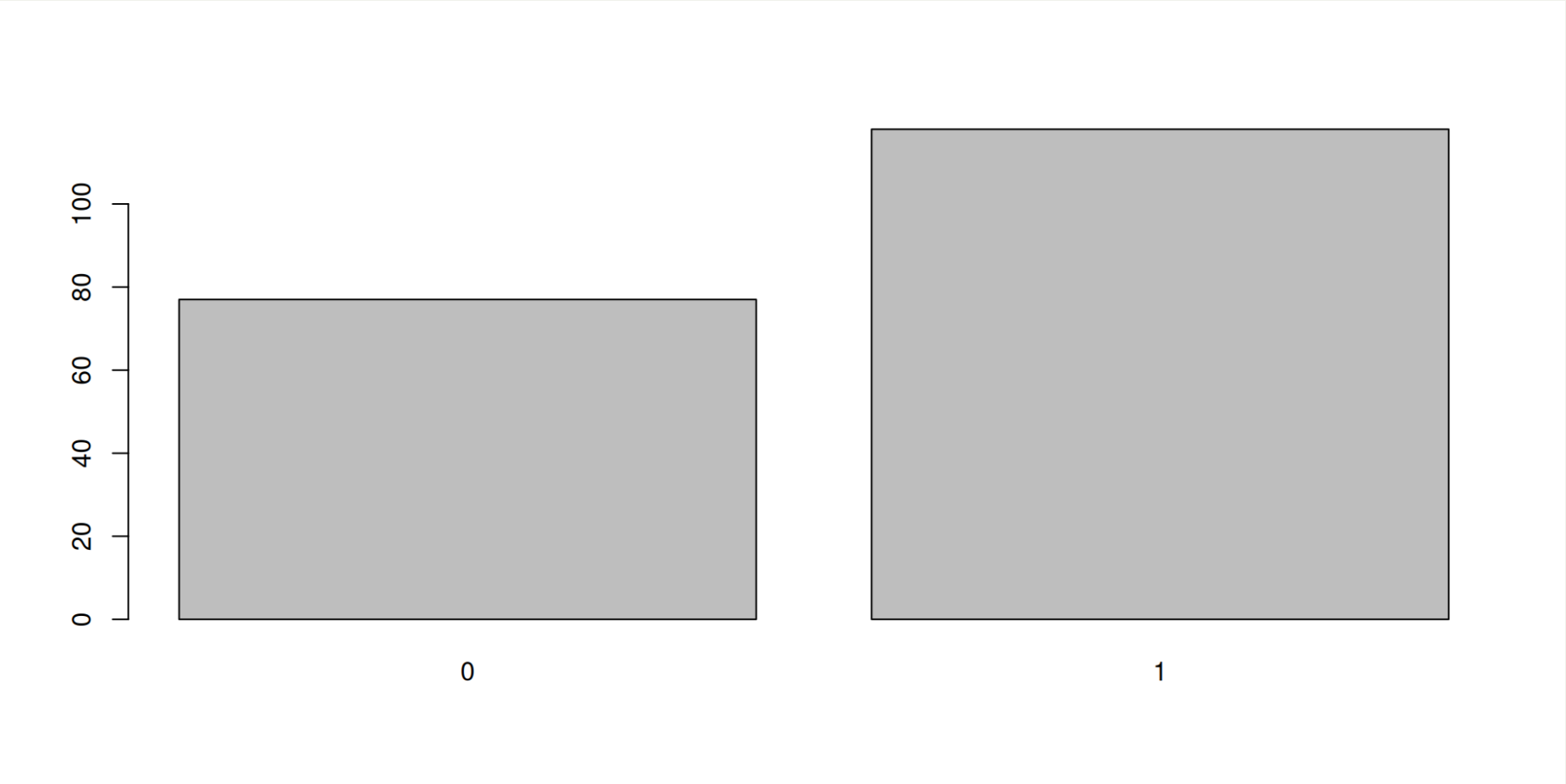
```r
# Calculate proportions
prop.table(table(democracy_2020$democracy))
```

```
        0         1
0.3948718 0.6051282
```

> **Source**
>
> Boix, Miller and Rosato (2013), (2020)

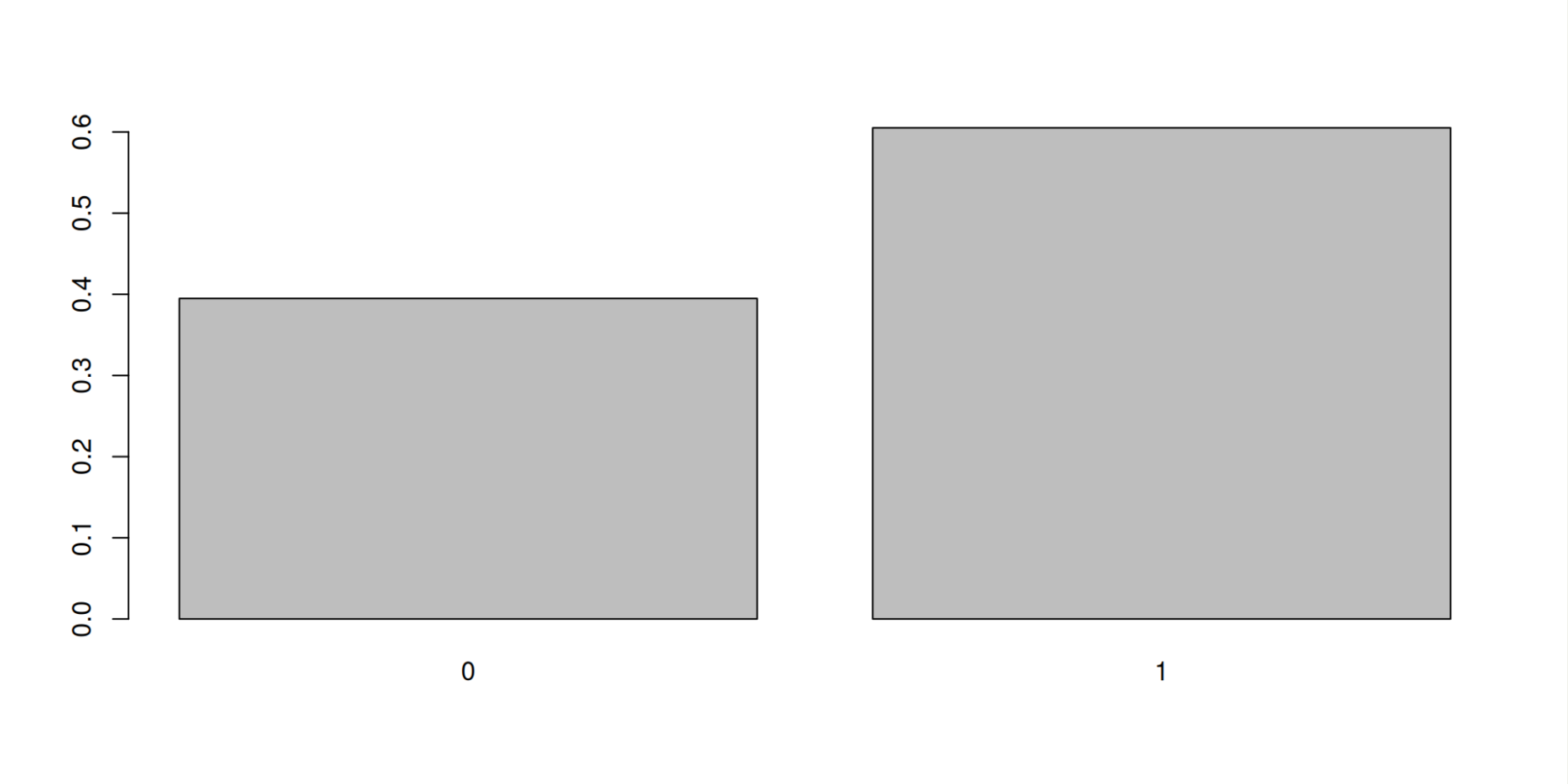
Visualising Categorical Data: Bar Graph

```
1 barplot(table(democracy_2020$democracy))
```



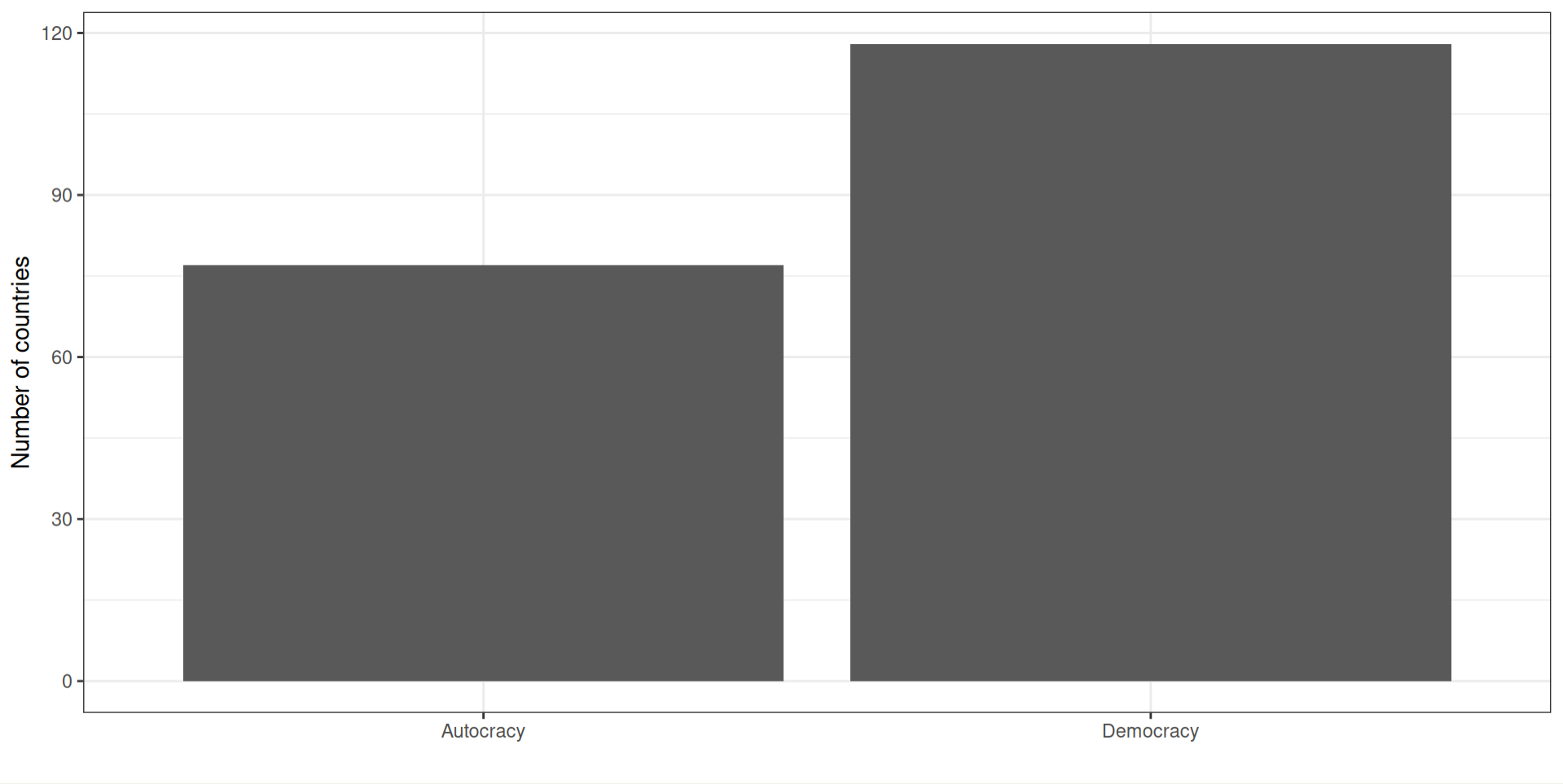
Visualising Categorical Data: Bar Graph

```
1 barplot(prop.table(table(democracy_2020$democracy)))
```



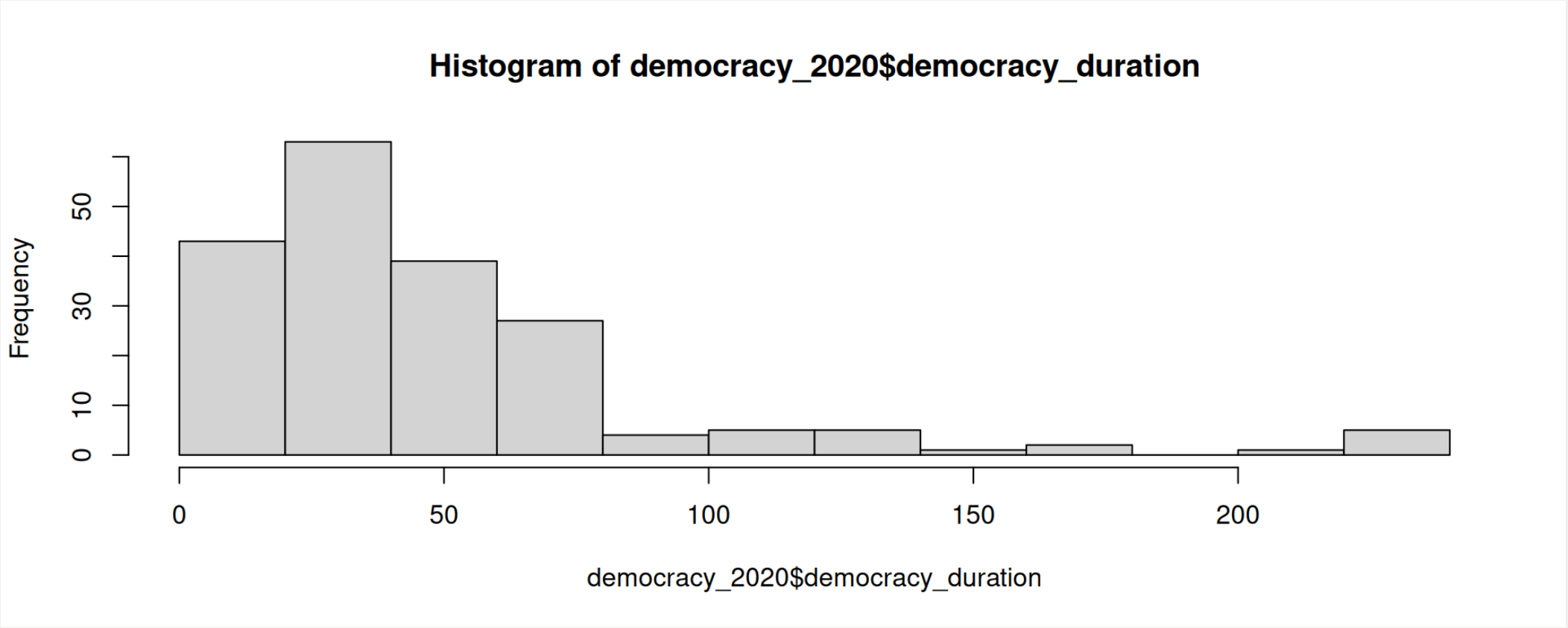
Prettyfing Bar Graph

Plot | Code



Visualising Quantitative Data: Histogram

```
1 hist(democracy_2020$democracy_duration)
```



Descriptive Statistics for Quantitative Data

- Key features to describe numerically:
  - **Centre** of the data - a typical observation
  - **Variability** of the data - the spread around the centre
- **Measures of central tendency** describe the centre.
- **Measures of variability** describe the variability.

Measures of Central Tendency

- **Mode** - the most common value.
- **Median** - the value of the observation in the middle.
- **Mean** - the average value of all observation.
- Measures of central tendency are some of sample *statistics*.
- Sample statistics are often our best estimates of population *parameters*.

Some Notation

- We refer to our **sample size** as N.
- If $N = 100$, we have 100 observations of the same variable:

$$(Y_1, Y_2, Y_3, \ldots, Y_N)$$

- We can then refer to the sum of that variable as:

$$Y_1 + Y_2 + Y_3 + \ldots + Y_N$$

- But it gets too cumbersome as N grows large.
- So, instead, we can write it as:

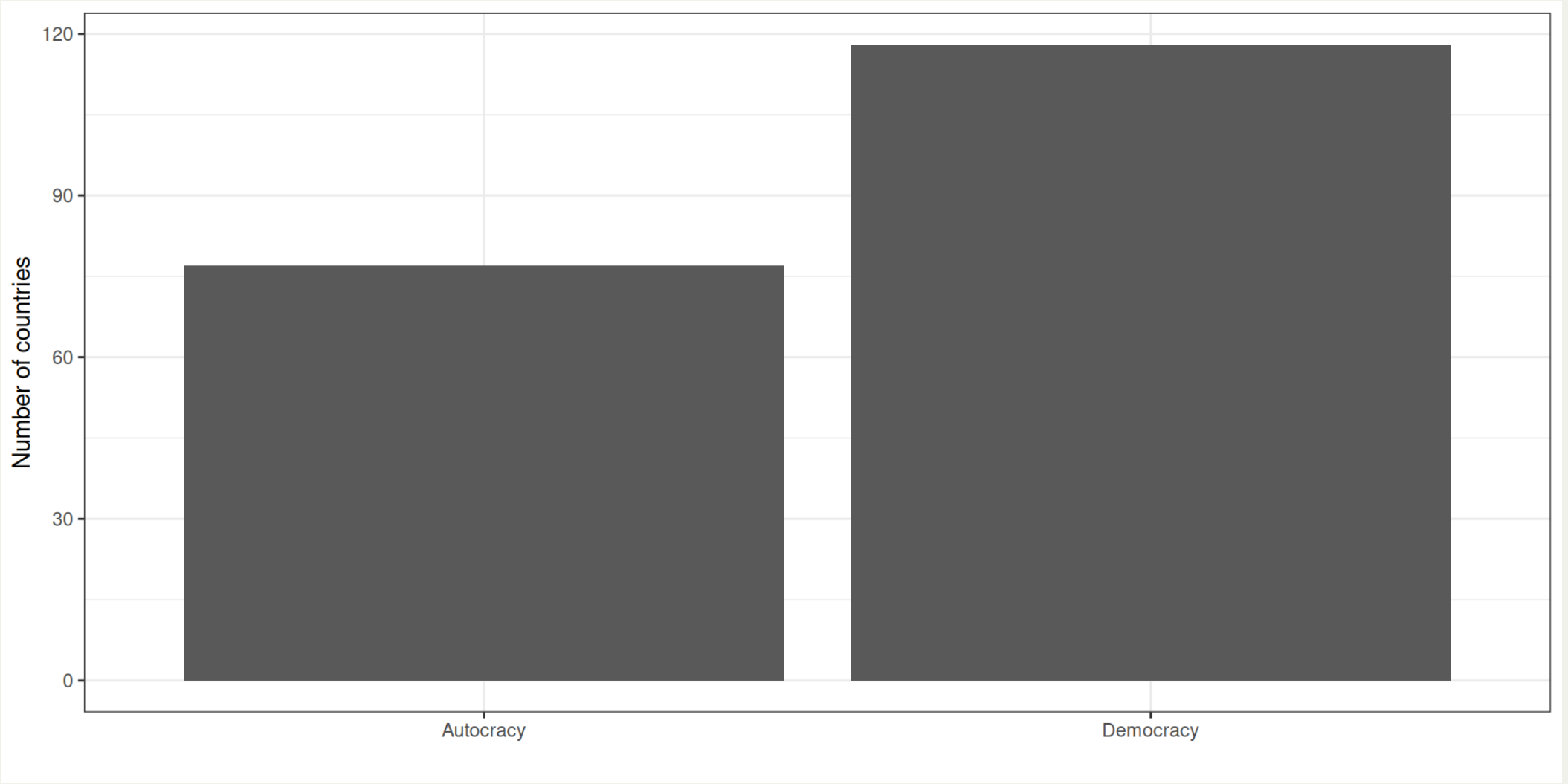
$$\Sigma_{i=1}^{N} Y_i = Y_1 + Y_2 + Y_3 + \ldots + Y_N$$

where $\Sigma_{i=1}^{N} Y_i$ means "sum up all the values of Y starting at 1 and ending at N"

Mode

- The value that occurs most often (has the highest frequency).
- It is appropriate for all scales of measurement.
- It is the only appropriate measure of central tendency for a nominal (categorical) variable.
- Most useful for discrete variables with few distinct values.

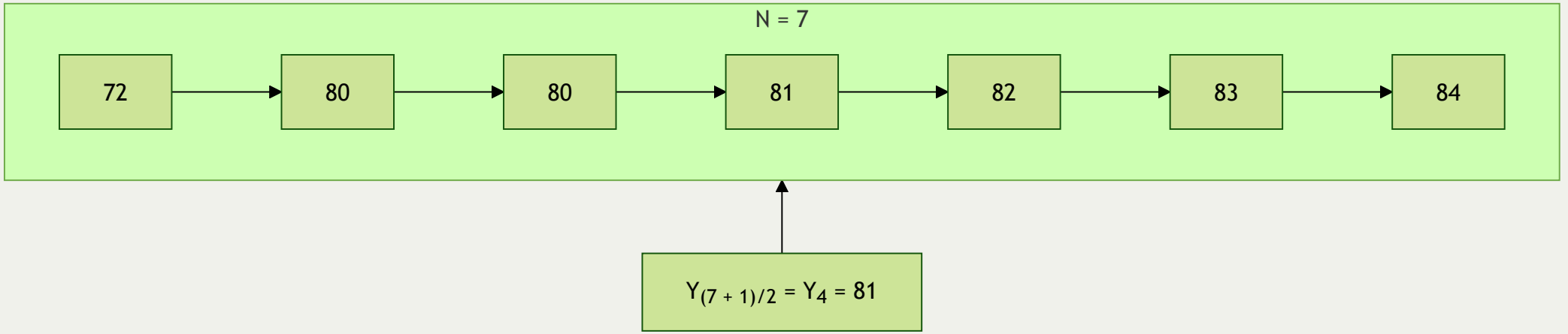
What Is The Mode Here?

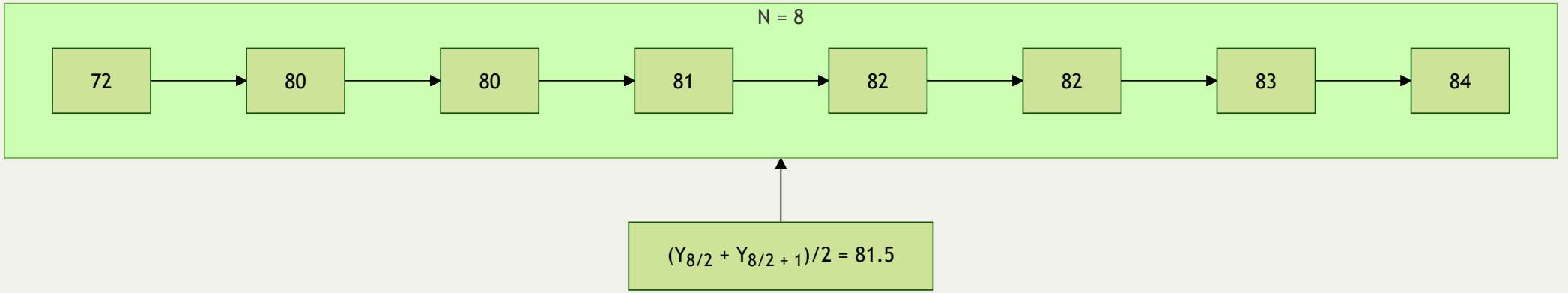


Median

- The value that falls in the middle of an ordered sample.

$$Median = \begin{cases} Y_{(N+1)/2} & \text{when N is odd} \\ \frac{1}{2}(Y_{(N/2} + Y_{N/2+1}) & \text{when N is even} \end{cases}$$




N = 8
72
80
80
81
82
82
83
84
$(Y_{8/2} + Y_{8/2+1})/2 = 81.5$

Mean

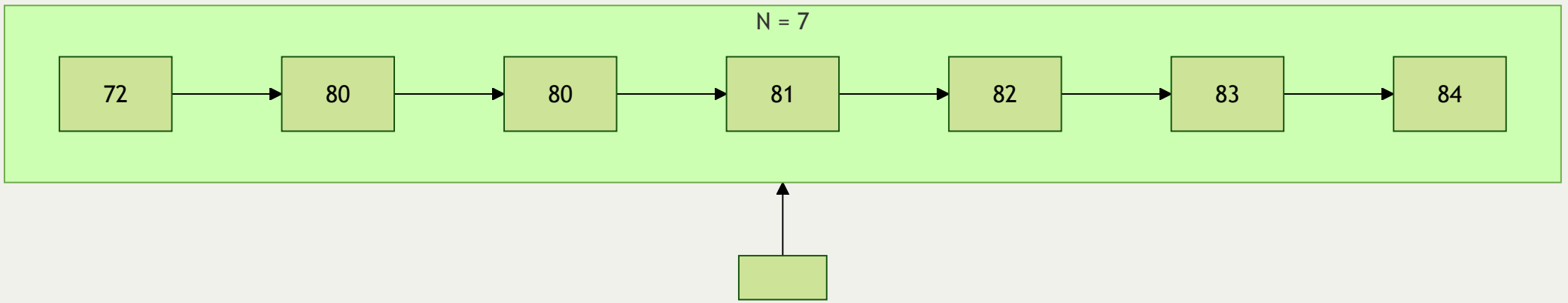
- Sum of observations' values divided by the number of observations.
- The mean is only appropriate for quantitative variables.
- The mean is often called the **average**.

```
1 y <- c(72, 80, 80, 81, 82, 83, 84)
2 sum(y)/length(y)
```

```
[1] 80.28571
```

```
1 mean(y)
```

```
[1] 80.28571
```



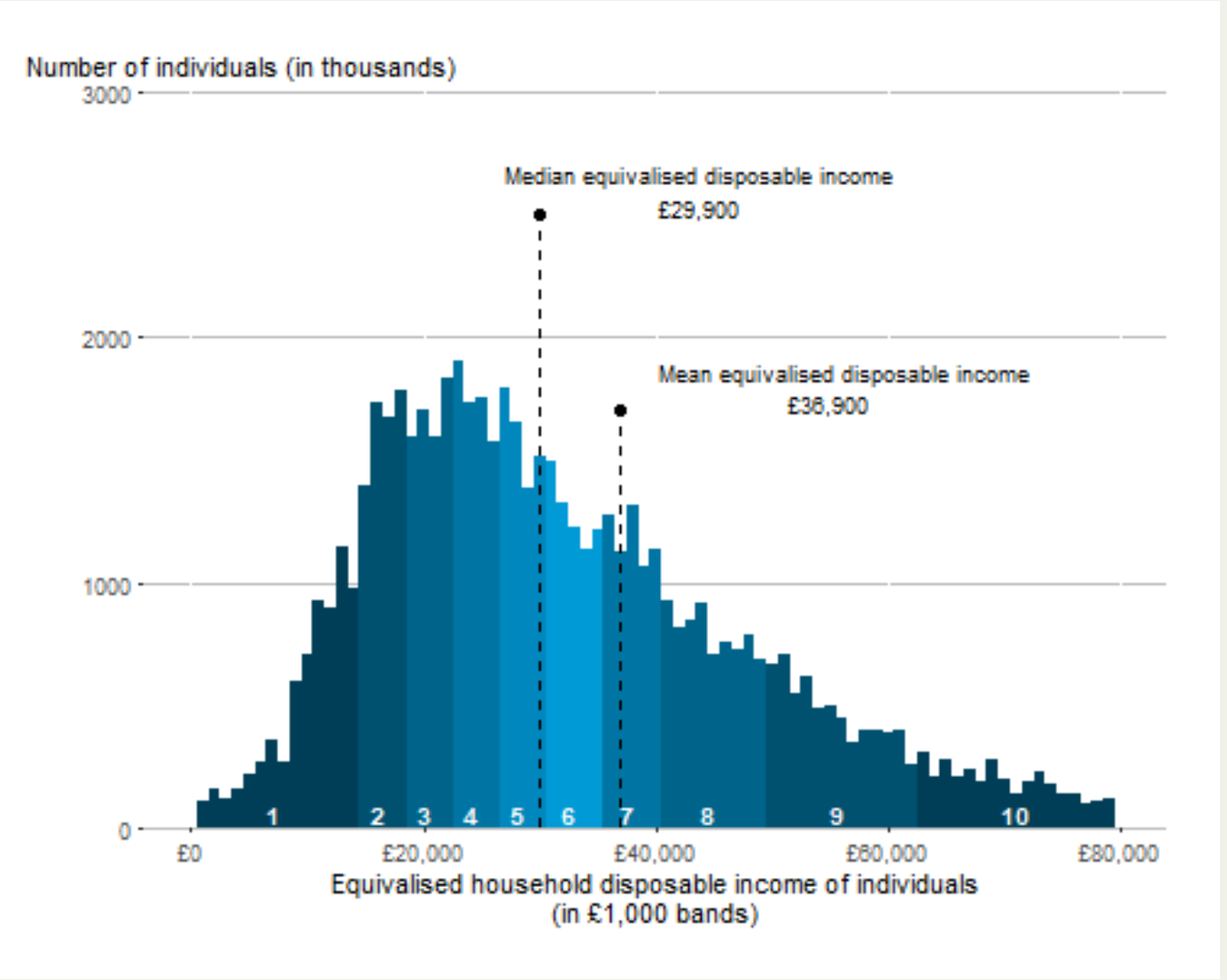
Mean as Expected Value

- Statistically, mean is also the **expected value** of some variable.
- E.g. for dependent (response) variable $E(Y) = \bar{Y}$ (pronounced y-bar).
- N is the sample size, subscript $1, 2, \ldots, N$ is the case (observation) number.
- We can then write the mean as:
  - $\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_N}{N}$
  - $\bar{Y} = \frac{\sum Y_i}{N}$
  - $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$
- Note that these 3 notations correspond to the same calculation!

Mean vs Median

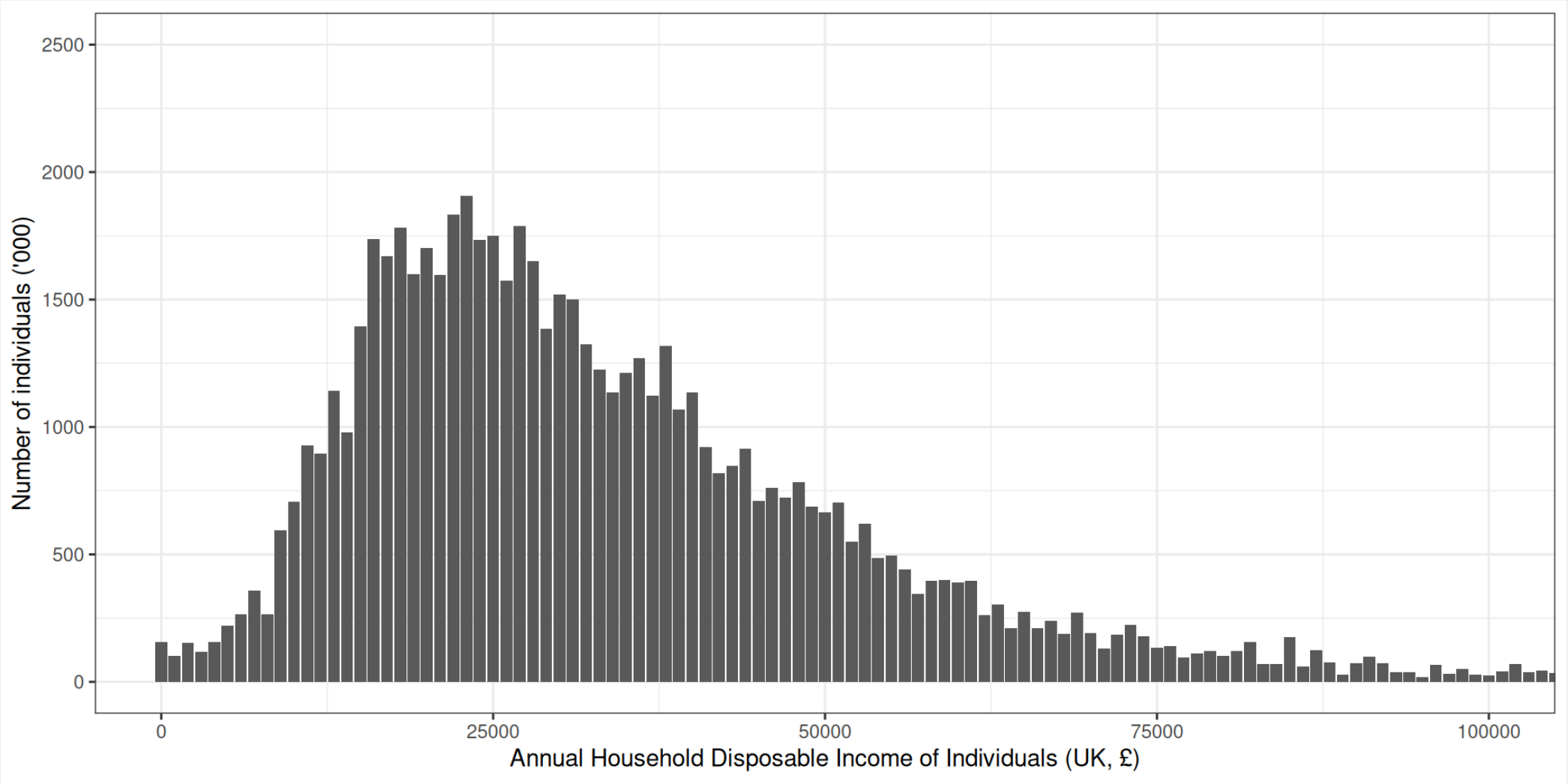
- Both median and mean are appropriate for quantitative data.
- For symmetric distributions mean and median are very similar.
- However, for distributions with long **tails** (**skewed**) median provides a more accurate location of the centre.
- Means are greatly affected by outliers.
- At the same time for discrete data that takes relatively few unique values, similar medians can be found for quite different patterns of the data.

Example: UK Income Data 2020

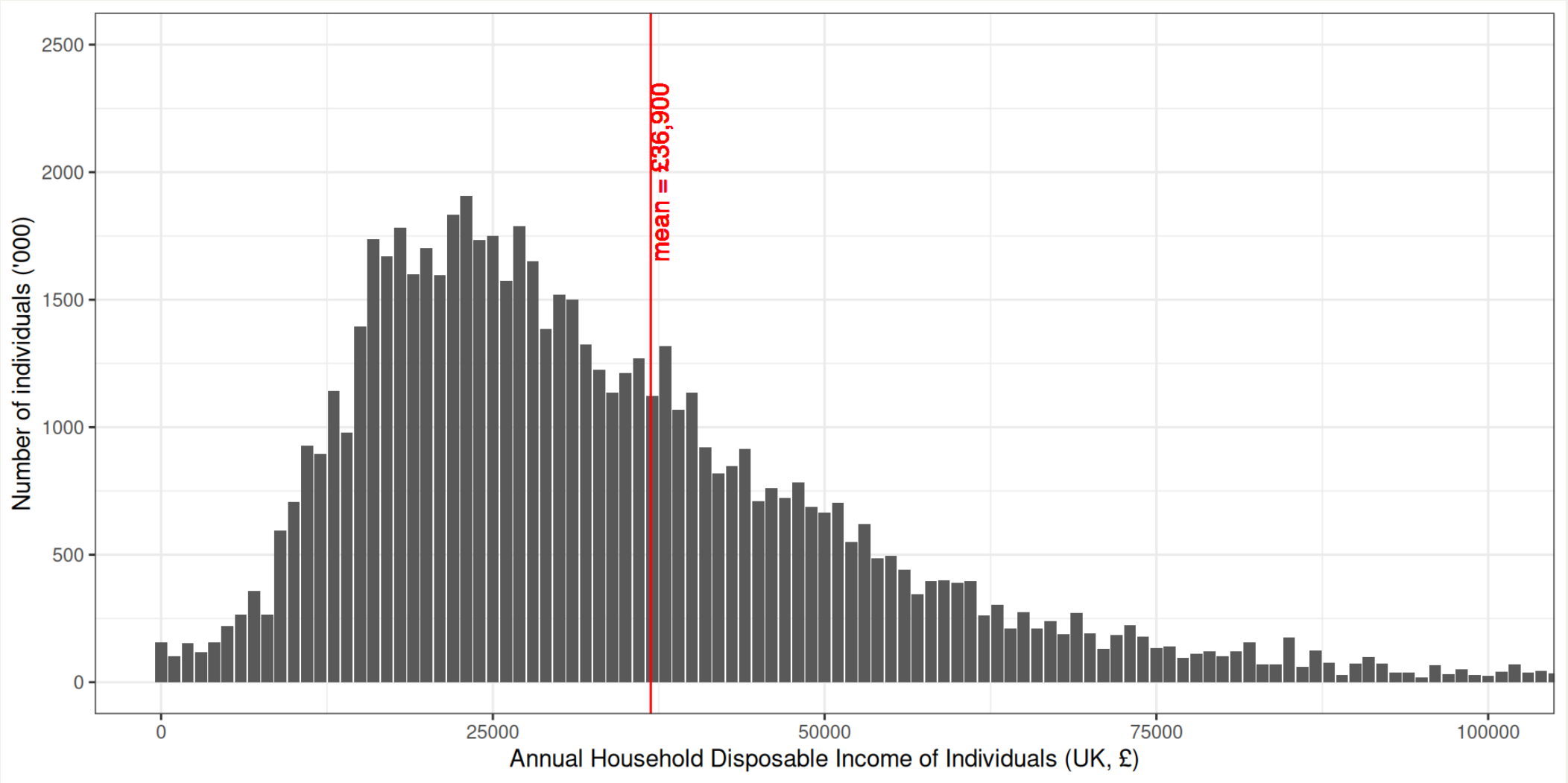


ONS Average household income, UK: financial year 2020

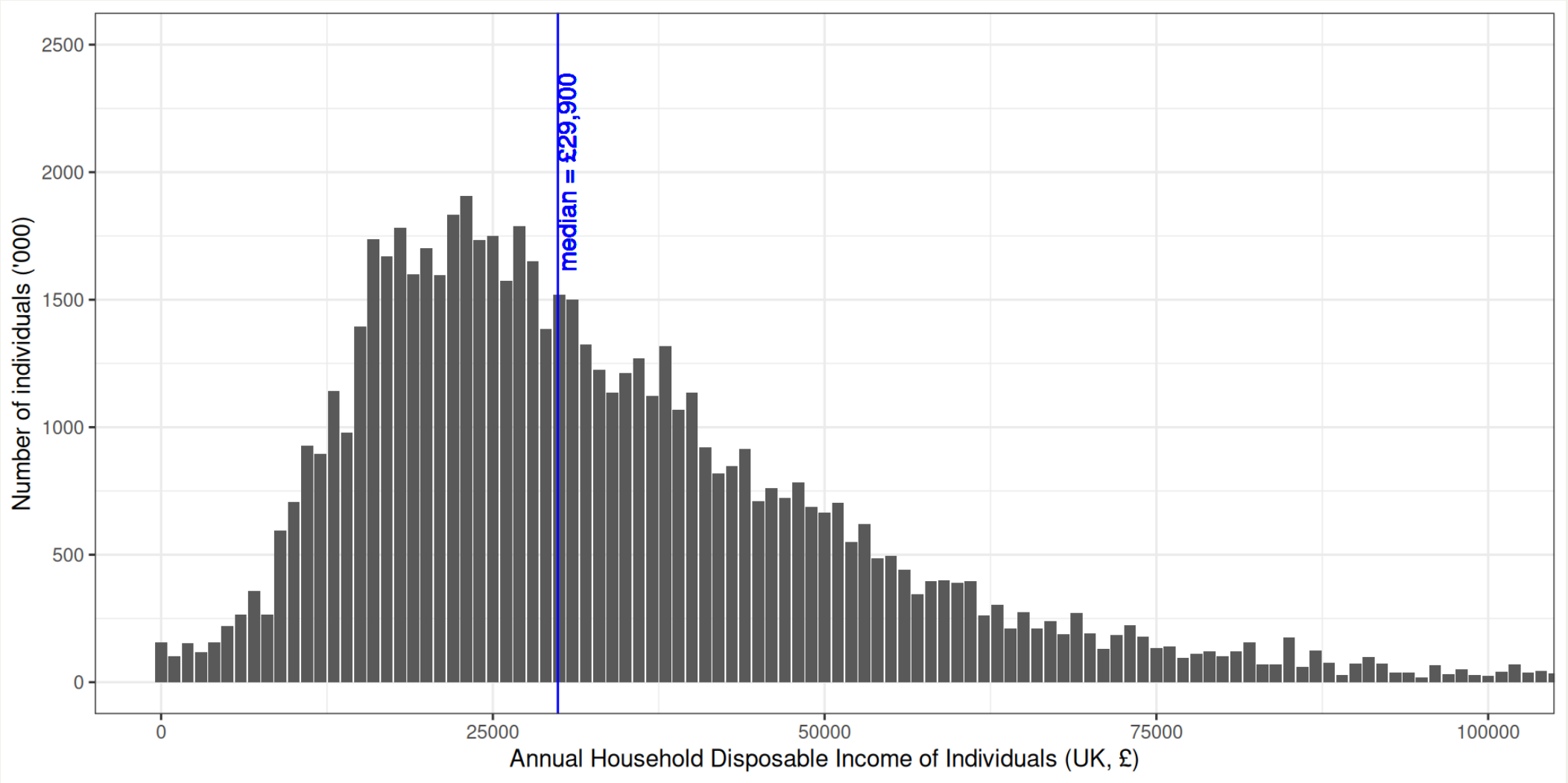
Visualising Quantitative Data: Histogram



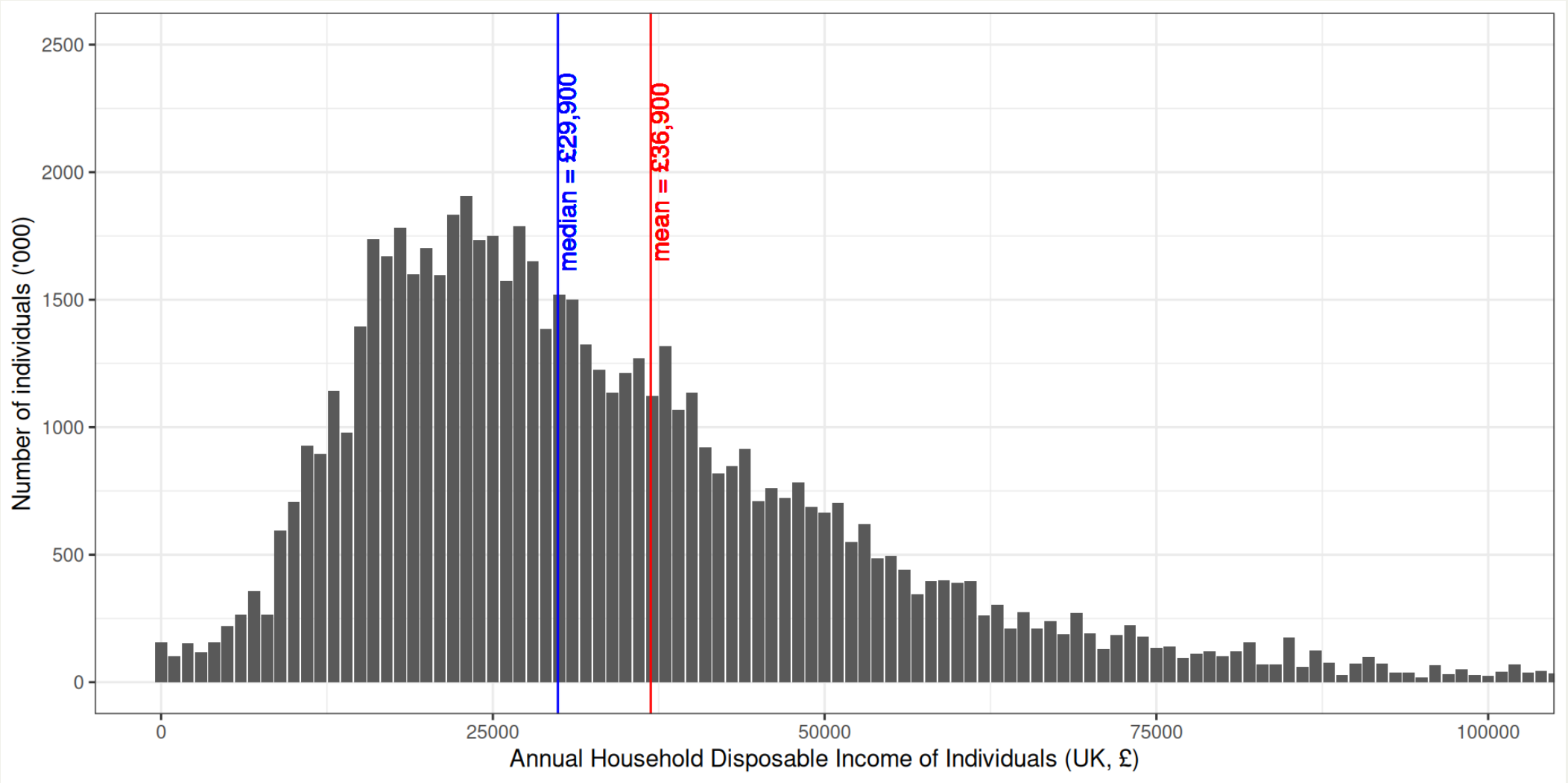
Describing Quantitative Data: Mean



Describing Quantitative Data: Median

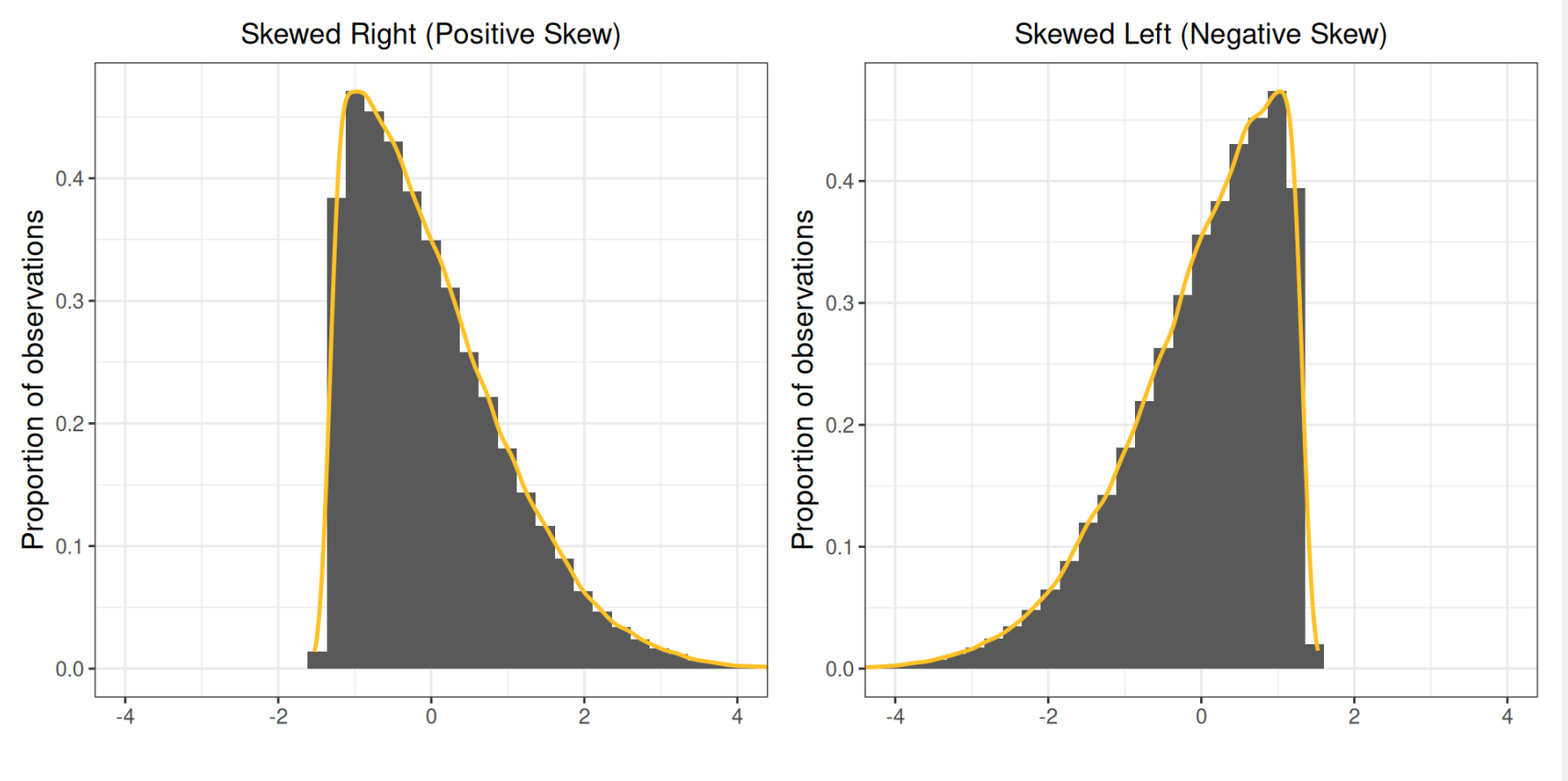


Describing Quantitative Data: Skew



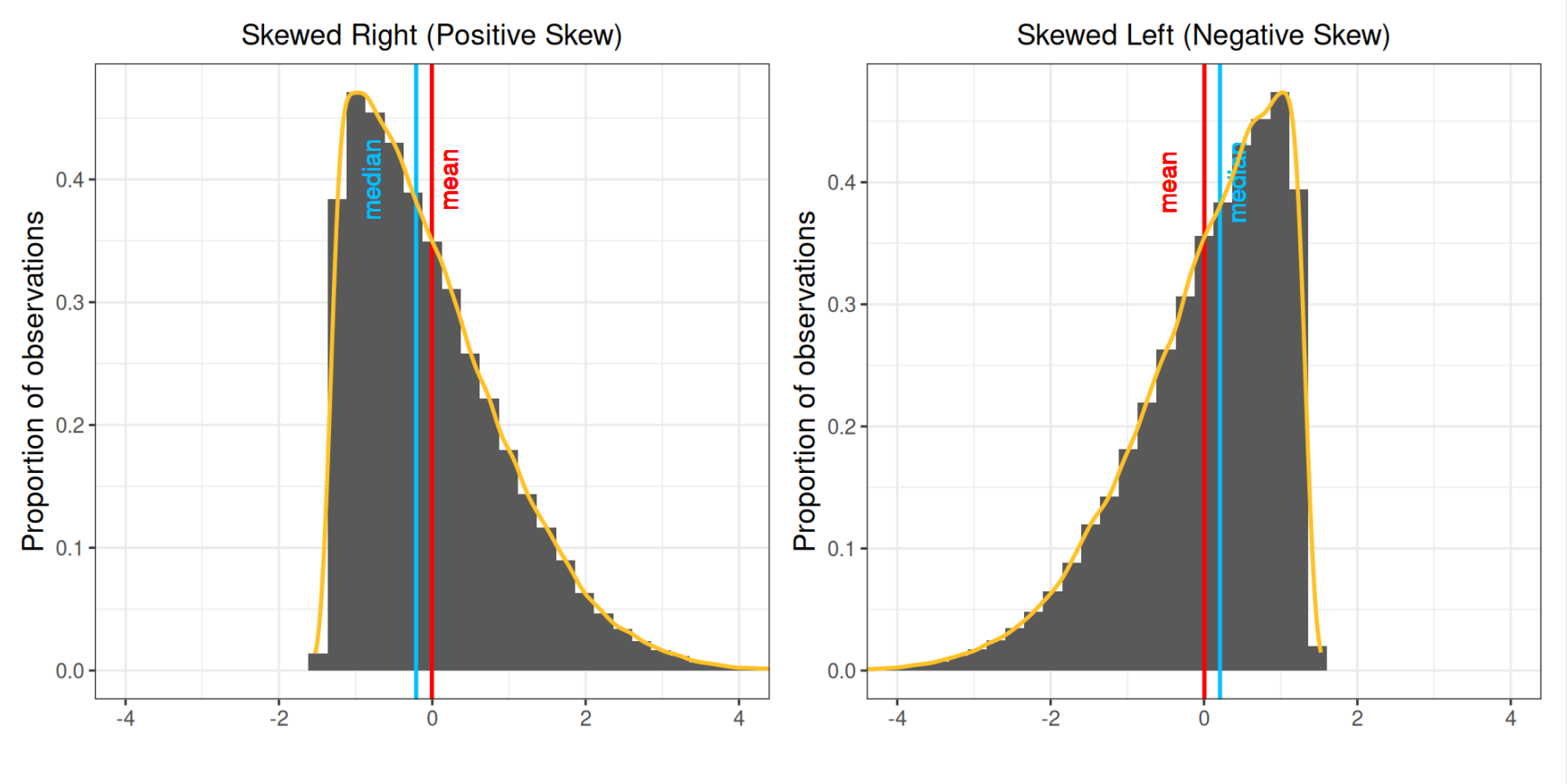
Skewness

- Quantitative variable distributed *asymmetrically* is described as *skewed*.

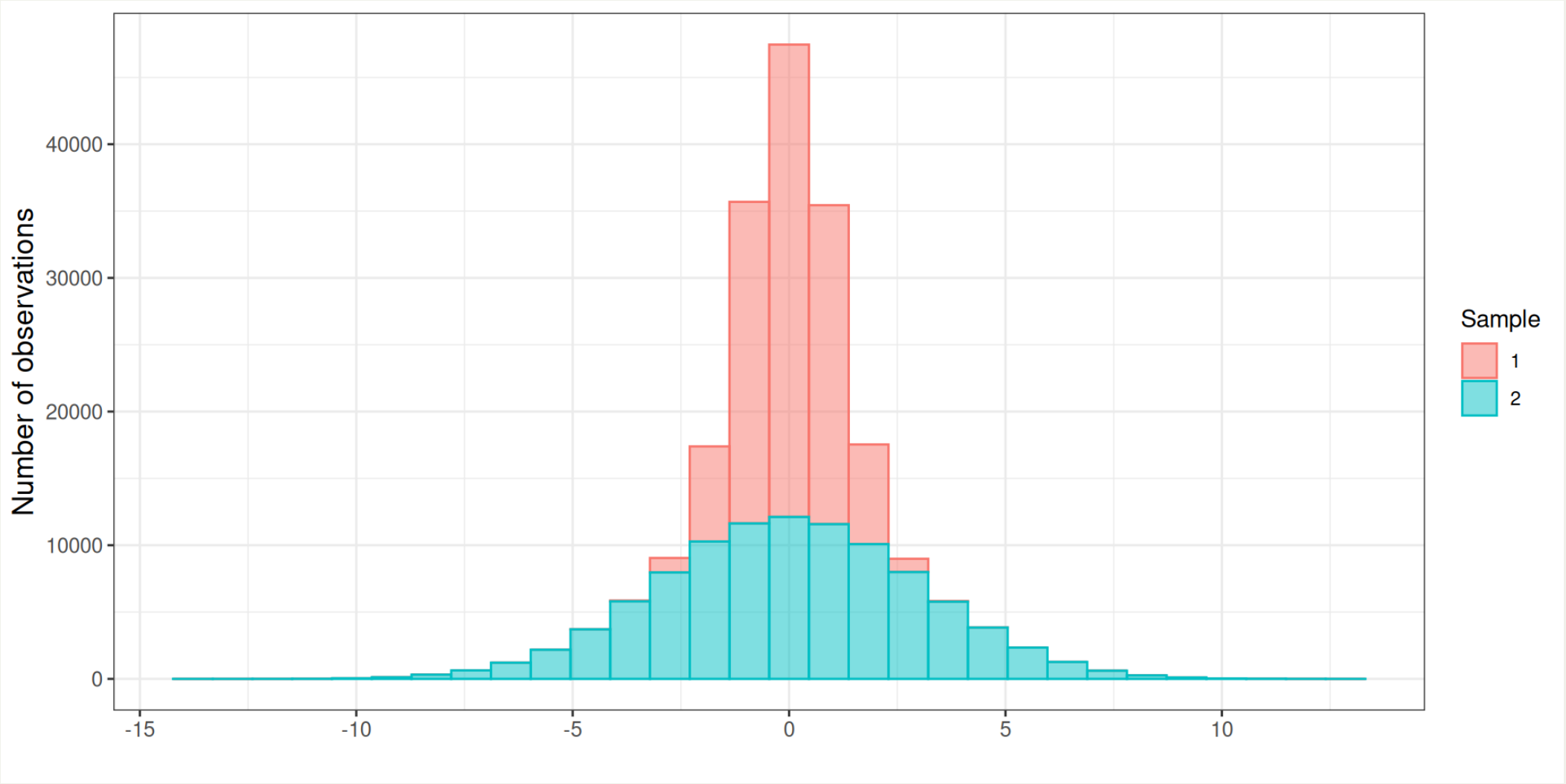


Skewness

- For *skewed* data mean will be towards the direction of skew relative to median.



Dispersion


Number of observations
40000
30000
20000
10000
0
-15
-10
-5
0
5
10
Sample
1
2

Measures of Variability (Dispersion)

- Measures of central tendency do not provide the full picture.
- We need to also describe the spread of the variable
- **Range** - difference between the smallest and largest values.
- **Deviation** - difference between an observed value and the mean of a variable.
- **Variance** - average of squared deviations.
- **Percentiles**, **Quartiles** and **Inter-quartile range** (IQR) - points at which a given percentage of the data falls below that point (median is $50th$ percentile).

Range

- The difference between the smallest and the largest values of a variable.
- Useful for detecting **outliers**, but can be misleading.

```
1 min(democracy_2020$democracy_duration)
```

```
[1] 2
```

```
1 max(democracy_2020$democracy_duration)
```

```
[1] 221
```

```
1 range(democracy_2020$democracy_duration)
```

```
[1]   2 221
```

Deviation

- The difference between an observed value and the mean of a variable:

$$Y_i - \bar{Y}$$

- A sample with little variation will have small deviations, and a sample with a lot of variation will have many large deviations.
- So, we might decide to summarise the variability by summing up all the deviations:

$$\sum_{i=1}^{N} Y_i - \bar{Y}$$

- But the sum of the differences between the mean and each of values is 0 by definition!

Solution?

- We could take the *absolute* values of the deviations and sum them up:

$$\sum_{i=1}^{N} |Y_i - \bar{Y}|$$

- But then the more observations we have, the larger the sum will be (e.g. in two samples of different sizes).
- We can normalize the measure by dividing by the number of observations (sample size N):

$$\frac{\sum_{i=1}^{N} |Y_i - \bar{Y}|}{N}$$

- This measure is called **mean average deviation** (MAD).
- But is rarely used in practice (largely, for technical reasons).

Variance

- Another way of turning negative numbers into positive is to square them (multiply by themselves).
- The sum of squared deviations normalised by the number of observations is called **variance**:

$$Var(Y) = \sigma^2$$

- σ^2 (sigma squared) denotes the *population variance*.
- *Sample variance* (denoted as s^2) is calculated like this:

$$s^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N - 1}$$

- Note that in a sample we have $N - 1$ rather than N in the denominator.

Standard Deviation

- Variance is expressed in the original units of measurement squared.
- To return to the original units we can take a square root of it.
- **Standard deviation** is the square root of variance.
- Conventionally denoted as s or SD.
- The most commonly used measure of deviation.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}}$$

```
1 var(democracy_2020$democracy_duration)
```

```
[1] 1918.217
```

```
1 sqrt(var(democracy_2020$democracy_duration))
```

```
[1] 43.79745
```

```
1 sd(democracy_2020$democracy_duration)
```

```
[1] 43.79745
```

Next

- Workshop:
  - Data Structures
- 1 R Assignment due:
  - **08:59 Tuesday, 4 February**
- Next week:
  - Probability Theory