Descriptive statistics

Quantitative and graphical techniques for summarizing data

Last time

1. Data semantics

- **categorical** data: ordinal (ordered) or nominal (unordered)
- **numeric** data: continuous (no 'gaps') or discrete ('gaps')

2. Data types and data structures in R

- basic types: numeric, character, logical, integer
- a **vector** is a collection of values of one type
- a **data frame** is a type-heterogeneous list of vectors of equal length

Techniques for summarizing data depend on the data type

Vectors can store observations of one variable:

```r
# 4 observations of age
ages <- c(18, 22, 18, 12)
ages
```

```
[1] 18 22 18 12
```

Data frames can store observations of many variables:

```r
# 3 observations of 2 variables
data.frame(subject.id = c(11, 2, 31),
           age = c(24, 31, 17),
           sex = c('m', 'm', 'f'))
```

```
  subject.id age sex
1         11  24   m
2          2  31   m
3         31  17   f
```

Today's agenda

1. Graphical and tabular summaries for numeric and categorical data
    - frequency distributions
    - barplots
    - histograms
2. Quantitative summaries for numeric data
    - percentiles
    - measures of center
    - measures of spread
3. Lab exploring descriptive statistics and robustness

What are descriptive statistics?

Descriptive statistics are quantitative summaries of the observations of one or more variables. They usually serve one of three aims:

1. Identify typical or “central” values
2. Characterize the variability or “spread” of values
3. Characterize relationships between variables

Descriptive statistics are often accompanied by **graphical summaries** that aid in visualizing these same characteristics.

Today's example data: FAMuSS

Observational study of 595 individuals comparing change in arm strength before and after resistance training between genotypes for a region of interest on the ACTN3 gene.

> Pescatello, L. S., *et al.* (2013). Highlights from the functional single nucleotide polymorphisms associated with human muscle size and strength or FAMuSS study. BioMed research international, 2013.

Example data rows

ndrm.ch	drm.ch	sex	age	race	height	weight	actn3.r577x	bmi
40	40	Female	27	Caucasian	65	199	CC	33.11
25	0	Male	36	Caucasian	71.7	189	CT	25.84
40	0	Female	24	Caucasian	65	134	CT	22.3
125	0	Female	40	Caucasian	68	171	CT	26

Categorical frequency distributions

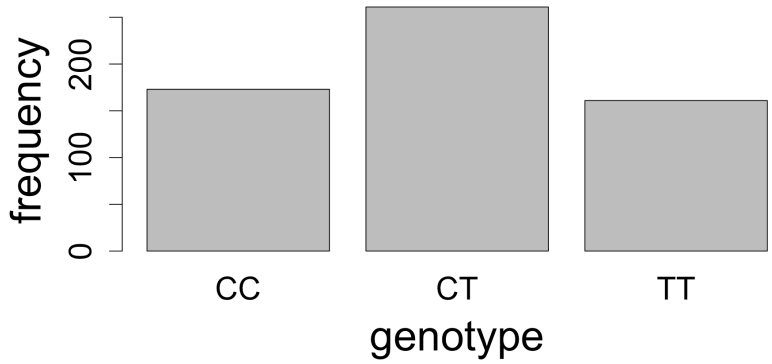
For categorical variables, the frequency distribution is simply an observation count by category. For example:

Data table

participant.id	genotype
494	TT
510	TT
216	CT
19	TT
278	CT
86	TT

Frequency distribution

CC	CT	TT
173	261	161



Numeric frequency distributions

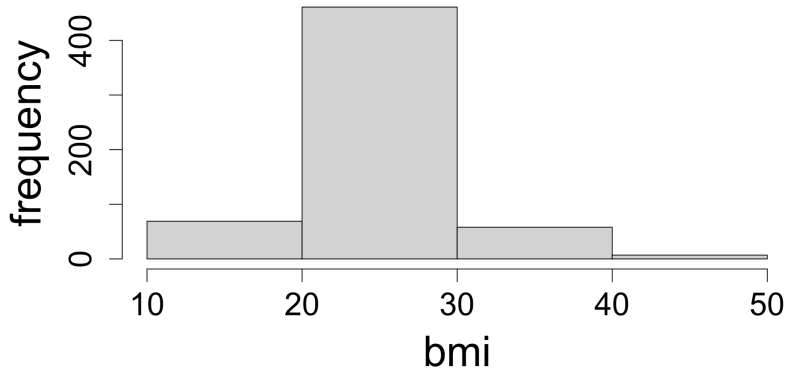
Frequency distributions of numeric variables are observation counts by *range*.

Data table

participant.id	bmi
194	22.3
141	20.76
313	23.48
522	29.29
504	42.28
273	20.34

Frequency distribution

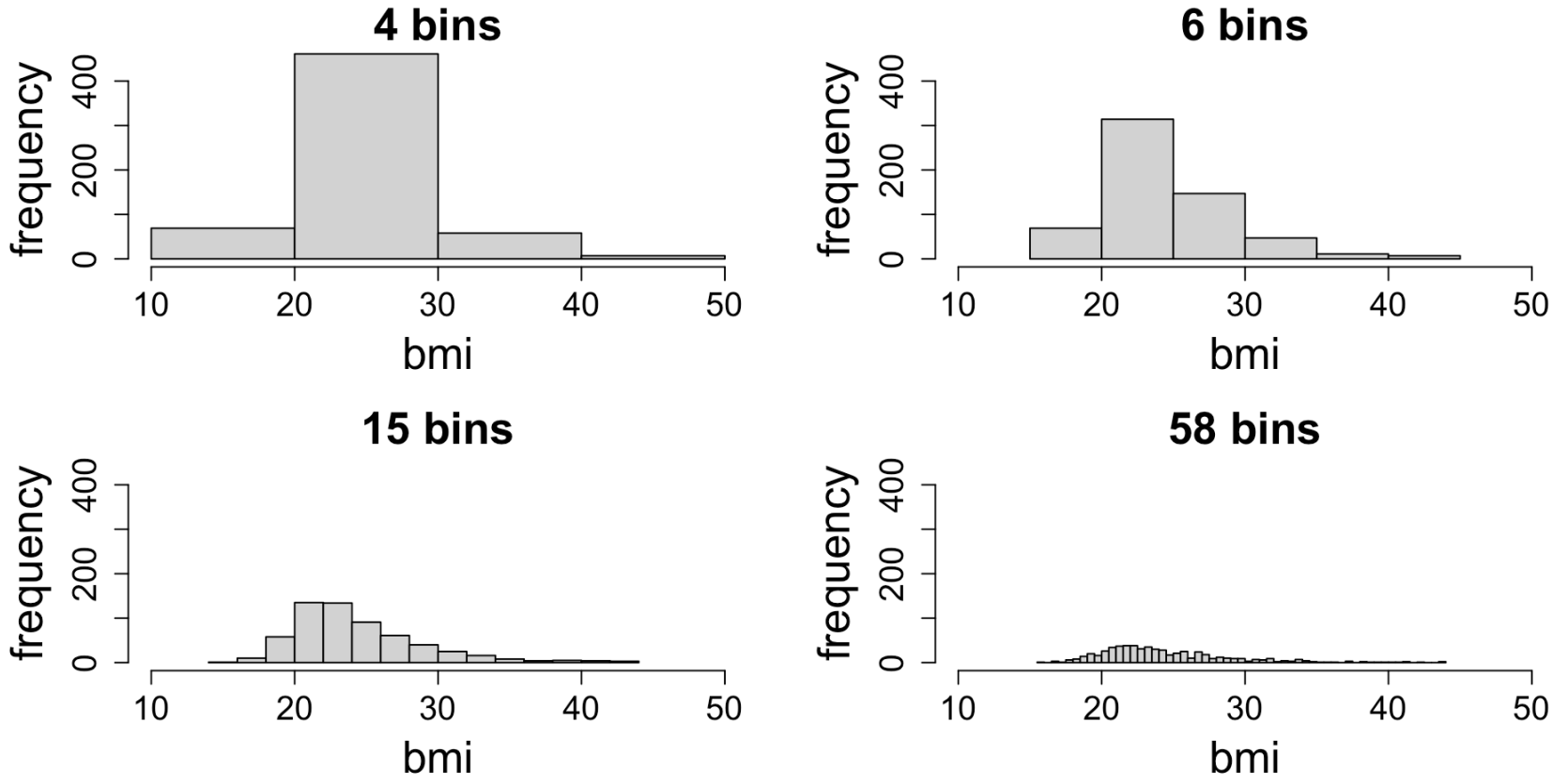
(10,20]	(20,30]	(30,40]	(40,50]
69	461	58	7



The operation of dividing a numeric variable into interval ranges is called **binning**.

Histograms

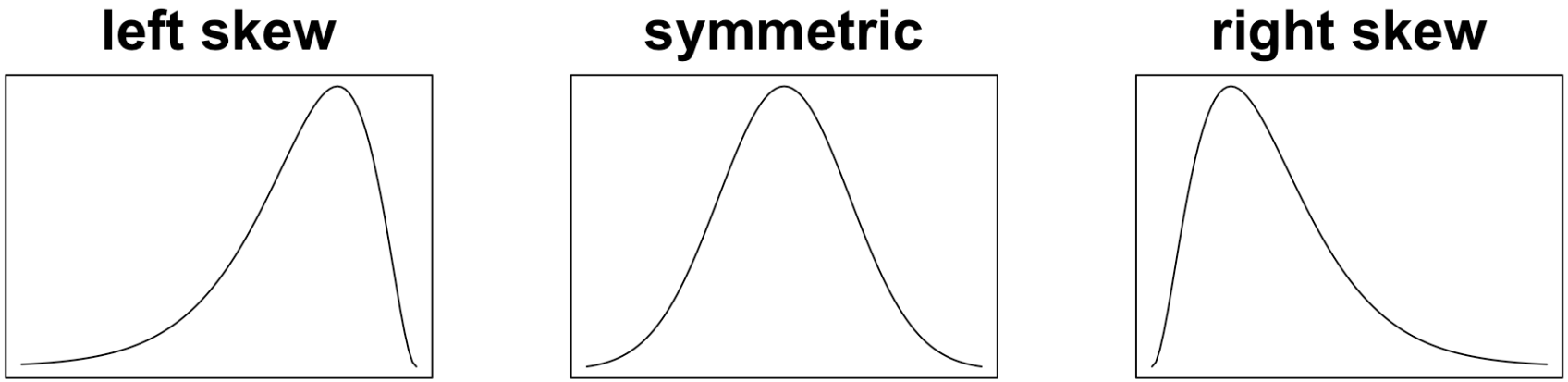
The graphical display of a frequency distribution for a numeric variable is called a **histogram**. Binning has a big effect on the visual impression.



Shapes

For numeric variables, the histogram reveals the **shape** of the distribution:

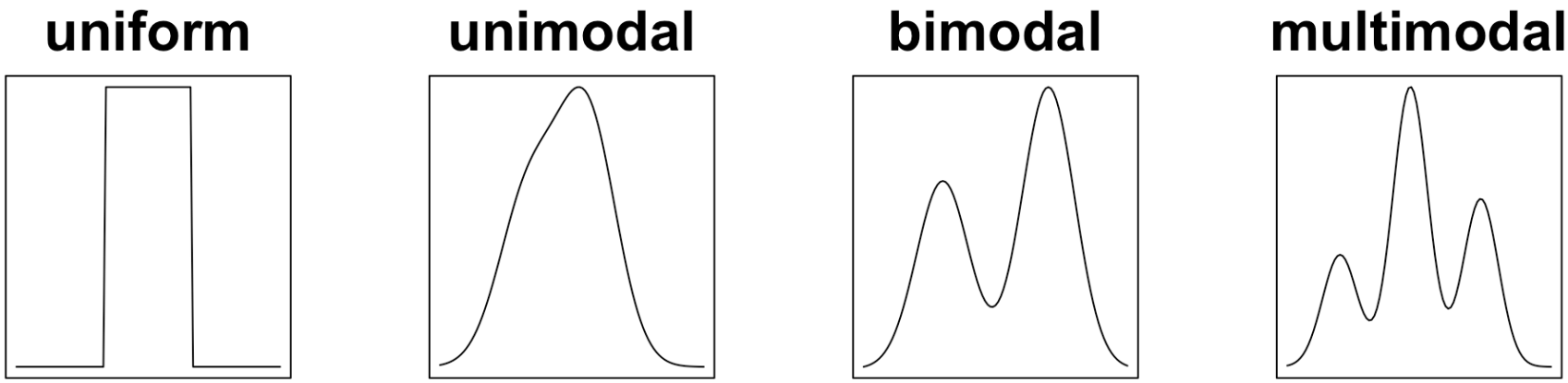
- **symmetric** if it shows left-right symmetry about a central value
- **skewed** if it stretches farther in one direction from a central value



Modes

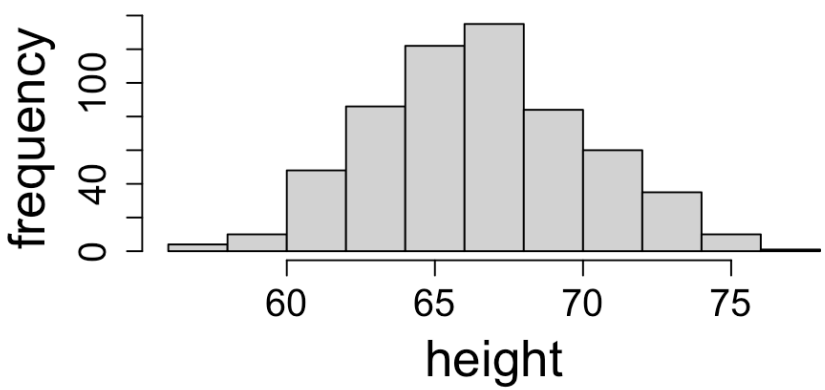
Histograms also reveal the number of **modes** or local peaks of frequency distributions.

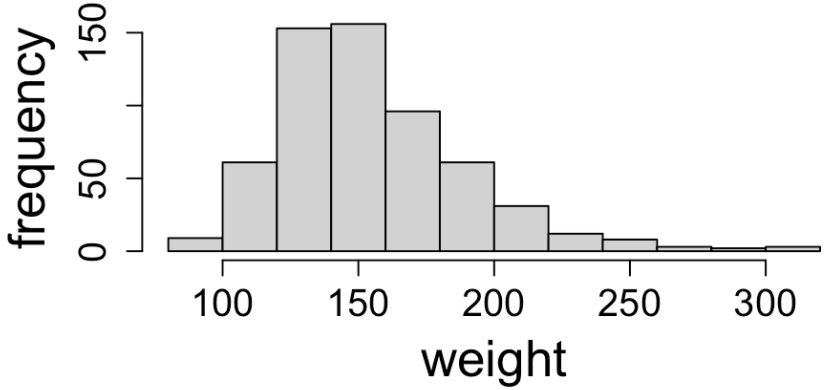
- **uniform** if there are zero peaks
- **unimodal** if there is one peak
- **bimodal** if there are two peaks
- **multimodal** if there are two or more peaks

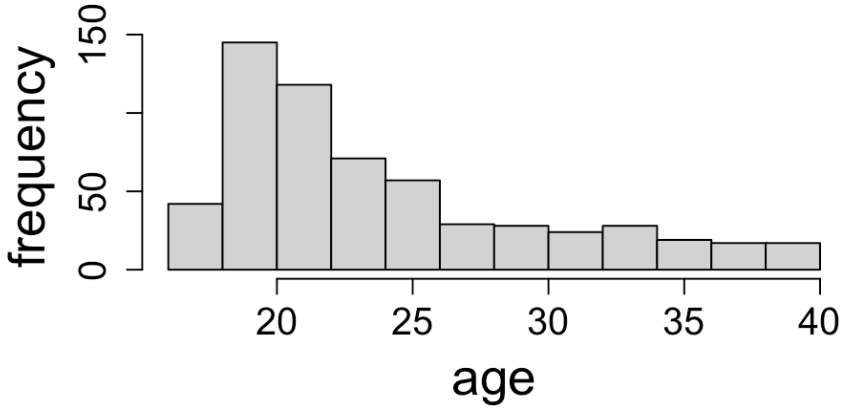


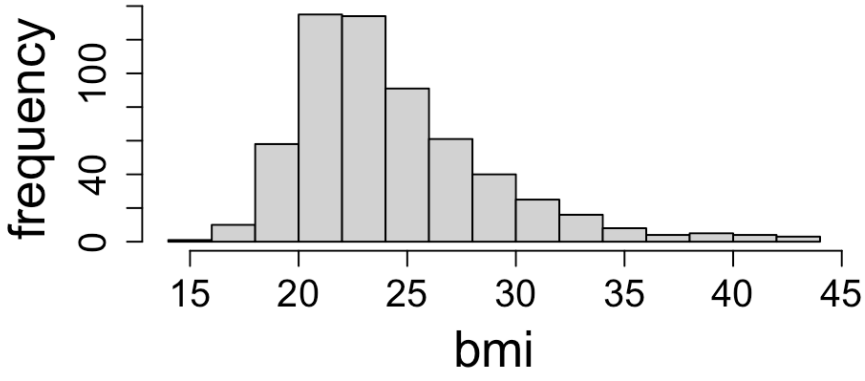
Your turn: characterizing distributions

Consider four variables from the FAMuSS study. Describe the shape and modality.



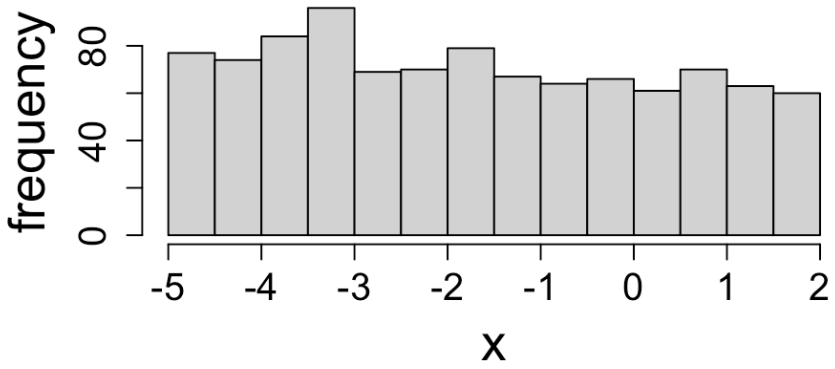


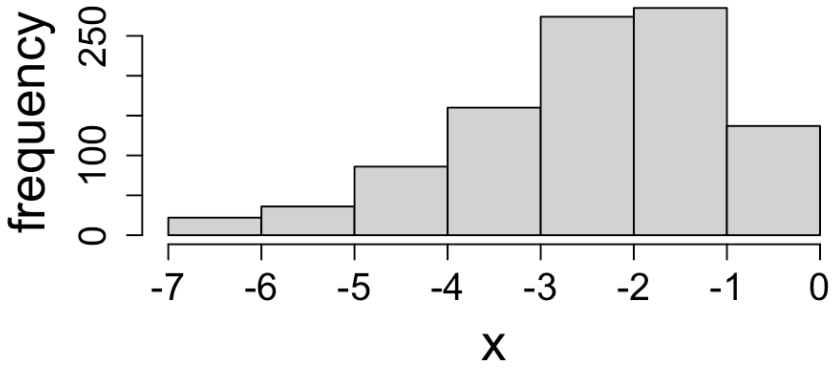


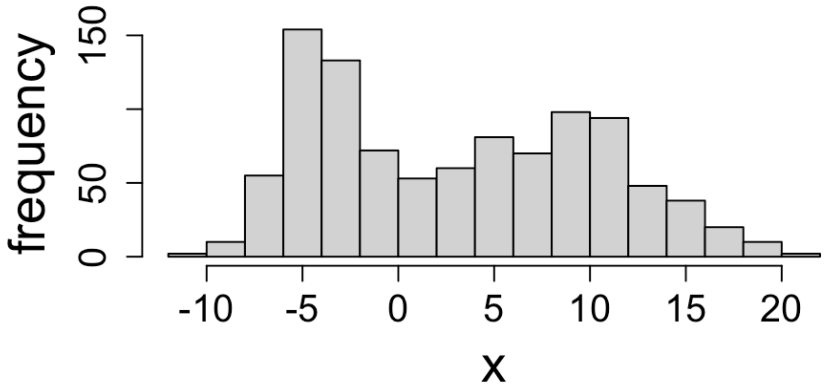


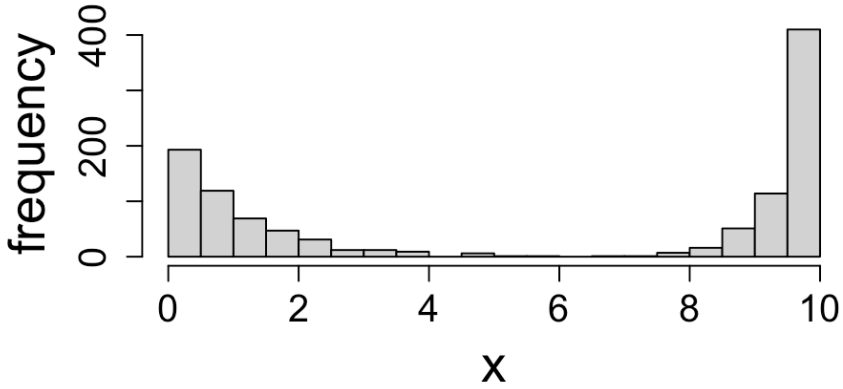
Your turn: characterizing distributions

Here are some made-up data. Describe the shape and modality.









Descriptive measures

Frequency distributions are great for many purposes but they have limitations:

- minimal “data reduction”, especially for many bins/categories
- sensitive to choice of binning
- perception of pattern is subjective

Descriptive measures, by contrast, reduce all observations of a variable down to just one number. There are two common types of measures:

- **measures of center**: mean, median, mode
- **measures of spread**: absolute deviation, standard deviation, interquartile range, range

Measures of center

A measure of center is a statistic that reflects the typical value of one or more variables.

There are three common measures of center, each of which corresponds to a slightly different meaning of “typical”:

Measure	Definition
Mode	Most frequent value
Mean	Average value
Median	Middle value

Suppose your data consisted of the following observations of age in years:

19, *19*, *21*, *25* and *31*

- the **mode** or most frequent value is 19
- the **median** or middle value is 21
- the **mean** or average value is $\frac{19+19+21+25+31}{5} = 23$

These measures are only used with numeric variables.

Your turn

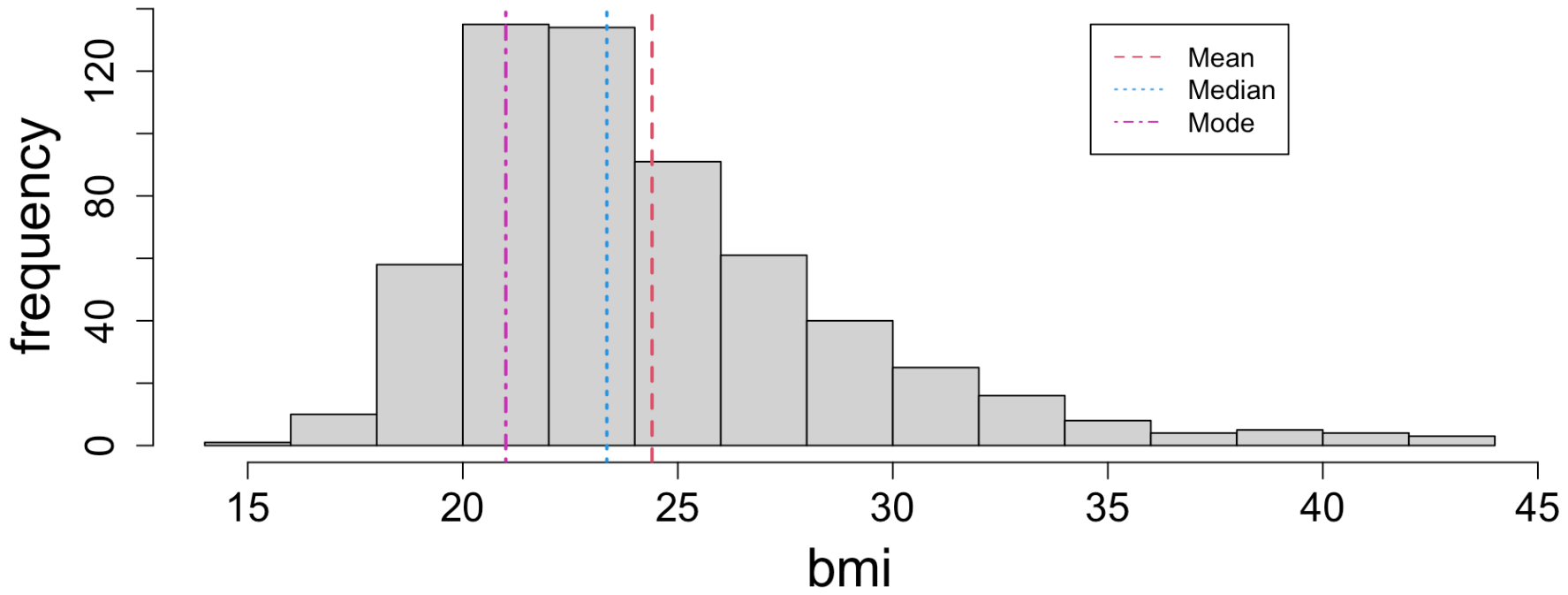
Consider the first 8 observations of change in nondominant arm strength from the FAMuSS study data:

40, *25*, *40*, *125*, *40*, *75*, *100* and *57.1*

Compute the mean, median, and mode.

Comparing measures of center

Each statistic is a little different, but often they roughly agree; for example, all are between 20 and 25, which seems to capture the typical age well enough.



How do you think the frequency distribution affects which one is "best"?

Percentiles

The median is an example of a **percentile**: a value with specified proportions of data lying both above and below. For example, the 20th percentile is the value with 20% of observations below and 80% of observations above.

Ranking observations helps to find this number. Suppose we have 5 observations:

age	19	20	21	25	31
rank	1	2	3	4	5

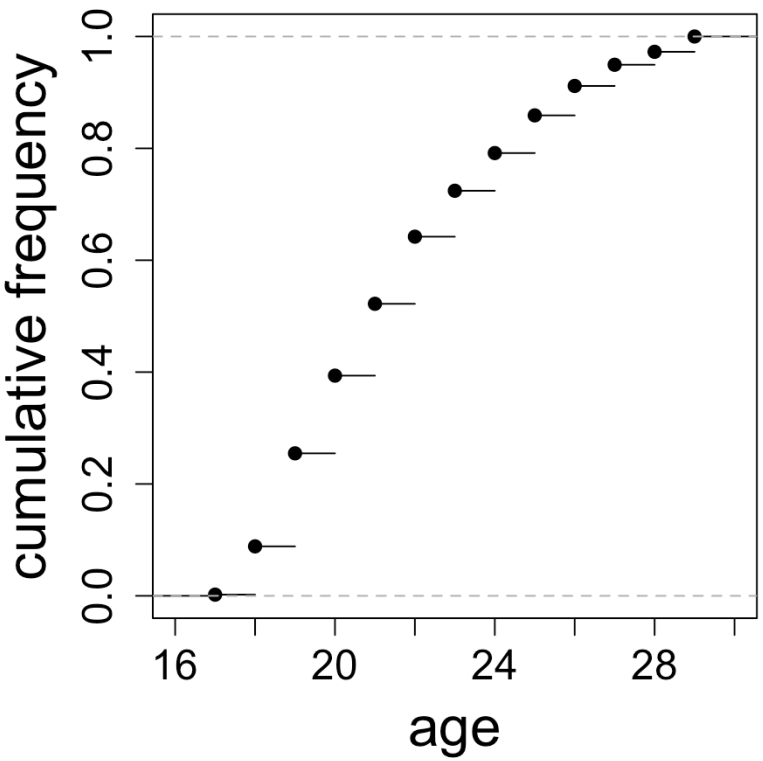
The 20th percentile is 20 since it is ranked second when observations are listed in order:

- 20% below (19, 20)
- 80% above (20, 21, 25, 31)

> Software implementations have a variety of ways for calculating percentiles when an exact solution isn't available due to ties (repeated values) or sample size.

Cumulative frequency distribution

The *cumulative frequency distribution* is a function showing all percentiles with an exact solution. Think of it as percentile (y) against value (x).



Interpretation of some specific values:

- about 40% of the subjects are 20 or younger
- about 80% of the subjects are 24 or younger

Your turn:

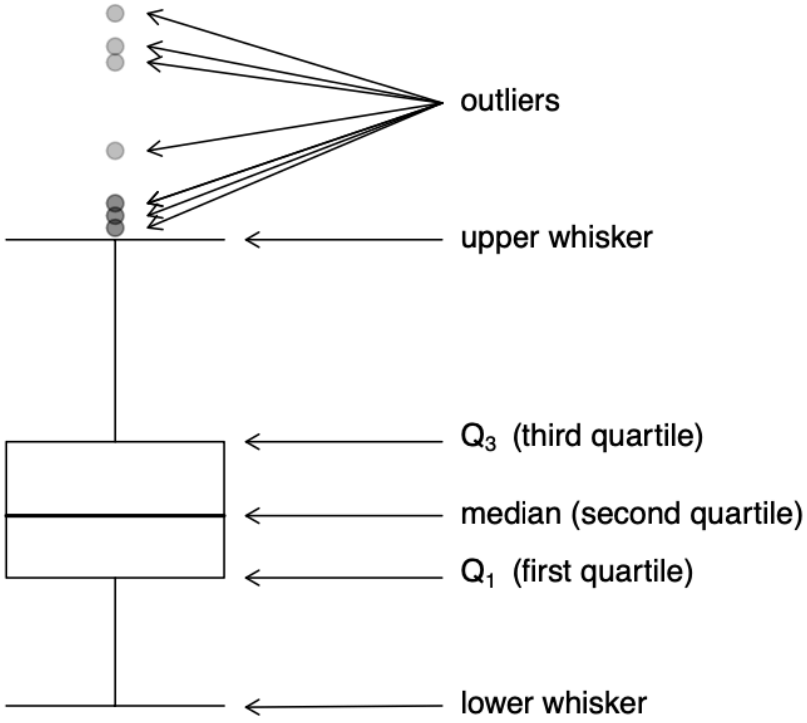
1. Roughly what percentage of subjects are 22 or younger?
2. About what age is the 10th percentile?

Common percentiles

The **five-number summary** is a collection of five percentiles that succinctly describe the frequency distribution:

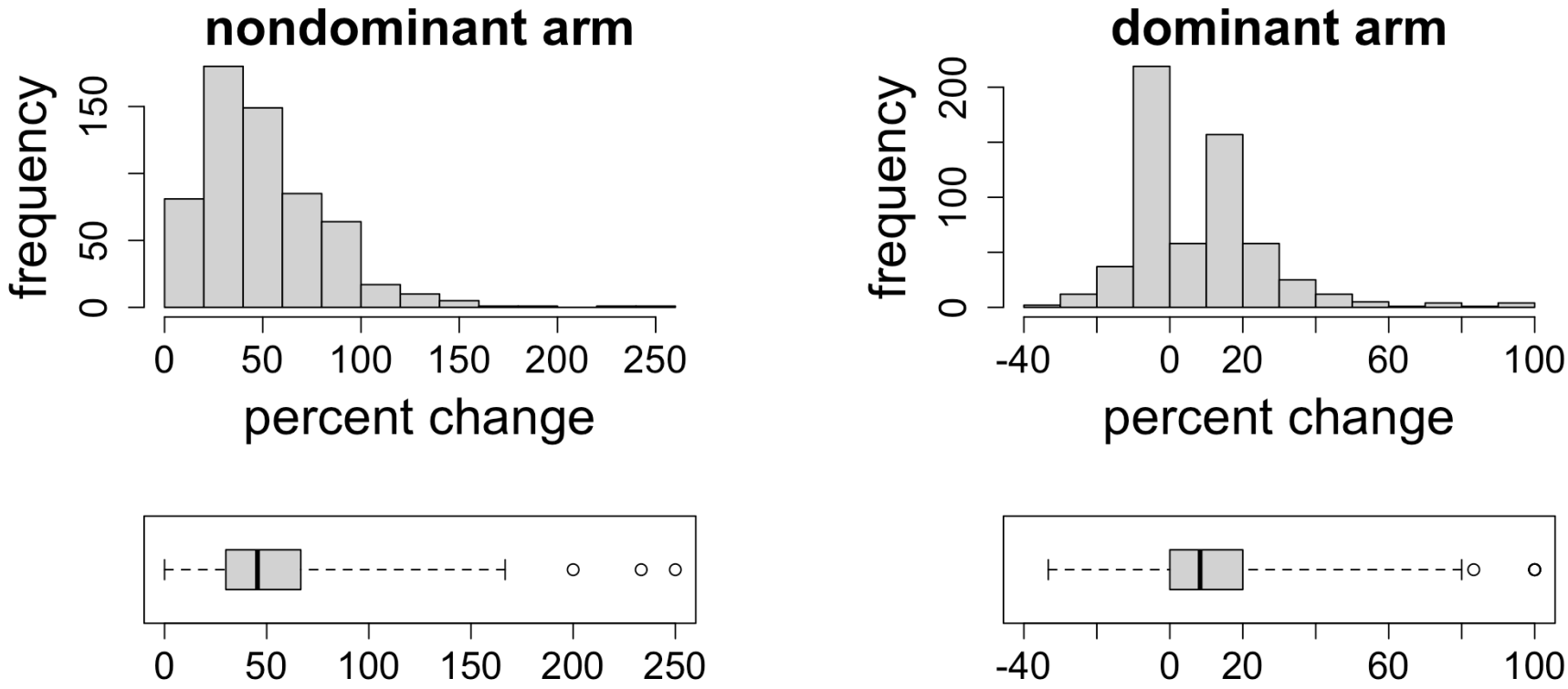
Statistic name	Meaning
minimum	0th percentile
first quartile	25th percentile
median	50th percentile
third quartile	75th percentile
maximum	100th percentile

Boxplots provide a graphical display of the five-number summary.



Boxplots vs. histograms

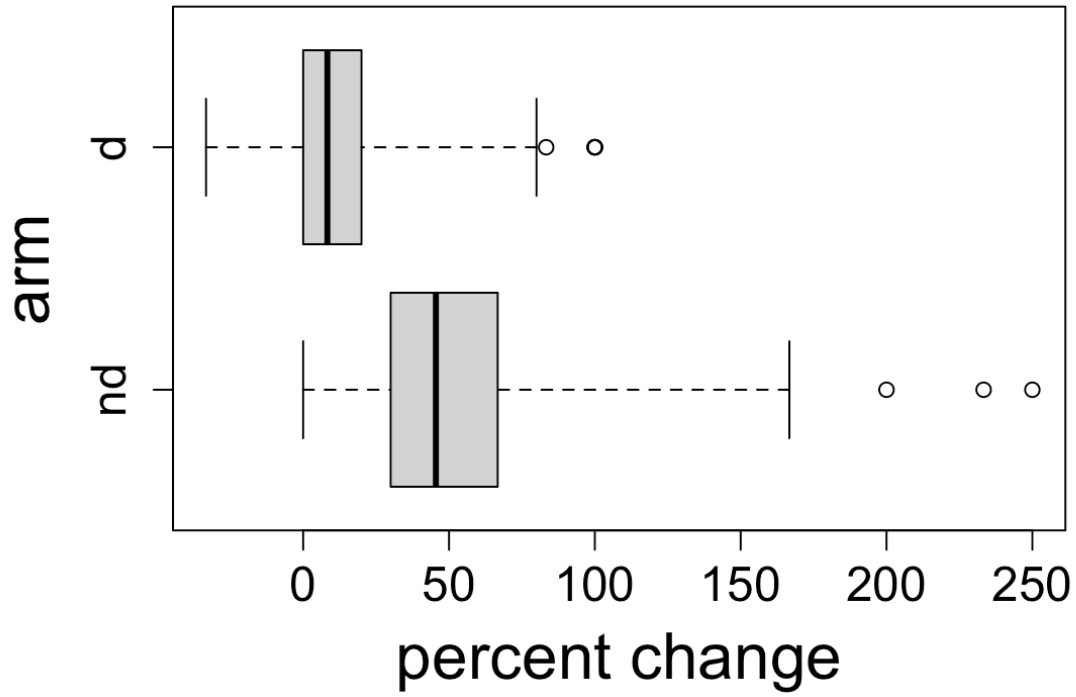
Notice how the two displays align, and also how they differ. The histogram shows shape in greater detail, but the boxplot is much more compact.



Boxplots vs. histograms

Suppose we wanted to compare the change in dominant arm strength with the change in nondominant arm strength in the FAMuSS (sports gene) study.

The boxplot is a cleaner display due to its compactness.



Lab: robustness

For this lab we'll continue to work with the FAMuSS data as we have throughout lecture.

This lab has two objectives:

1. Teach you to compute descriptive statistics and prepare graphical summaries for a single variable in R
2. Learn when and why to use certain descriptive statistics in place of others

Up next: multivariate summaries

Today we discussed numeric and graphical summaries for a *single* variable. These are **univariate** techniques.

- **Will** reveal basic statistical properties (shape, skew, outliers)
- **Won't** reveal relationships

What if you wish to understand relationships? The fact is, most data are **multivariate** because several variables are measured together.

Next time we'll discuss

1. Measures of spread/variability
2. Bivariate descriptive and graphical techniques