[IGP 484] Quantitative Biology: Data Analysis for Life Scientists

Dr. Katie Evans

🕒 Tu/Th 1:30-3:00, Hughes Auditorium

💻 github.com/katiesevans/IGP_biostatistics

biostats-484

🕒 Tu/Th 1:30-3:00, Hughes Auditorium*

💻 github.com/katiesevans/IGP_biostatistics

- Weekly homework assignments (40%)
- Class participation (5%)
- Midterm exam (15%)
- Final exam (15%)
- Final project (25%)

9/23 and 11/11 — Class will meet in Daniel Hale Williams, McGaw 2-320

🕒 Tu/Th 1:30-3:00, Hughes Auditorium

💻 github.com/katiesevans/IGP_biostatistics

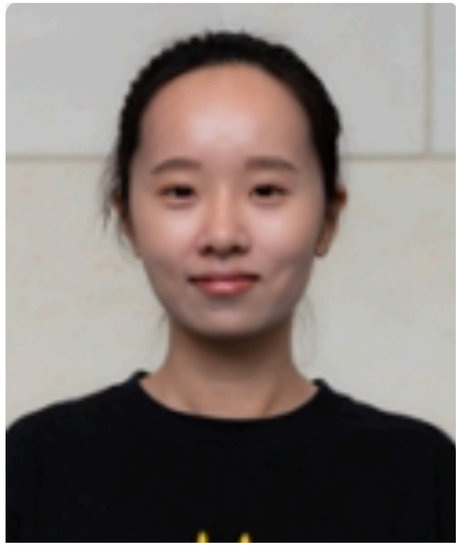
- **Homework assignments**
  - Must be completed using R
  - Make sure to show ALL your work for credit
  - Submitted electronically as PDF or markdown
- **Final project**
  - Choose a dataset (rotation, internet, old publication…)
  - Perform a statistical analysis
  - Write up a report with conclusions
  - *More to come in second half of class, but start thinking about ideas!*

Meet the class TA's!

Saya

Jiexi

Sam



Yidan

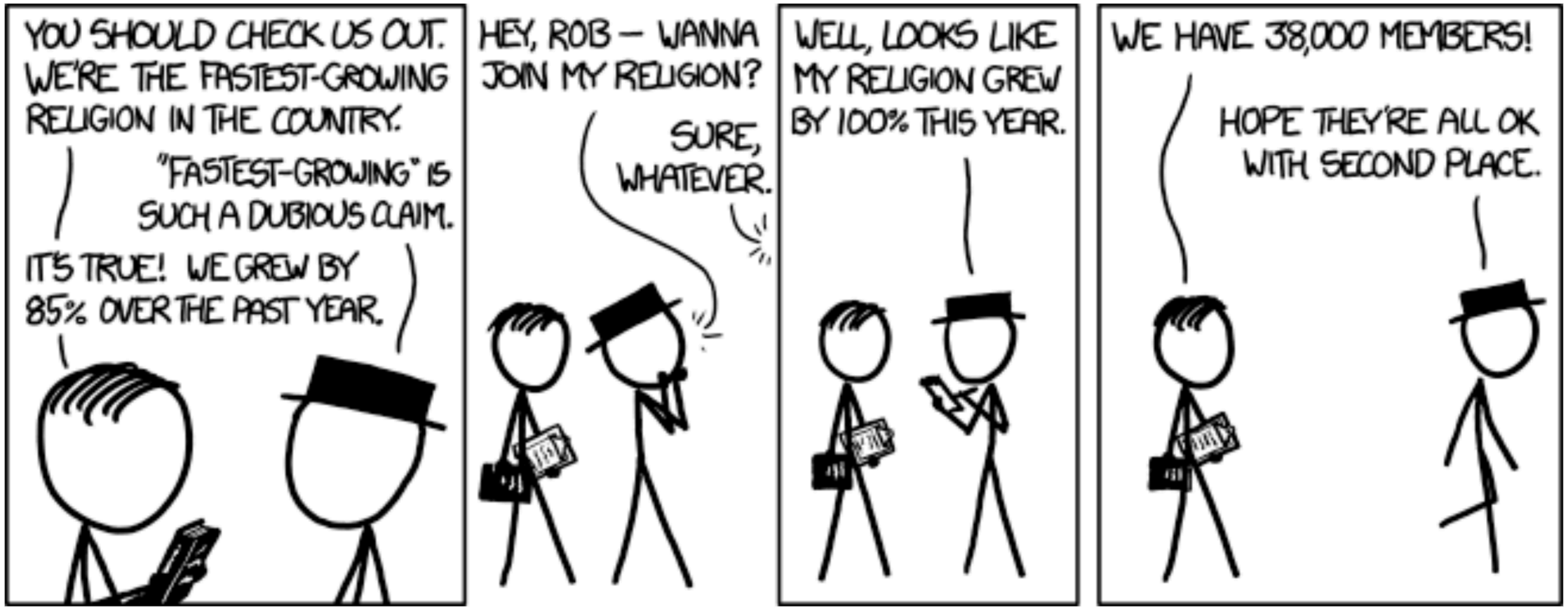
🕒 **TA office hours:** Tuesday 3-4 (TBD); Friday 1:30-2:30 (Zoom)

(Or by appointment)

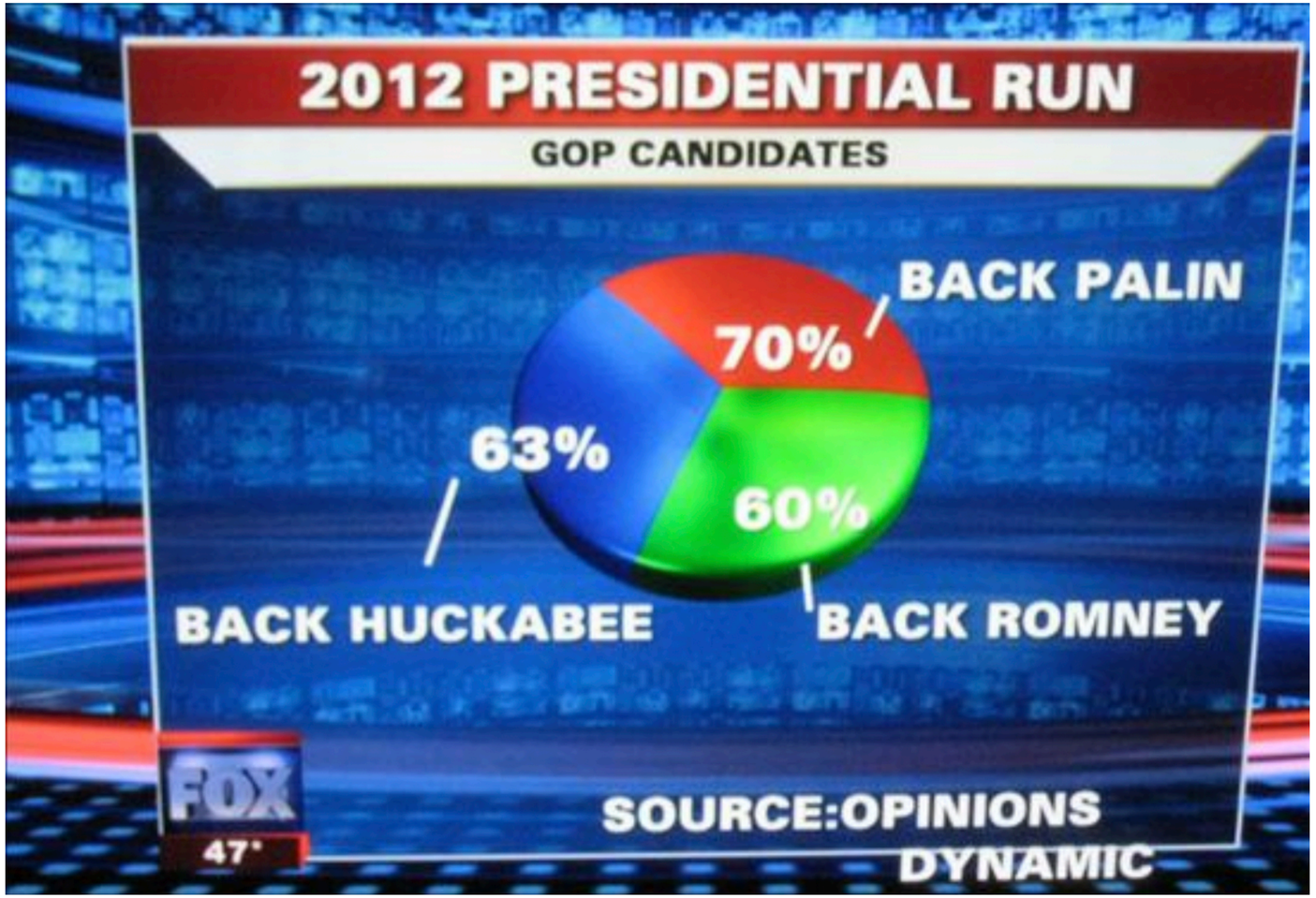
Date	Class	Lecture	Topics	Textbook Sections
September 21, 2021	Lecture	Lecture 1: Introduction	Class intro, what is "statistics"?, displaying and summarizing data, getting started with R	1.1-1.3, 2.1-2.7
September 23, 2021	Lecture	Lecture 2: Probability	notation, independence, conditional probability	3.1-3.3
September 28, 2021	Lecture	Lecture 3: Probability (part 2)	conditional probability, Bayes theorem, practical implications	3.3-3.5
September 30, 2021	Lecture	Lecture 4: Distributions	binomial and normal	3.6, 4.1-4.3
October 5, 2021	Practicum	Practicum 1: Learning R	Data tidying with tidyverse, plotting with ggplot2, thinking like a data scientist, Rmarkdown	NA
October 7, 2021	Lecture	Lecture 5: Distributions (part 2)	Central Limit Theorem, sampling	5.1-5.4
October 12, 2021	Lecture	Lecture 6: Estimation, testing, and p-values	estimation, testing, p-values	4.4, 6.1-6.5
October 14, 2021	Lecture	Lecture 7: One-sample t-tests	estimating and testing means	6.6-6.7, 7.2-7.6, 7.8
October 19, 2021	Lecture	Lecture 8: Two-sample comparisons	unpaired t-tests, paired t-tests, power	7.2-7.6, 8.2-8.3, 7.7
October 21, 2021	Lecture	Lecture 9: Nonparametric alternatives	permutation, Wilcoxin/Mann-Whitney, signed rank	7.1, 7.10, 8.4-8.5
October 26, 2021	Review	Q&A: Project discussion and Midterm review	Bring your questions!	Ch. 1-8
October 28, 2021	Exam	Midterm exam	NA	NA

November 2, 2021	Lecture	**Lecture 10: Multiple hypotheses**	Urn models, "enrichment", Fischer's exact test	10.4
November 4, 2021		**Practicum 2: Analyzing gene expression data with R**	NA	NA
November 9, 2021	Lecture	**Lecture 11: Categorical data**	contingency tables, chi-squared tests, relative risks and odds ratio	9.1-9.2, 9.4, 10.1-10.3, 10.5-10.6, 10.9-10.10
November 11, 2021	Lecture	**Lecture 12: Relationships in data**	independence, covariance, and correlation	12.1-12.2
November 16, 2021	Lecture	**Lecture 13: Regression models**	linear assumptions, interpretation, limitations, transformations	12.4-12.6
November 18, 2021	Lecture	**Lecture 14: Models with categorical covariates**	indicator/"dummy" variables	11.1-11.5
November 23, 2021	Lecture	**Lecture 15: ANOVA**	one-way ANOVA, two-way ANOVA	11.7-11.8
November 25, 2021		**No class: Thanksgiving!**	NA	NA
November 30, 2021	Practicum	**Practicum 3: Model fitting in R**	NA	NA
December 2, 2021	Review	**Q&A: "Overflow" topics and final review**	Cumulative!	NA
December 7, 2021	Exam	**Final Exam**	Cumulative!	NA

Misleading statistics



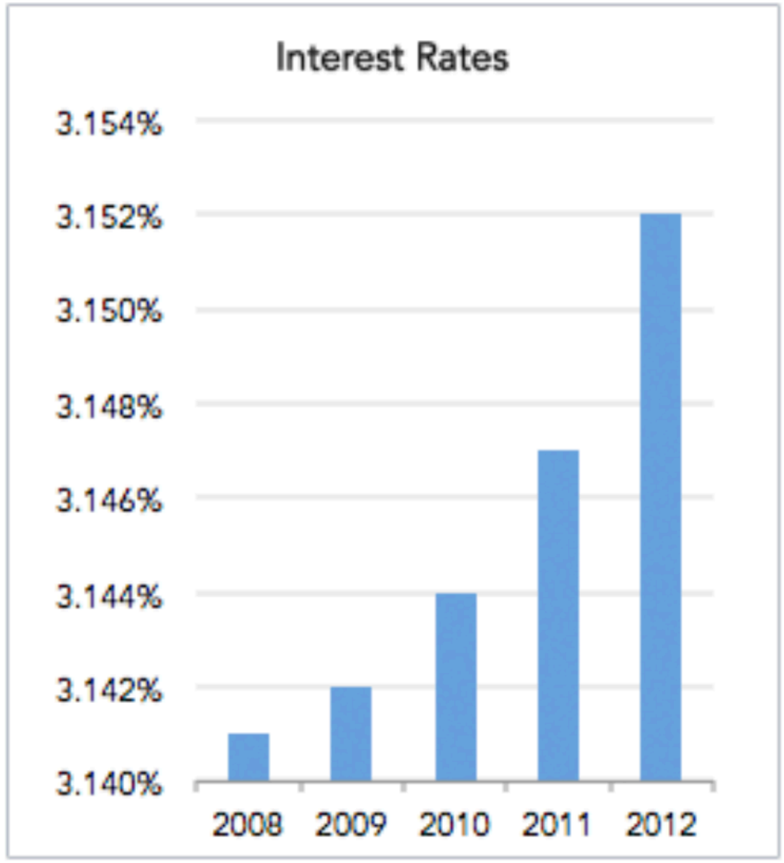
Misleading statistics



https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/

Misleading statistics

Same Data, Different Y-Axis



Misleading statistics

Reproducible research

The New York Times

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

FORTUNE

Big data's dirty problem

by Verne Kopytoff @vkopytoff JUNE 30, 2014, 10:58 AM EDT

PHYS.ORG

Science is in a reproducibility crisis: How do we resolve it?

Sep 20, 2013 by Fiona Fidler & Ascelin Gordon, The Conversation

Validation BY THE SCIENCE EXCHANGE NETWORK About Projects Press Contact

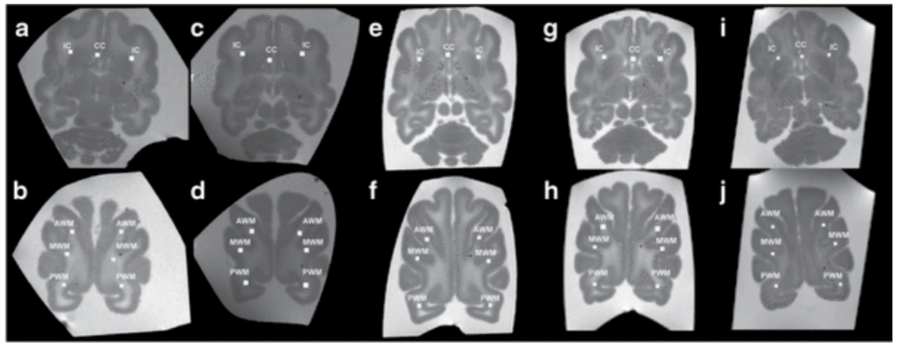
Validating key experimental results via independent replication

Learn more »

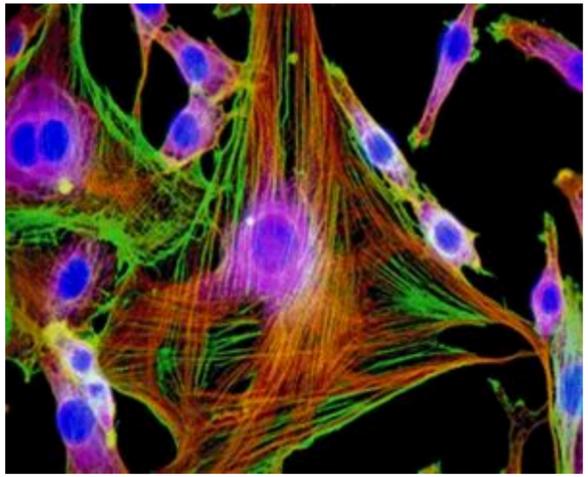
Reproducibility initiatives will invite increased scrutiny into data cleaning methods.

Cr: Erik Andersen

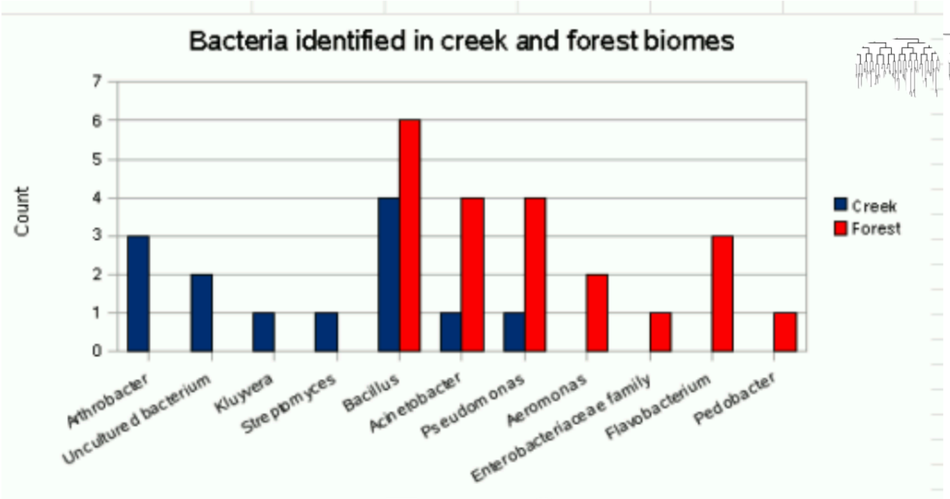
Reproducible research



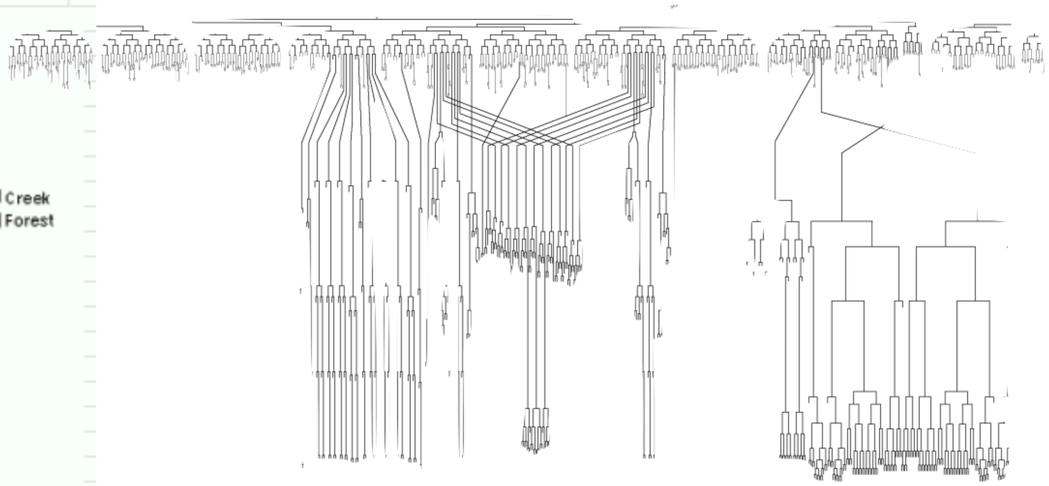
Reproducibility?



Sampling?



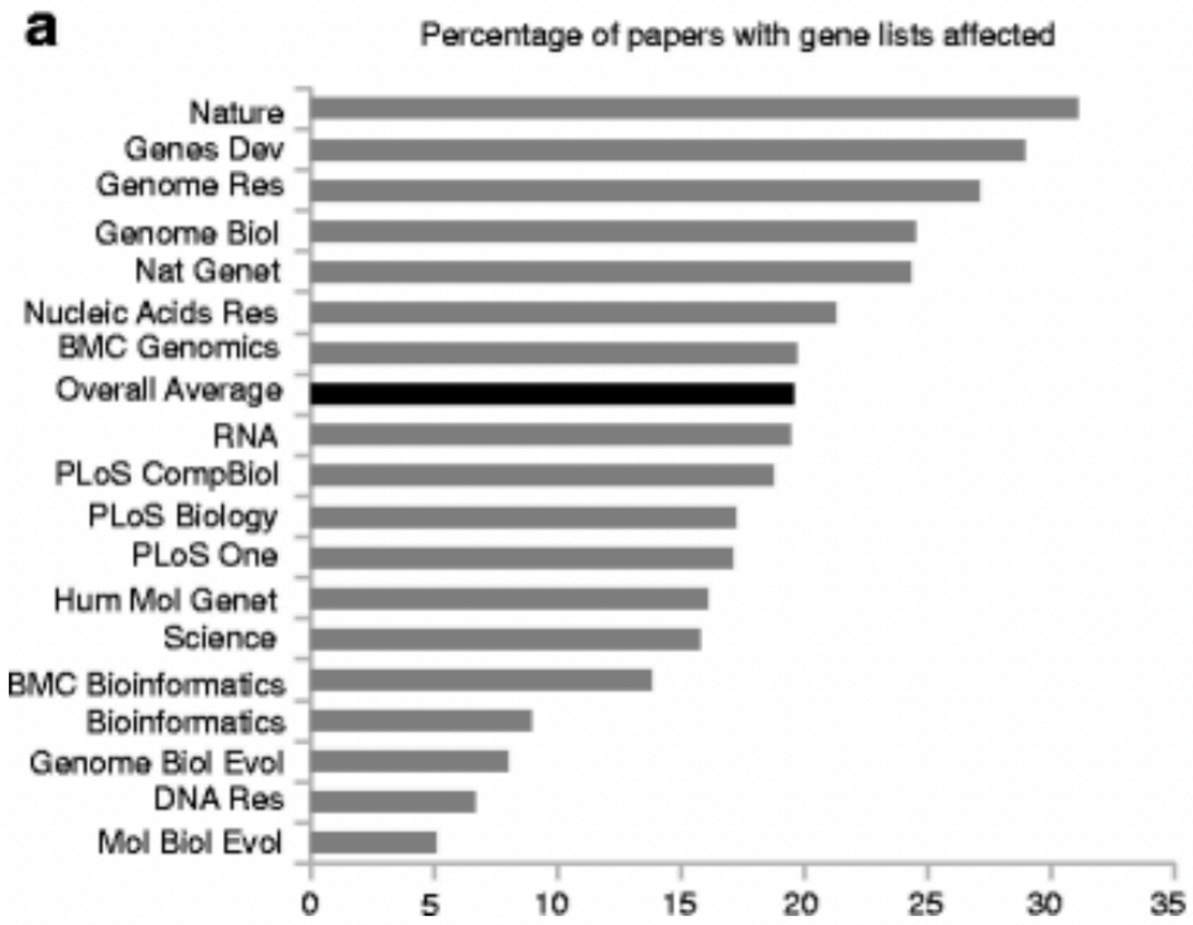
Error?

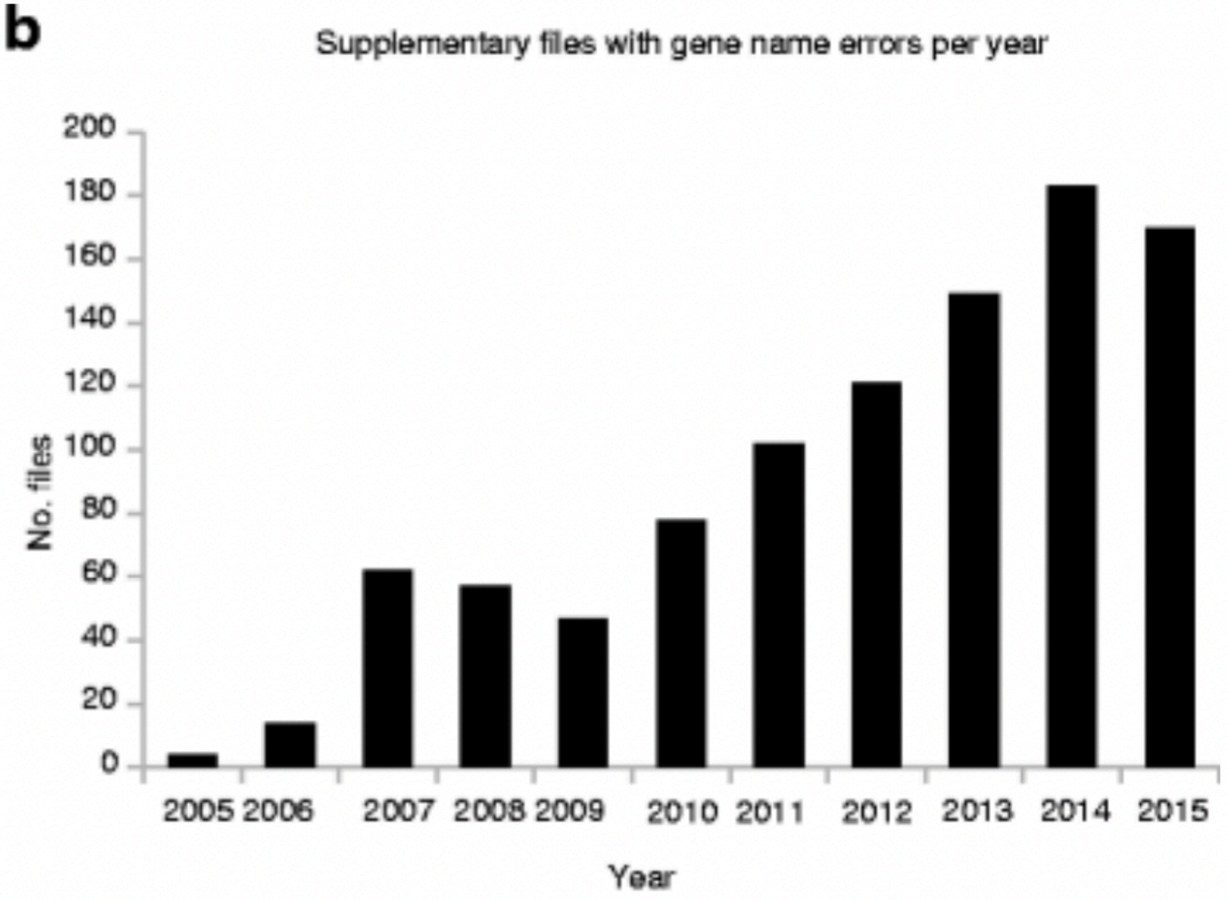


Automation?

Cr: Erik Andersen

Reproducible research





Ziemann et al. Genome Biology (2016) 17:177

What I hope you learn from this class:

- Biology is becoming more and more **quantitative**, so data analysis, reproducible research, and proper statistics skills are more important than ever

- Excel is great for some things, but data analysis should be **scripted** (in R or another language) in a way that any person could take your raw data and reproduce the figures in your manuscript

- Statistics can be slimy! (1) you must be **cautious and critical** when reviewing others' statistics and (2) you must be **transparent and honest** when providing your own analysis

Science, variability, and statistics

Response of sheep to anthrax

Response	Vaccinated	Not vaccinated
Died of anthrax	0	24
Survived	24	0
Total	**24 (100%)**	**0 (0%)**

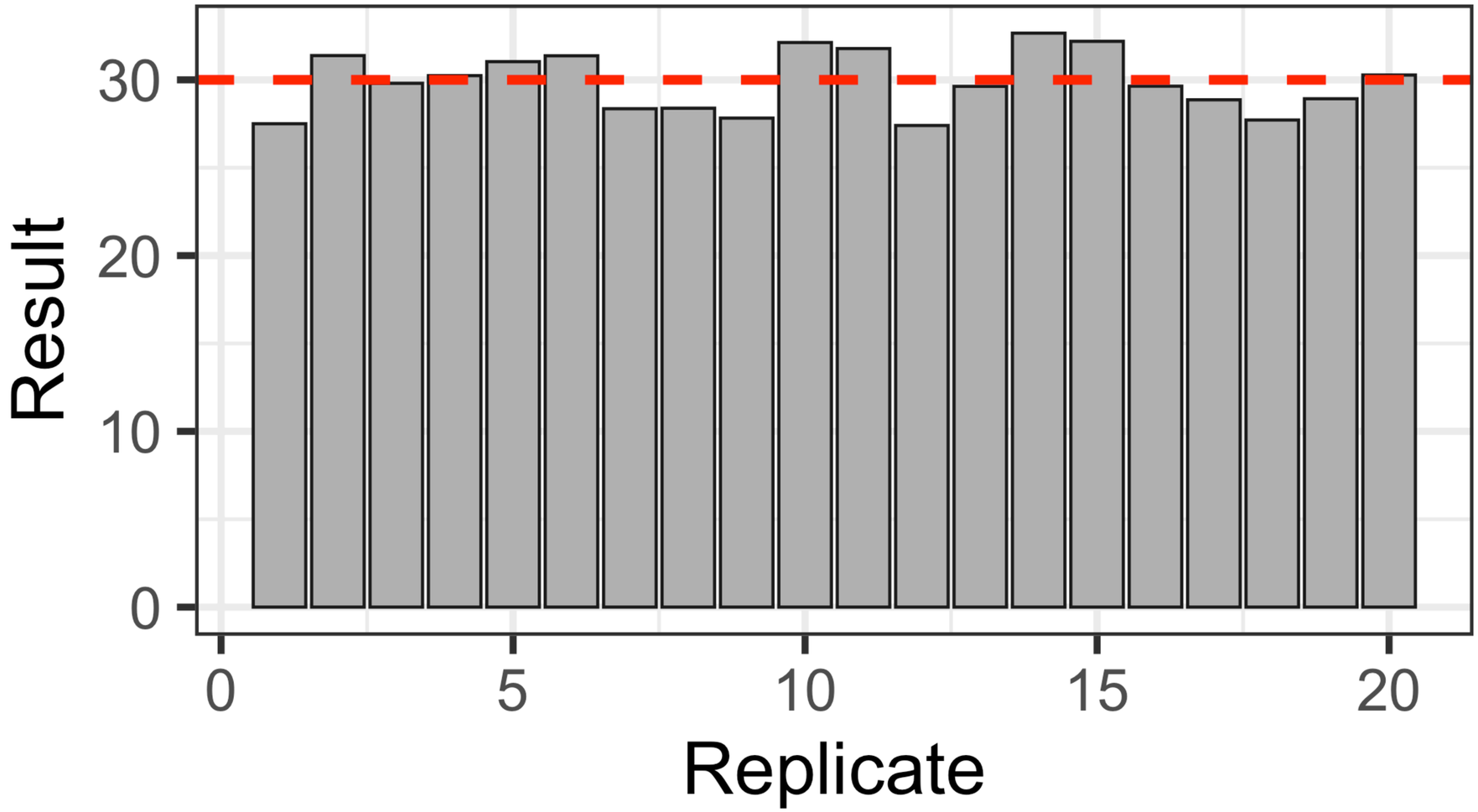
✓ ***Significant!***

Incidence of liver tumors in mice

Response	*E. coli*	Germ free
Liver tumors	8	19
No tumors	5	30
Total	**13 (62%)**	**49 (39%)**

Significant?

Science, variability, and statistics



Science, variability, and statistics

Response of sheep to anthrax

Response	Vaccinated	Not vaccinated
Died of anthrax	0	3 ~~24~~
Survived	3 ~~24~~	0
Total	3 ~~24~~ (100%)	0 (0%)

Incidence of liver tumors in mice

Response	*E. coli*	Germ free
Liver tumors	8	19
No tumors	5	30
Total	13 (62%)	49 (39%)

Significant?

Significant?

Observational studies v. experiments

Collecting data from subjects as an observer, not manipulating conditions

Survey about smoking habits and health

Performing a GWA/disease risk analysis on individuals with different genotypes

Collecting data in a controlled environment where researchers impose the conditions

Clinical trial for a new drug treatment

Knocking out a gene and then testing growth/survival with and without the gene

Controls: an essential part of any study

Collecting data from subjects as an observer, not manipulating conditions

(Smokers & non-smokers)
Survey about smoking habits and health

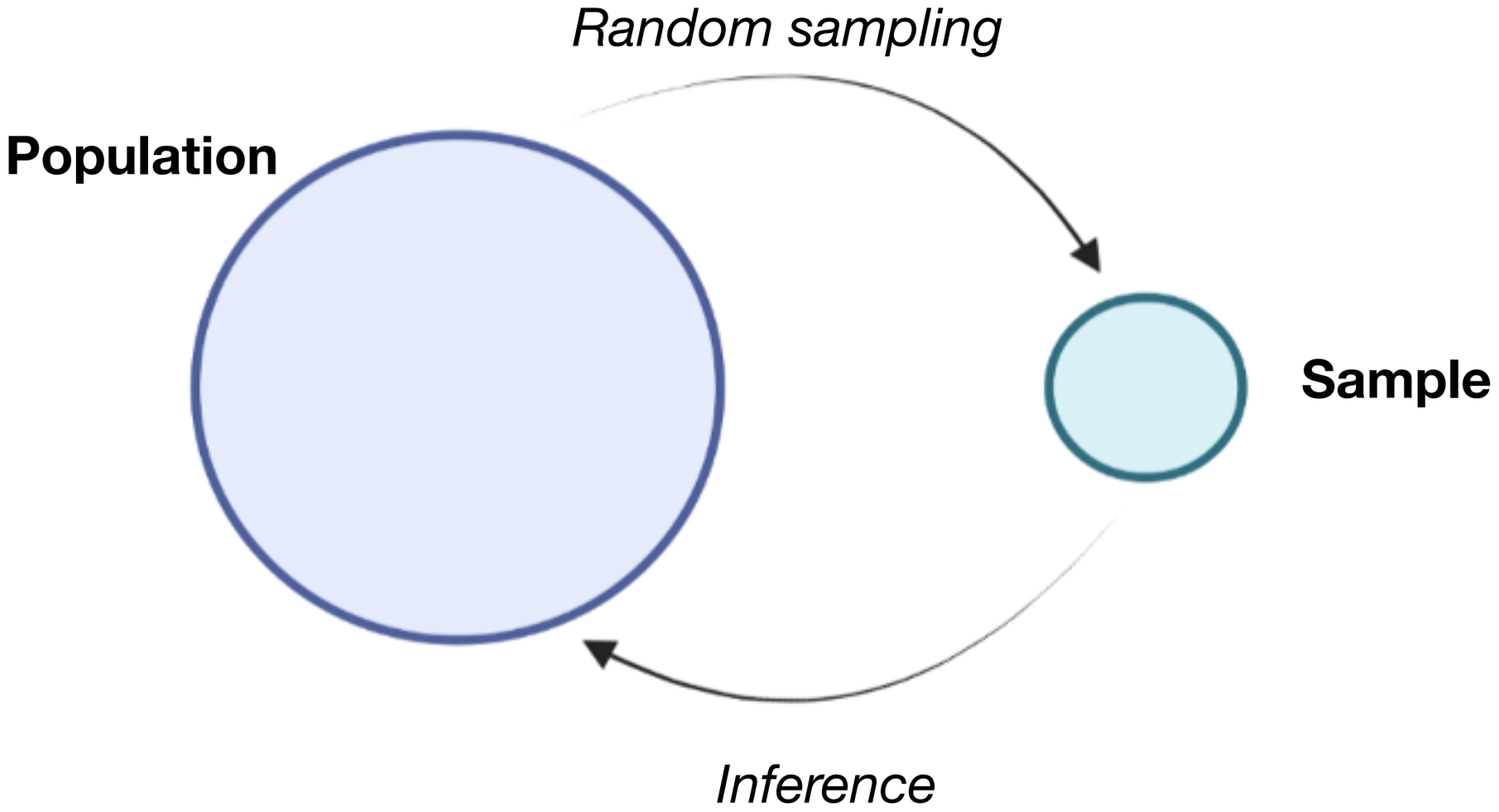
(Disease & non-disease)
Performing a GWA/disease risk analysis on individuals with different genotypes

Collecting data in a controlled environment where researchers impose the conditions

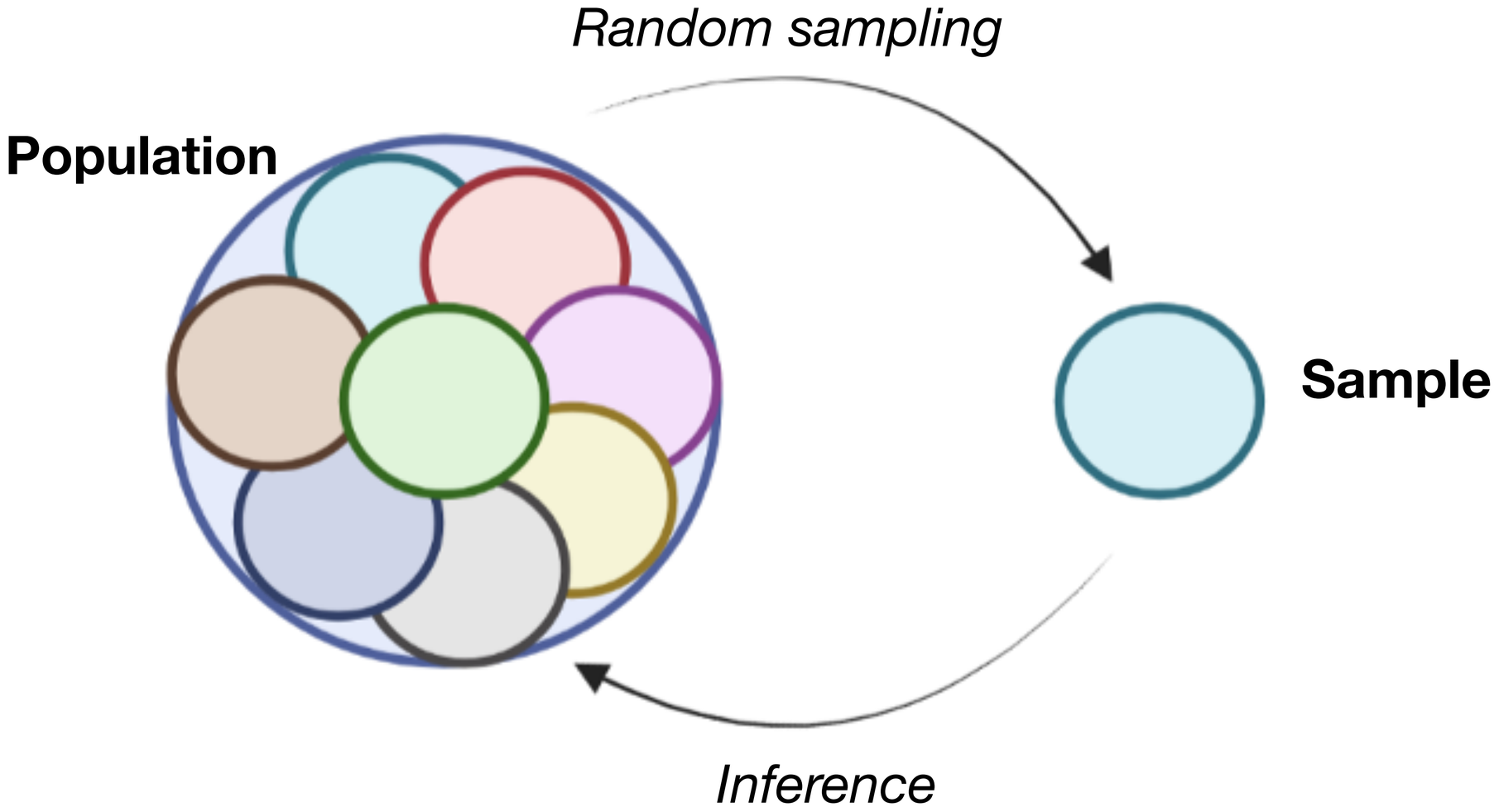
(Drug & placebo)
Clinical trial for a new drug treatment

(Knockout & wild-type)
Knocking out a gene and then testing growth/survival with and without the gene

Populations and sampling



Populations and sampling



Simple random sample

- Every member of the population has the **same** chance of being included in the sample
- Members of the sample are chosen **independently** of each other
  - *Not dependent on which other members are chosen*
- How we gather our data has ***tremendous*** implications on our choice of analysis methods and validity of our studies
  - ***There is no replacement for good/clean data!!!***

Types of variables

Categorical

Blood type (A, B, AB, O)

Fish sex (male, female)

Shape of pea (smooth, wrinkled)

Success in trial (Alive, dead)

Numeric

Human height

Blood cholesterol of patient

Number of bacterial colonies

Length of DNA segment

Types of variables

Discrete

Continuous

Categorical

Blood type (A, B, AB, O)

Fish sex (male, female)

Shape of pea (smooth, wrinkled)

Success in trial (Alive, dead)

Numeric

Human height

Blood cholesterol of patient

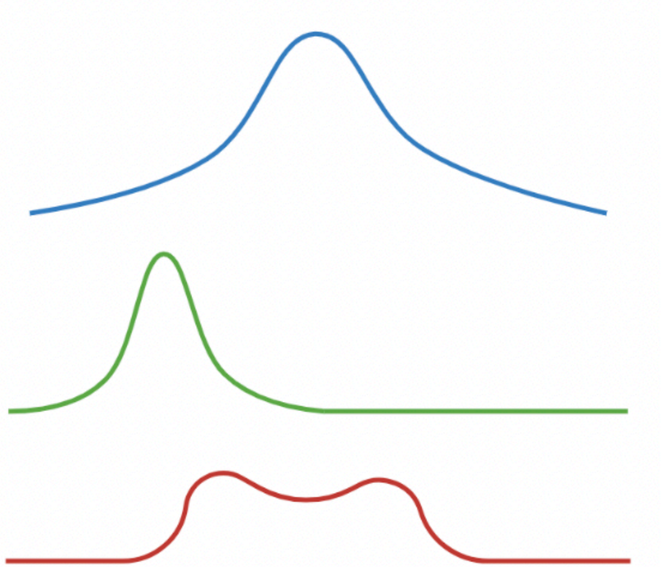
Number of bacterial colonies

Length of DNA segment

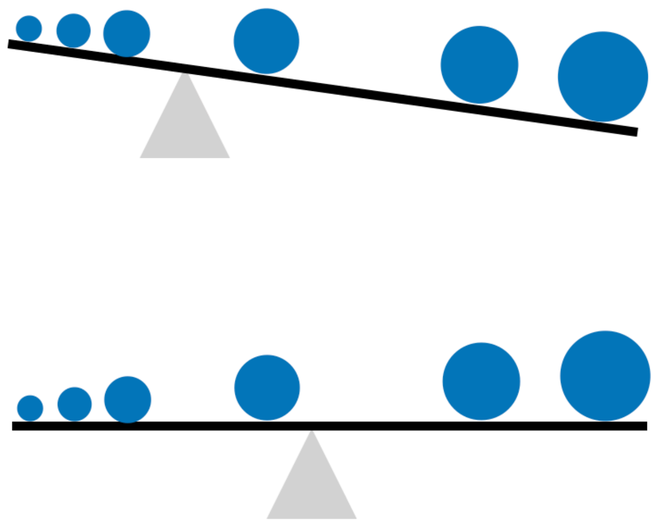
Discrete

Data exploration and summarization

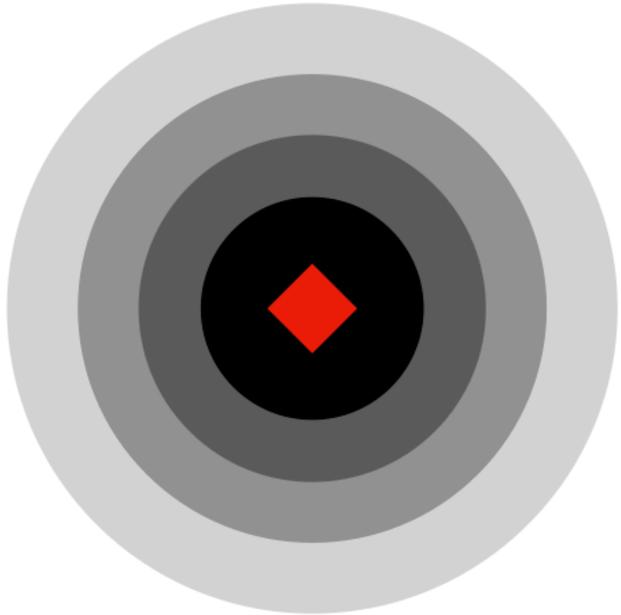
Shape



Center

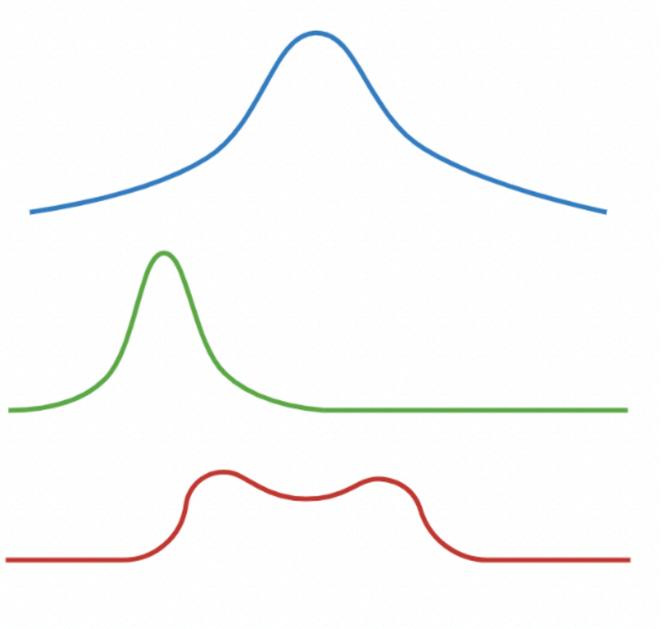


Spread



Data exploration and summarization

Shape



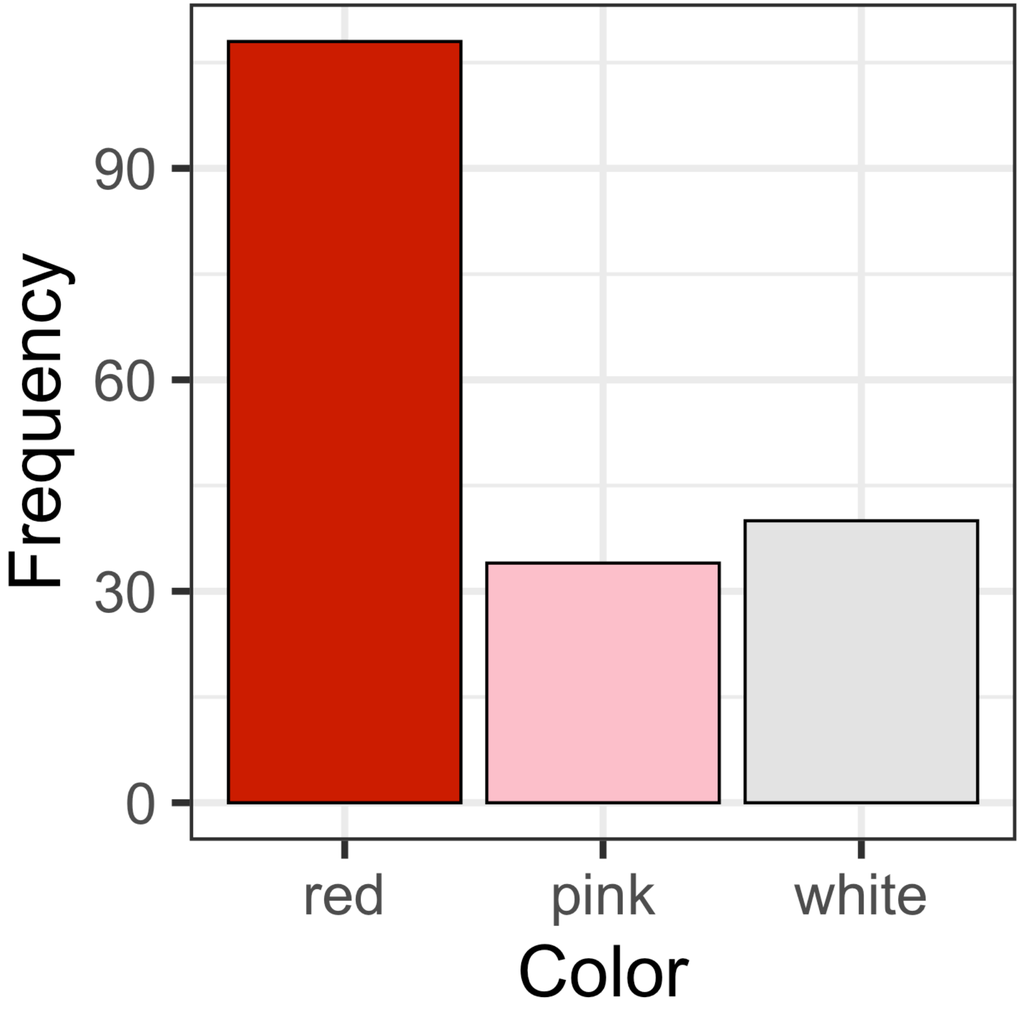
Center

Spread

Visualizing frequency distributions

Color	Frequency	Relative frequency	Percent frequency
Red	108	0.59	59.34
Pink	34	0.19	18.68
White	40	0.22	21.98
Total	**182**	**1**	**100**

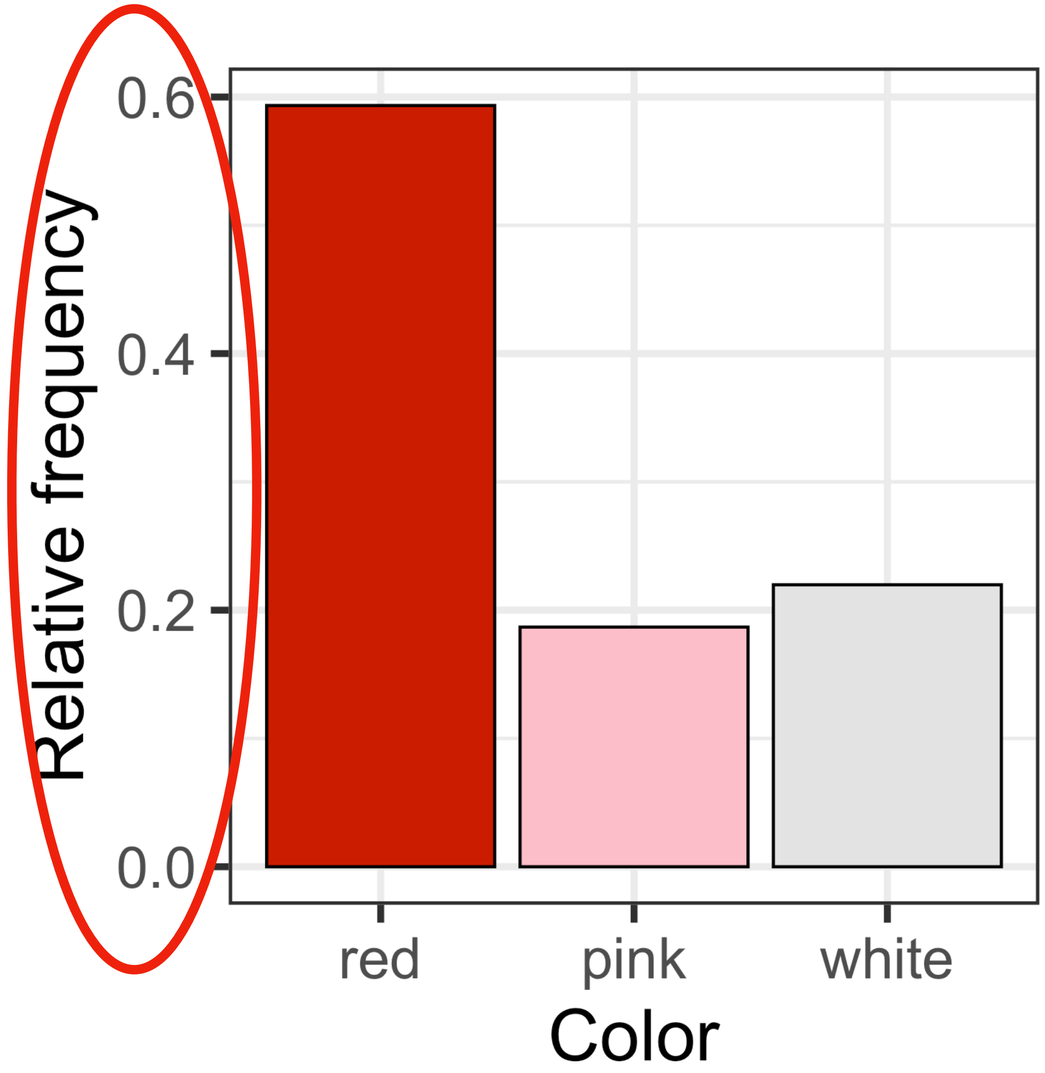
- Categorical data can be represented with a **bar chart**



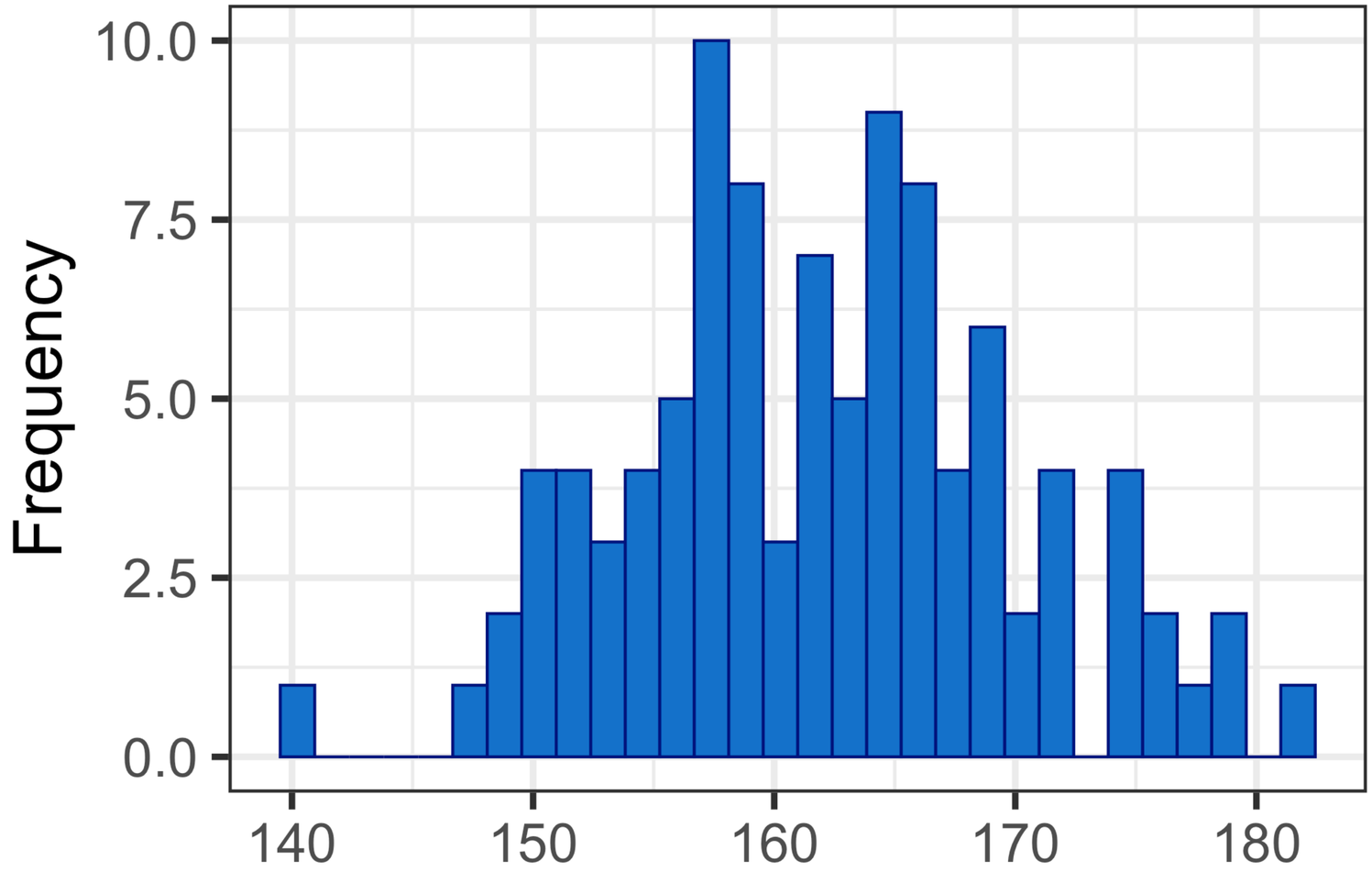
Visualizing frequency distributions

Color	Frequency	Relative frequency	Percent frequency
Red	108	0.59	59.34
Pink	34	0.19	18.68
White	40	0.22	21.98
Total	**182**	**1**	**100**

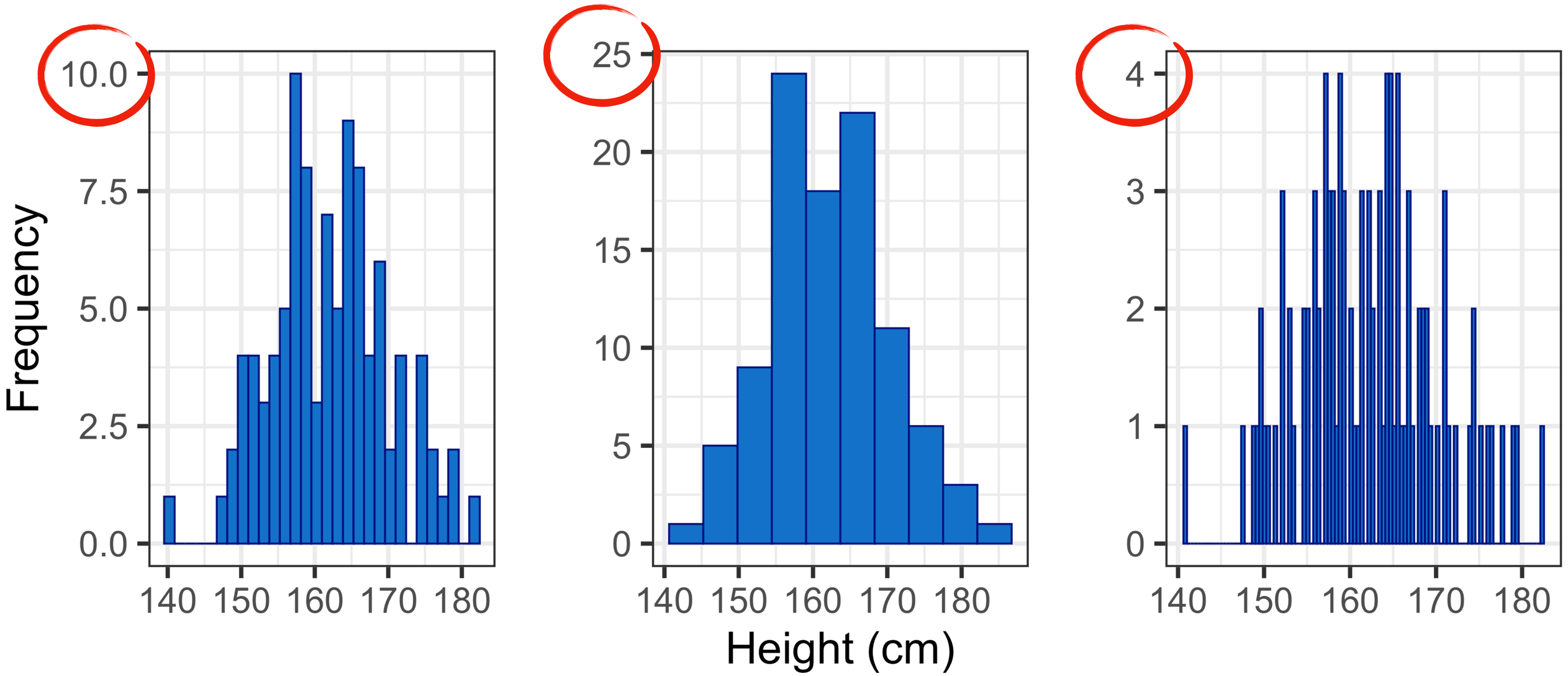
- Categorical data can be represented with a **bar chart**
- **Relative frequency** can be useful to compare datasets of different sizes



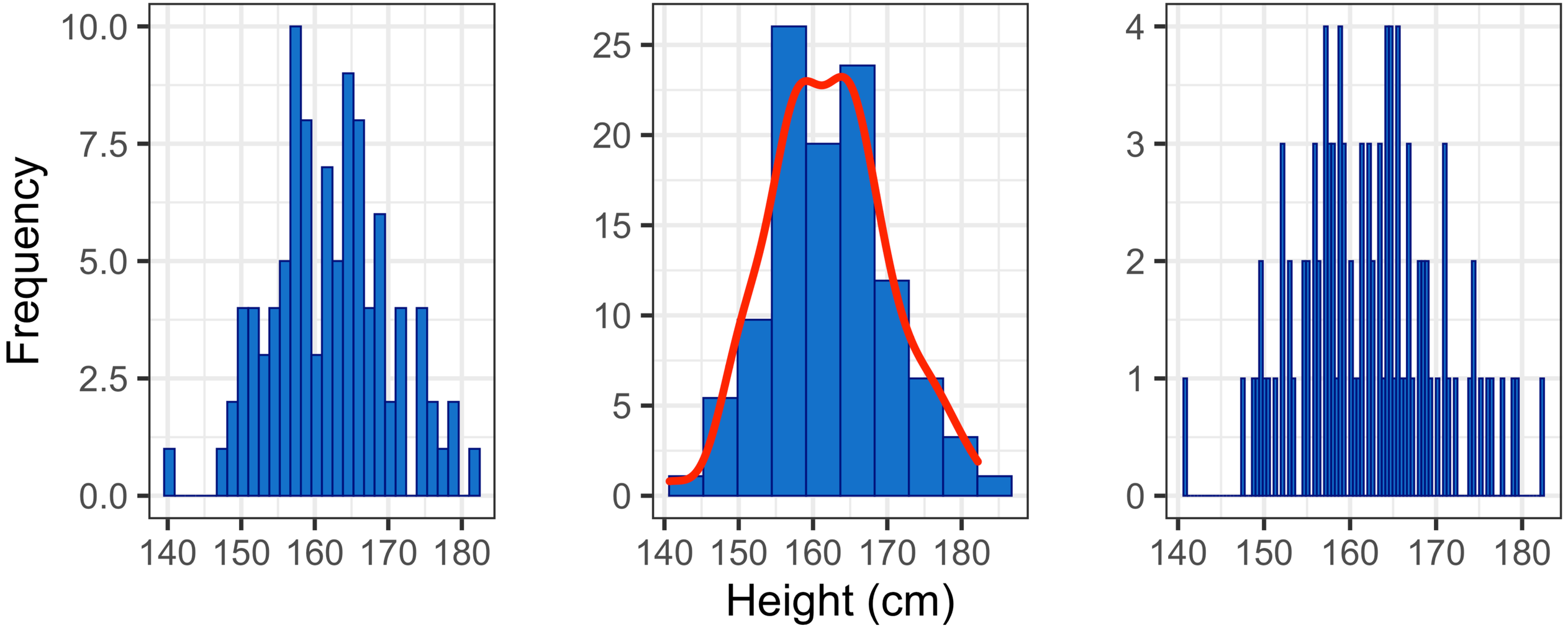
Histograms show shape of numerical data



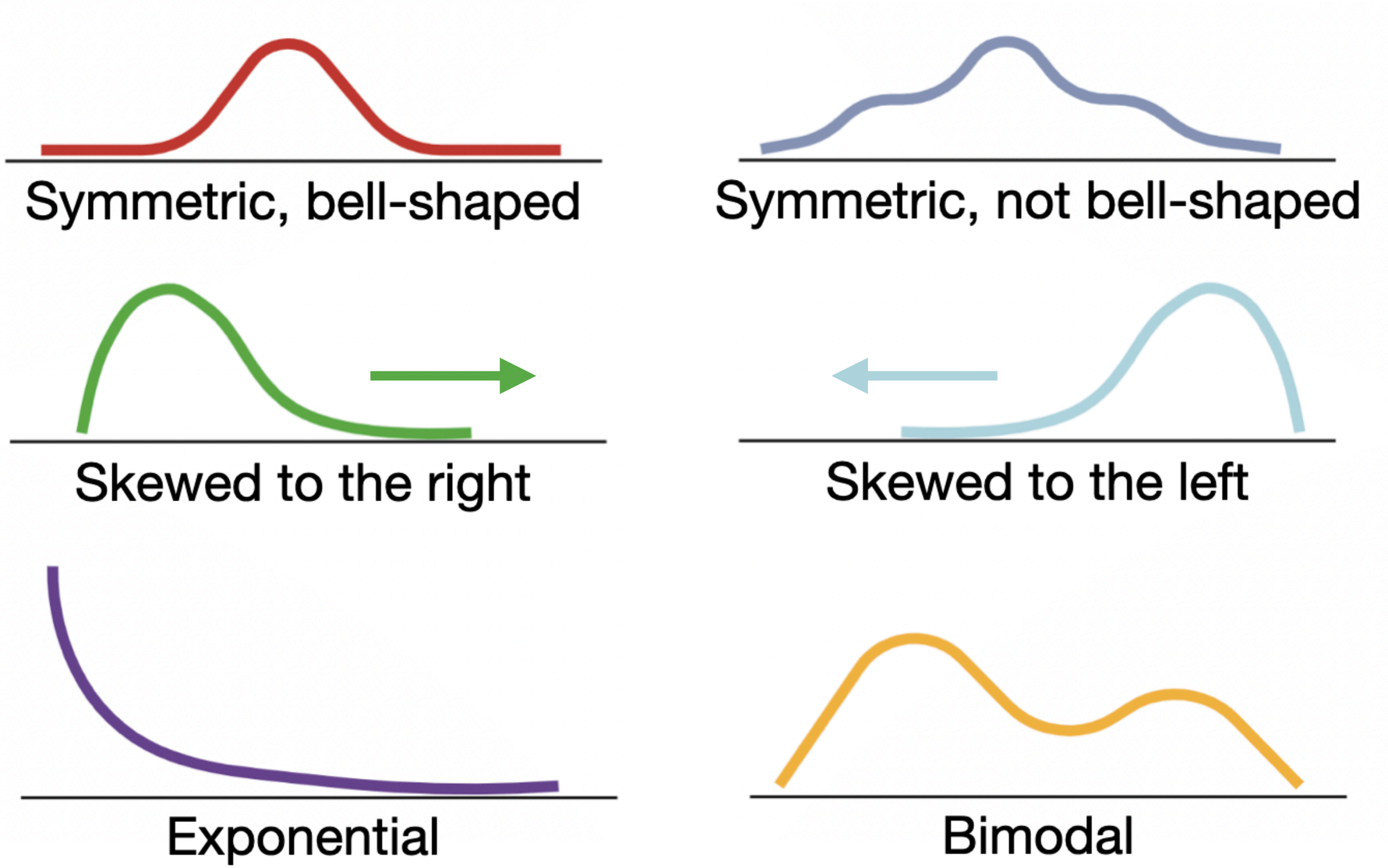
Histograms show shape of numerical data



Histograms show shape of numerical data



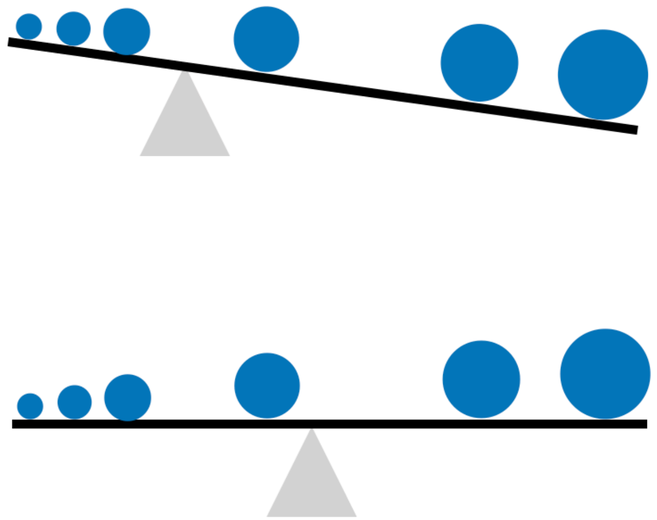
Biological frequency distribution shapes



Data exploration and summarization

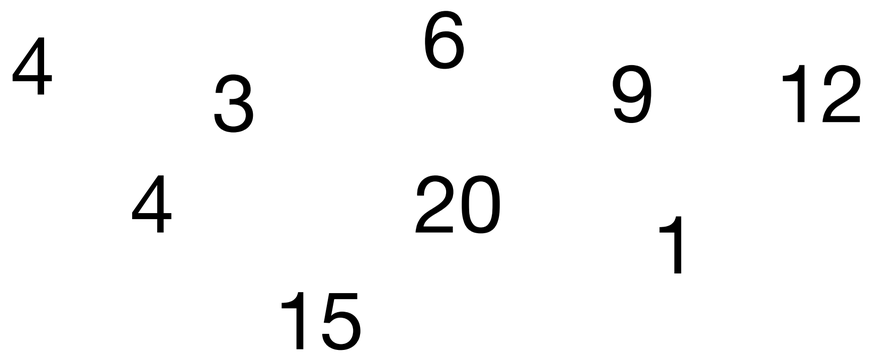
Shape

Center

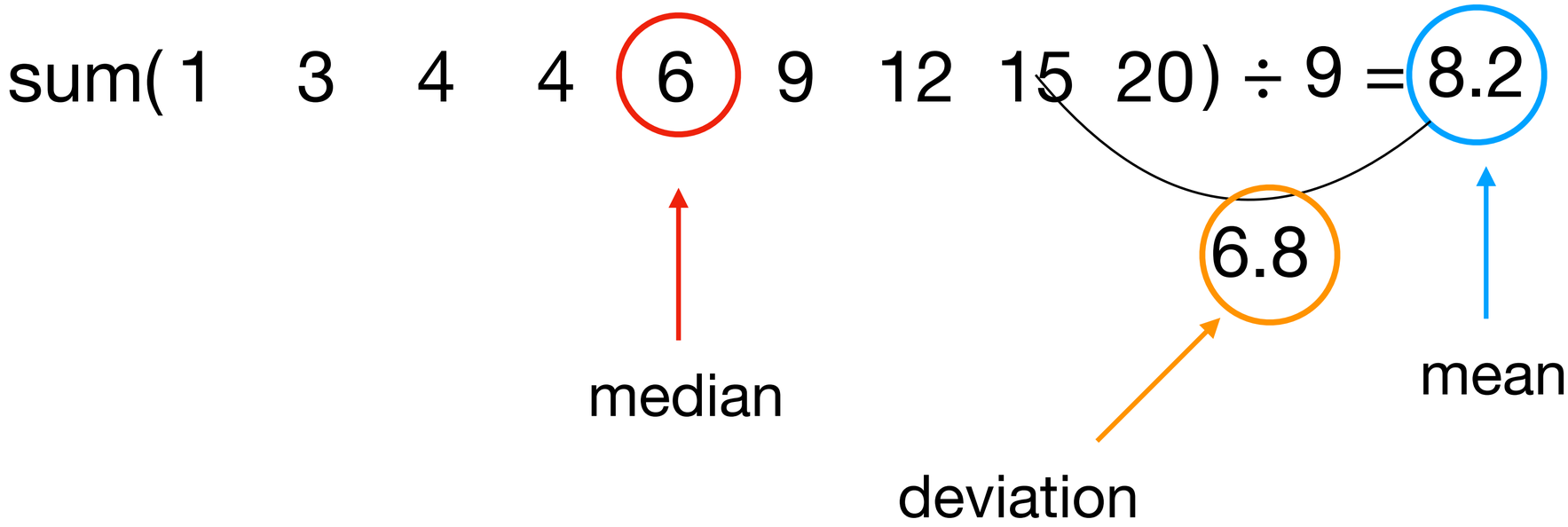


Spread

Measures of center: **median** vs. **mean**



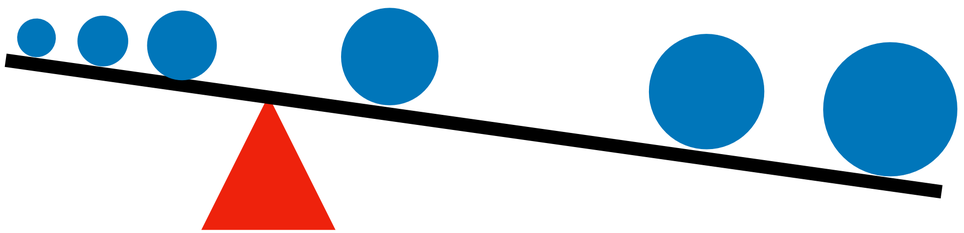
Measures of center: **median** vs. **mean**



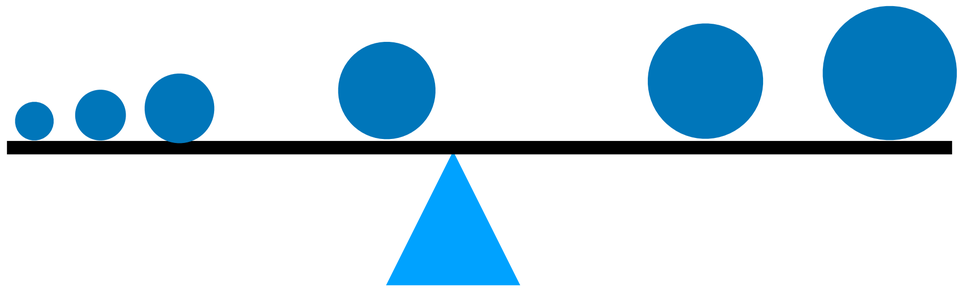
- Median is a more **robust** statistic
- `median(x)`

- Mean can be influenced by extremes
- `mean(x)`

Measures of center: **median** vs. **mean**



- Median is a more **robust** statistic
- `median(x)`



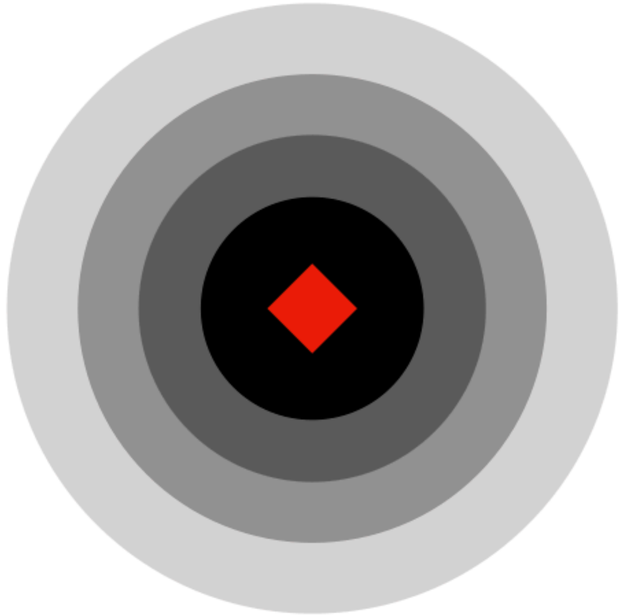
- Mean can be influenced by extremes
- `mean(x)`

Data exploration and summarization

Shape

Center

Spread



Measures of dispersion: **range** and **IQR**

-10, 0, 10, 20, 30

median: 10

mean: 10

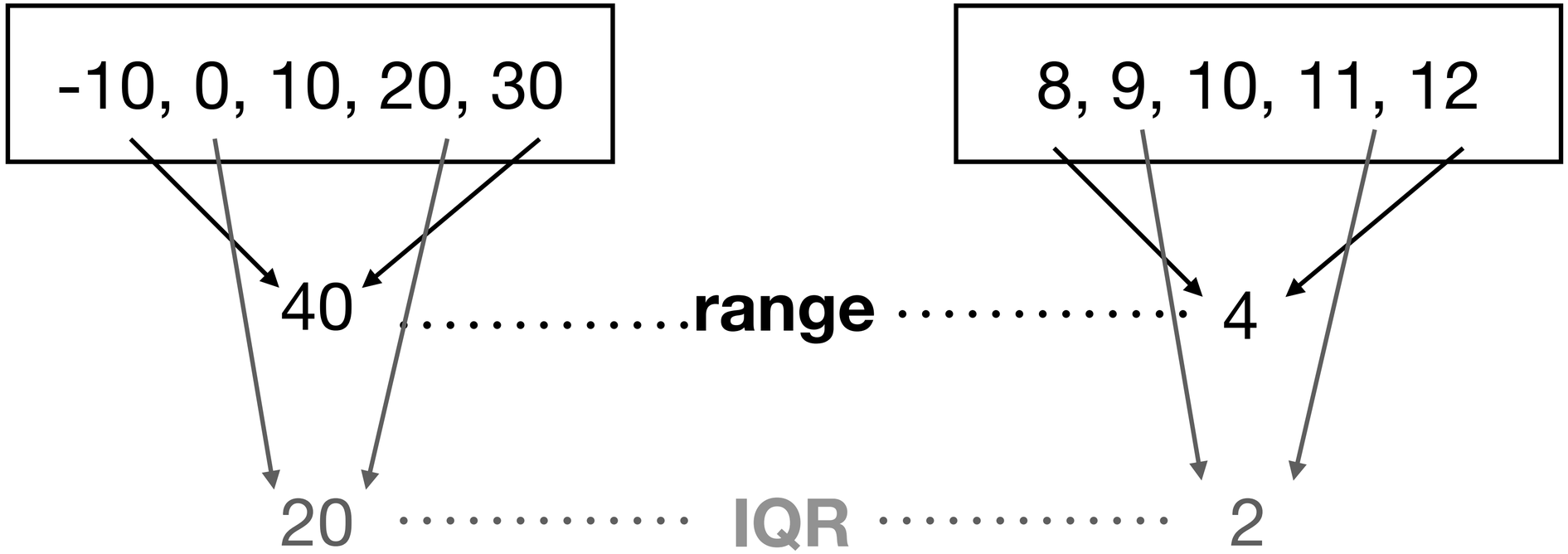
???

8, 9, 10, 11, 12

median: 10

mean: 10

Measures of dispersion: **range** and **IQR**



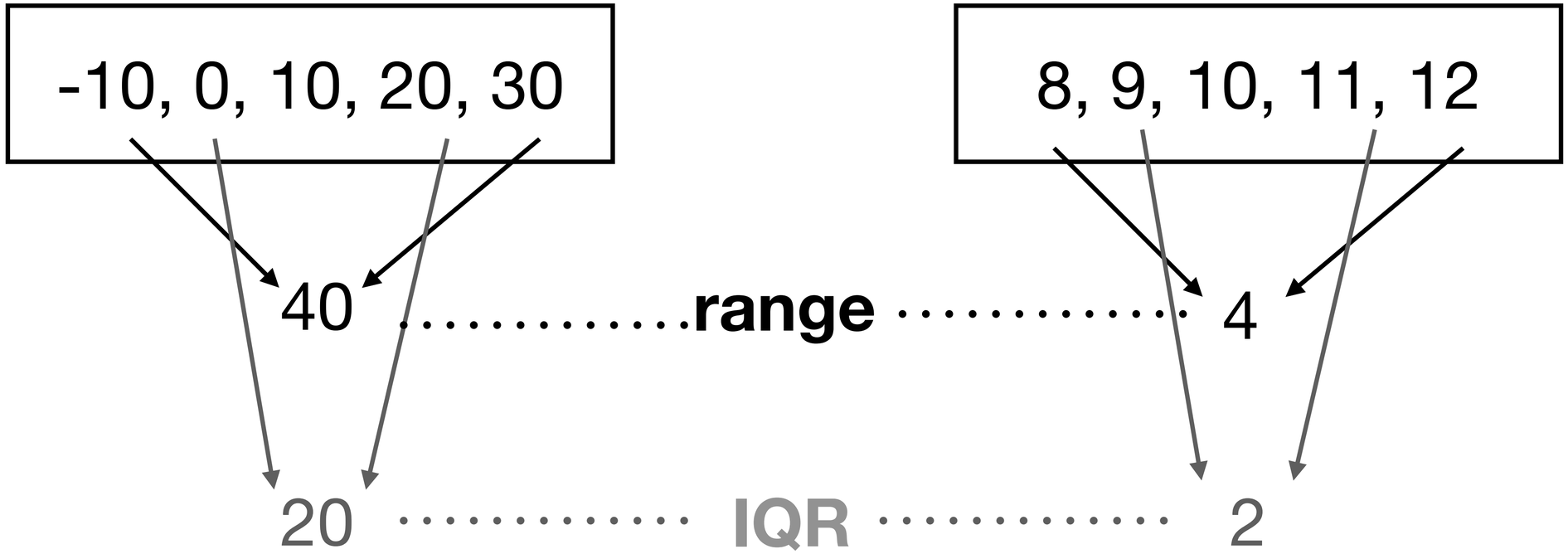
median: 10

mean: 10

median: 10

mean: 10

Measures of dispersion: **range** and **IQR**



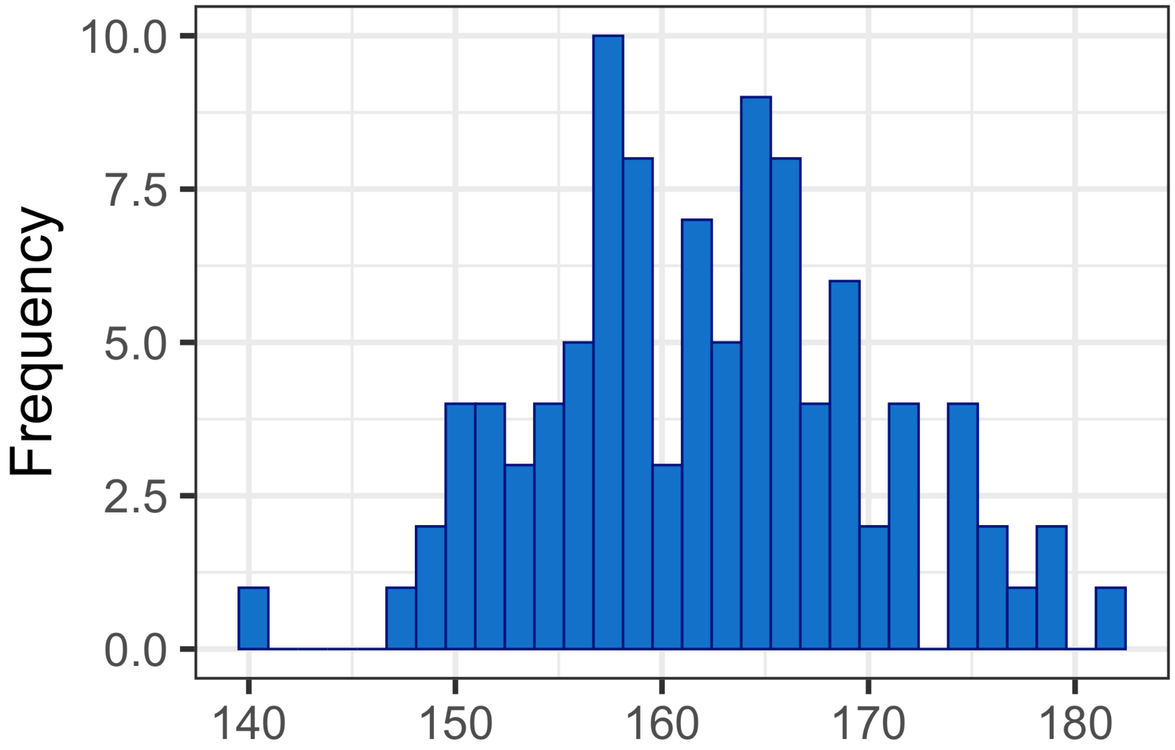
> median: 10

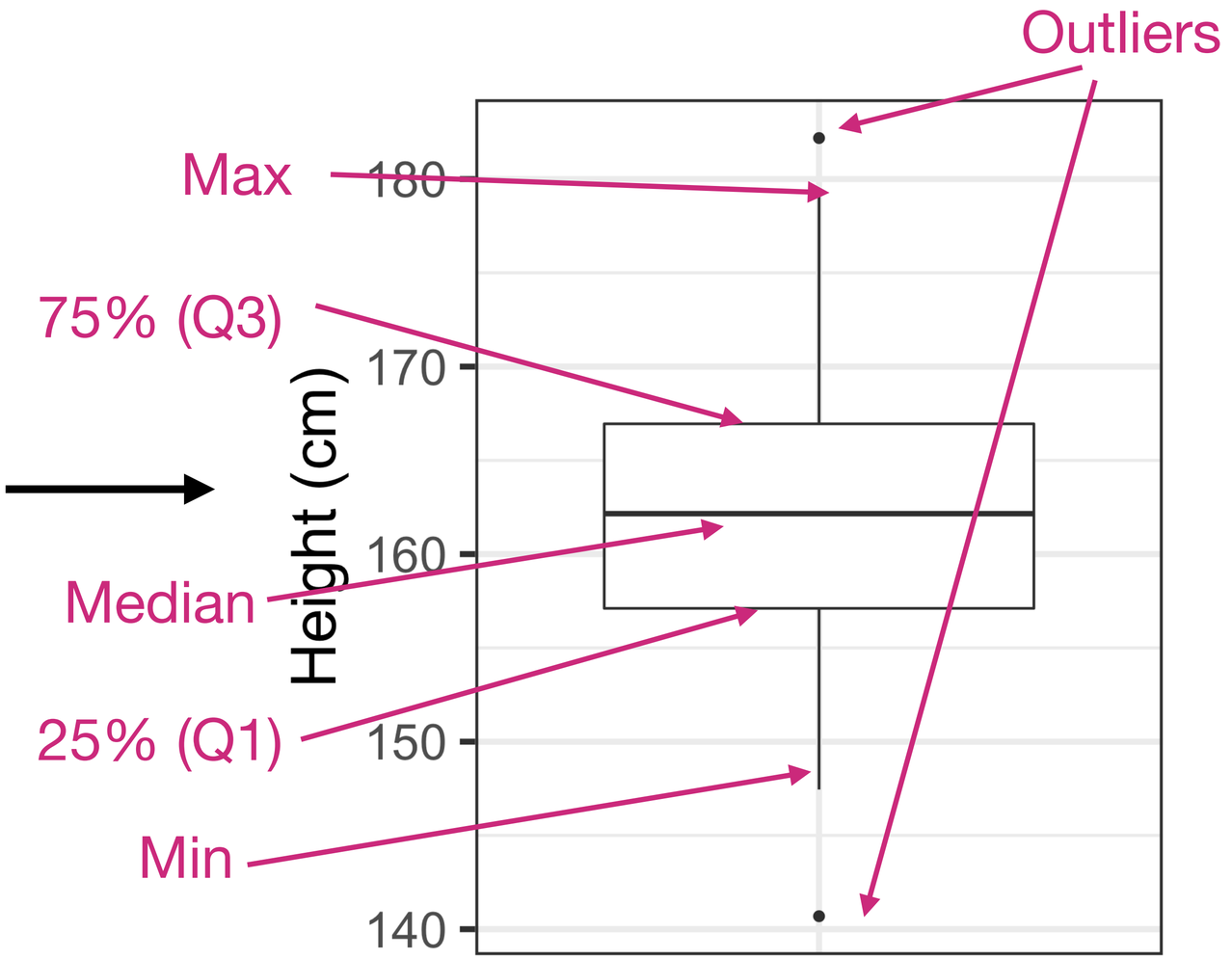
mean: 10

> median: 10

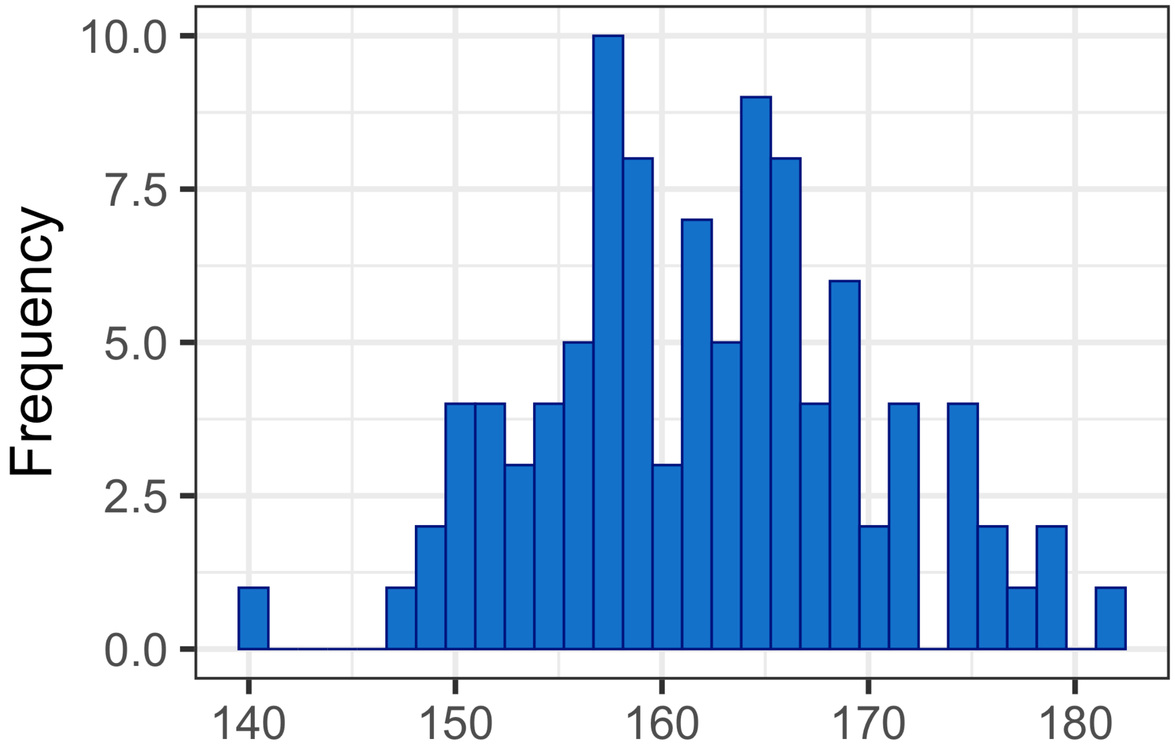
mean: 10

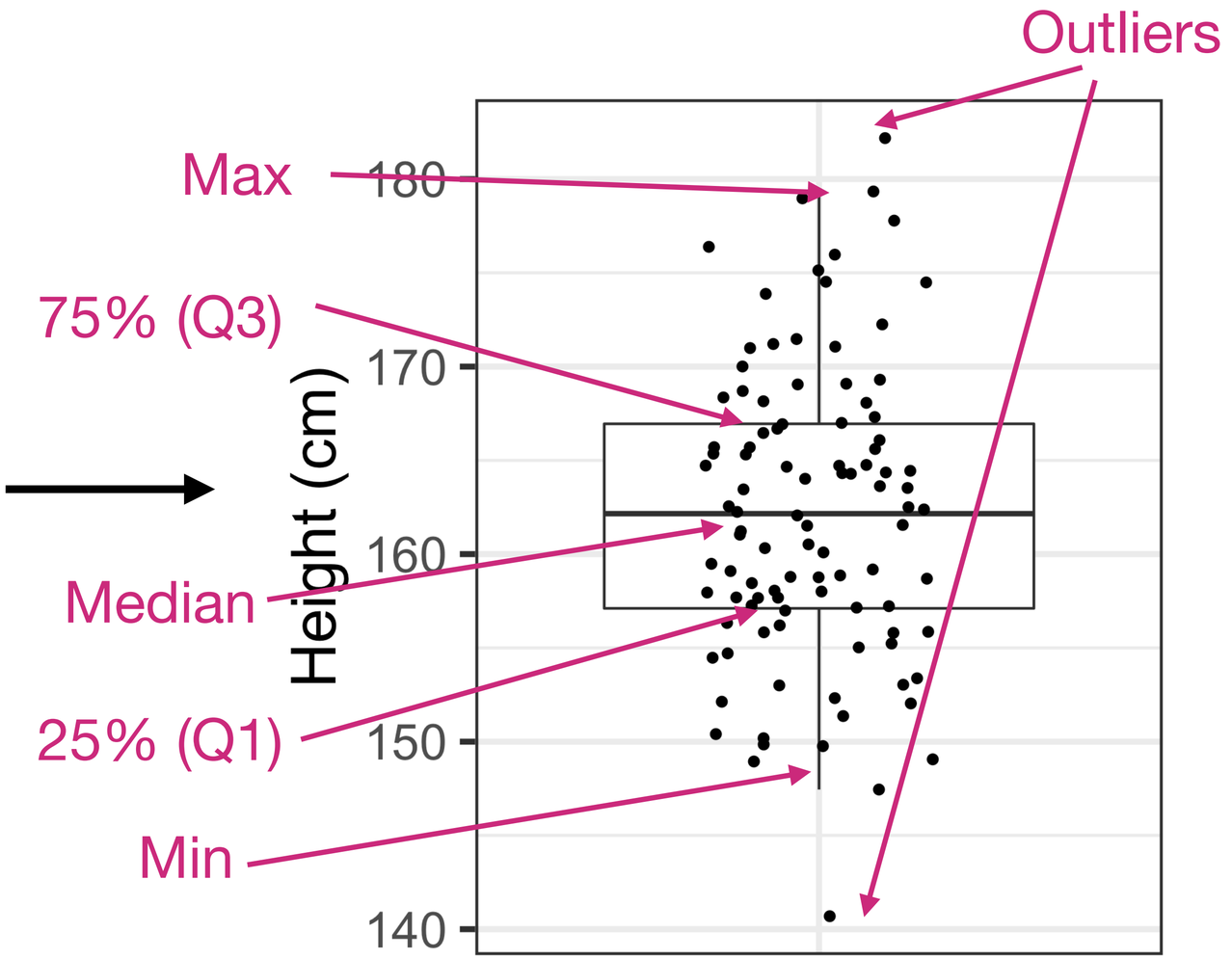
Plotting distributions with boxplots





Plotting distributions with boxplots





Measures of dispersion: **standard deviation**

-10, 0, 10, 20, 30

8, 9, 10, 11, 12

Variance: Mean of the squared deviations

Standard deviation: square root of the variance

median: 10

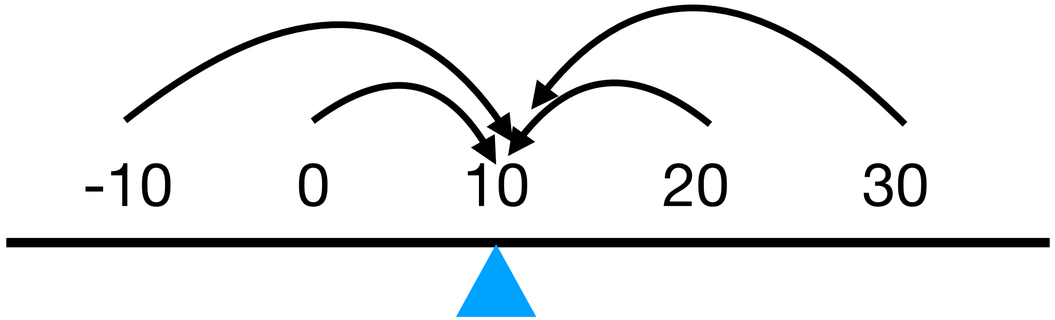
> mean: 10

median: 10

> mean: 10

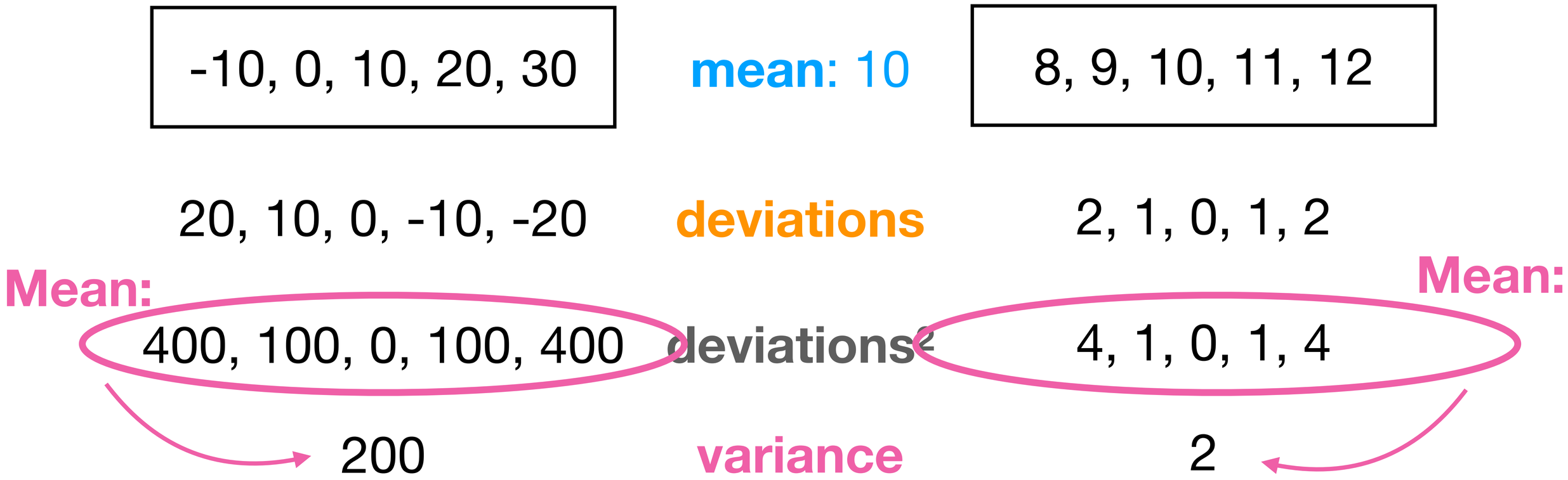
Measures of dispersion: **standard deviation**

-10, 0, 10, 20, 30	**mean:** 10	8, 9, 10, 11, 12
20, 10, 0, -10, -20	**deviations**	2, 1, 0, 1, 2



Variance: Mean of the squared deviations

Measures of dispersion: **standard deviation**



Variance: Mean of the squared deviations

Measures of dispersion: **standard deviation**

-10, 0, 10, 20, 30	**mean**: 10	8, 9, 10, 11, 12
20, 10, 0, -10, -20	**deviations**	2, 1, 0, 1, 2
400, 100, 0, 100, 400	**deviations^2**	4, 1, 0, 1, 4
200	**variance**	2
14.1	**standard deviation**	1.41

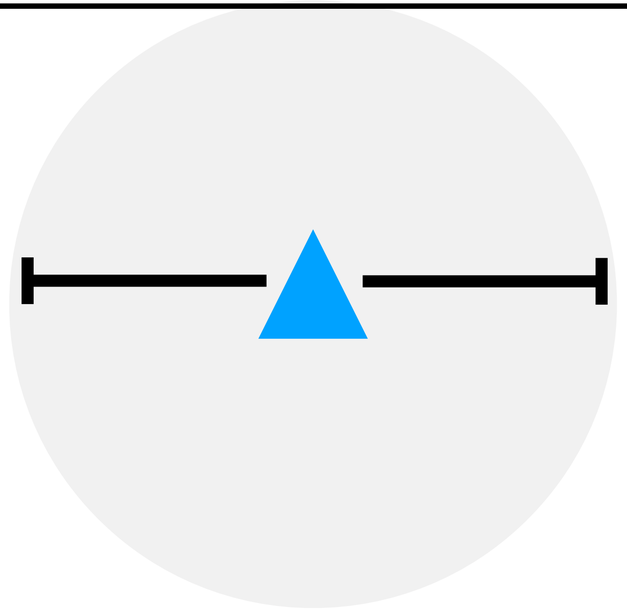
Standard deviation: square root of the variance

Measures of dispersion: **standard deviation**

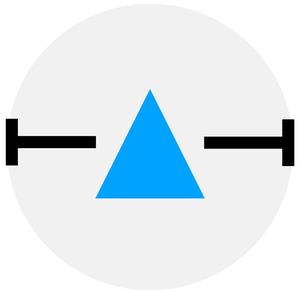
-10, 0, 10, 20, 30

mean: 10

8, 9, 10, 11, 12



"typical distance of the observations from their mean"



14.1

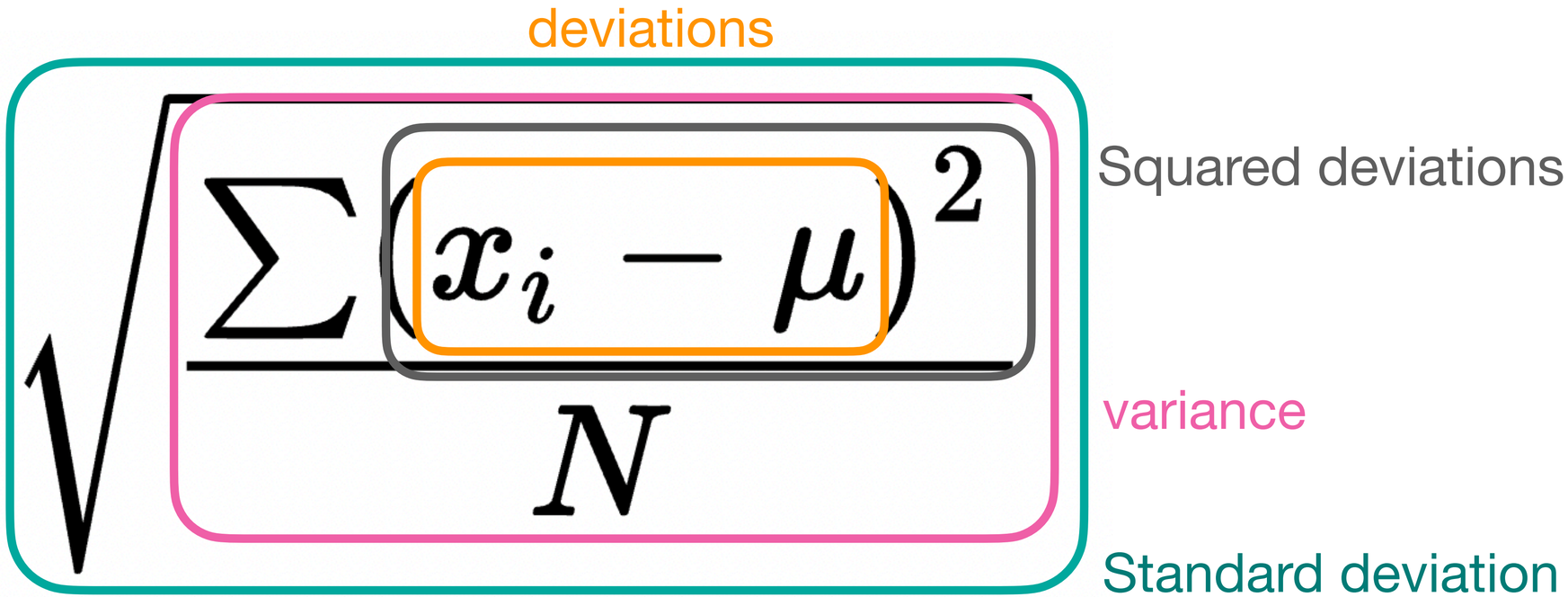
standard deviation

1.41

Standard deviation: square root of the variance

Measures of dispersion: **standard deviation**

```
> var(x)
> sd(x)
```



$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ**:** population standard deviation

x_i **:** observed values

N: Size of population

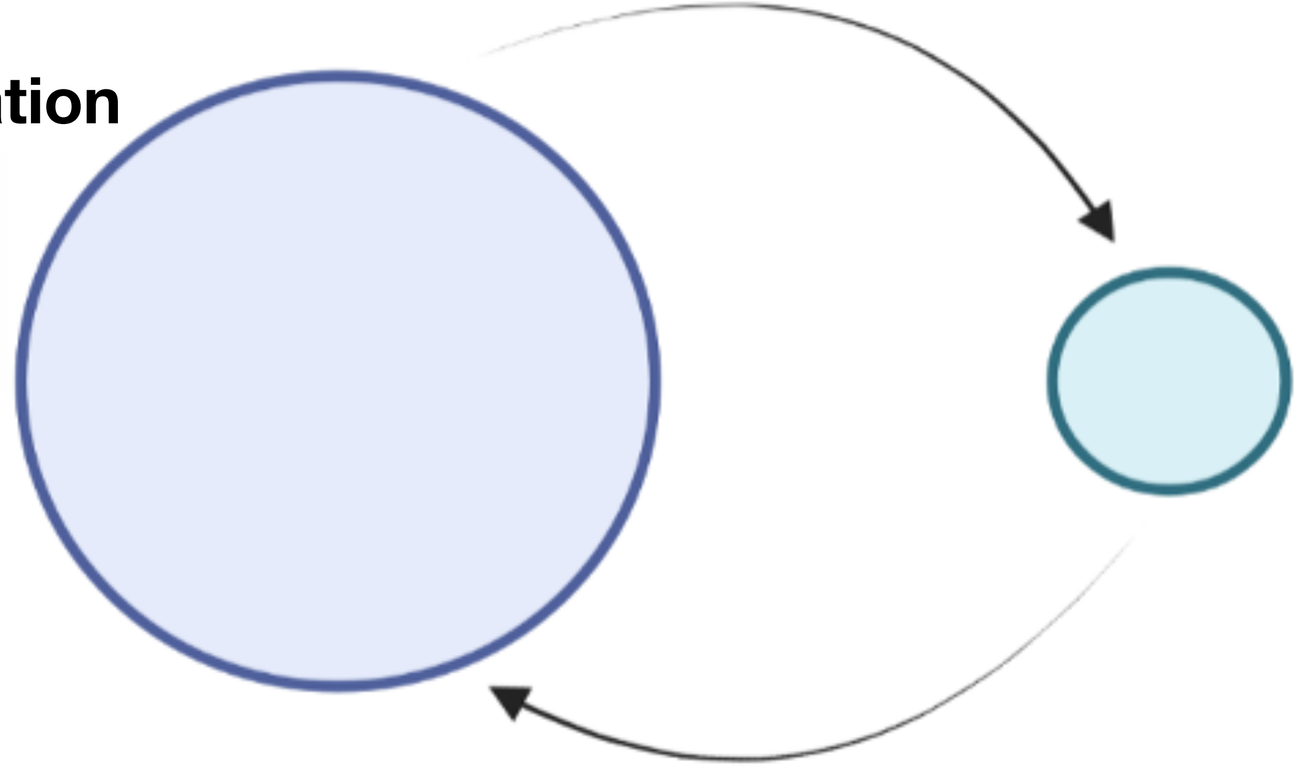
μ **:** population mean

Similar, but different, equations describe a population vs. a sample

Random sampling

Population

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$



Sample

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

Inference

Measures of dispersion: **standard deviation**

```
> var(x)
> sd(x)
```

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

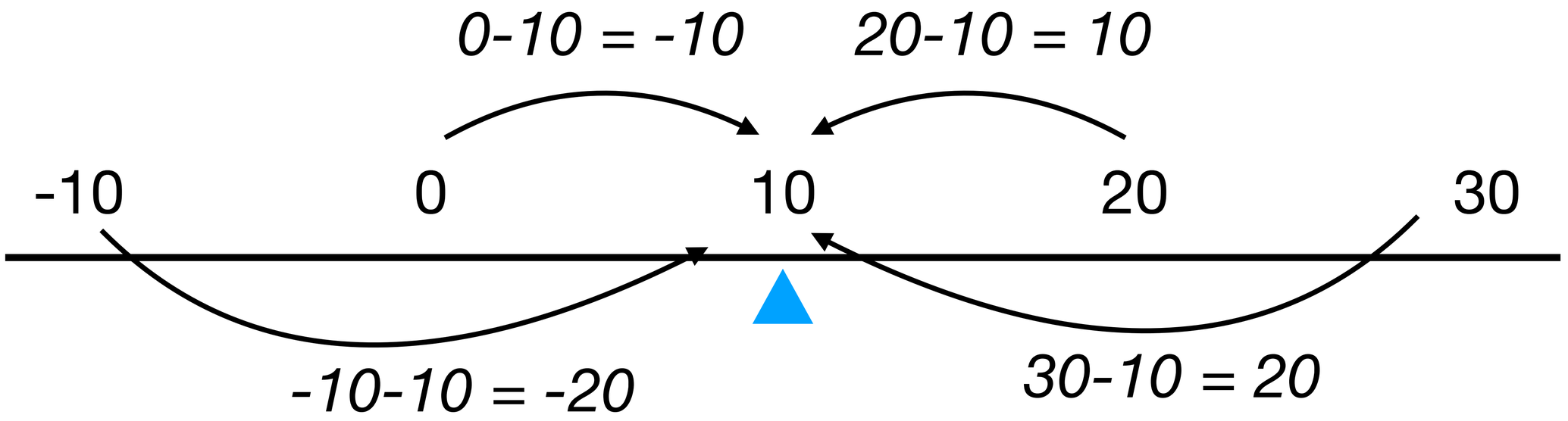
s: sample standard deviation

N: number of observations

x_i **:** observed values

$\bar{x}$ **:** mean value of samples

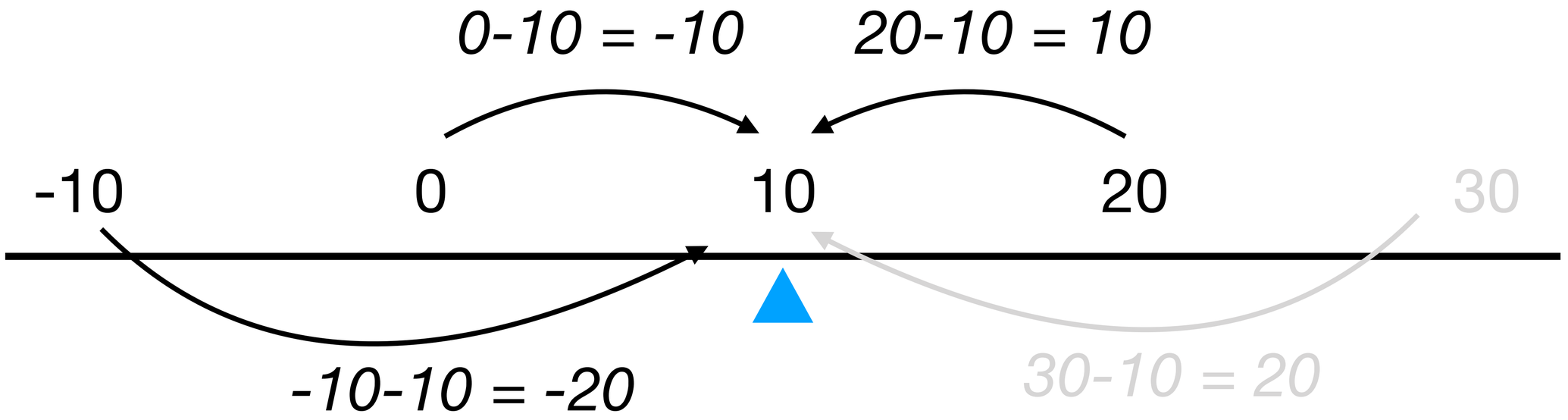
N-1: **degrees of freedom** explained



(-20) + (-10) + (10) + (20) = 0

Sum of deviations is always zero!

N-1: **degrees of freedom** explained

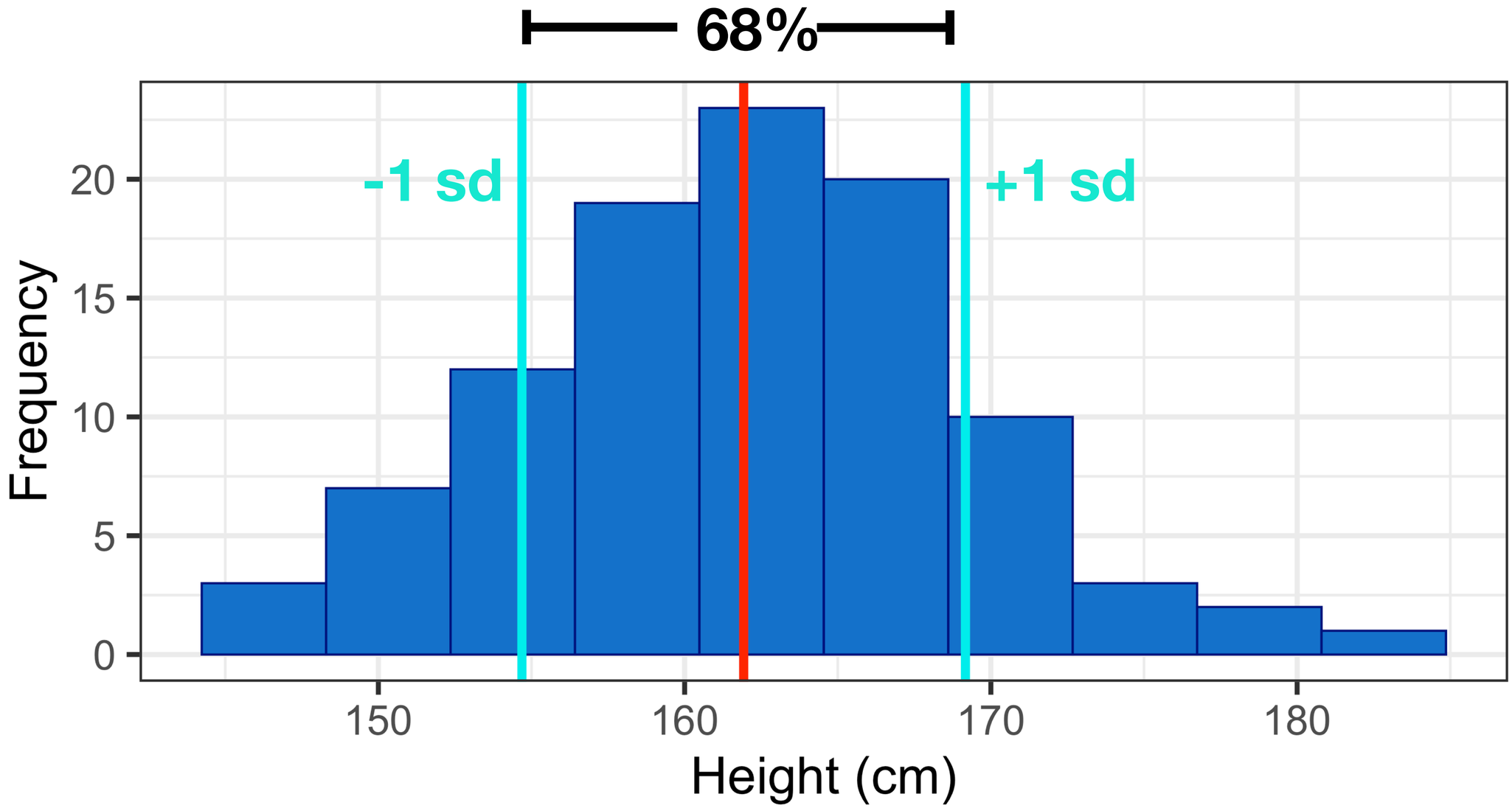


(-20) + (-10) + (10) + X = 0

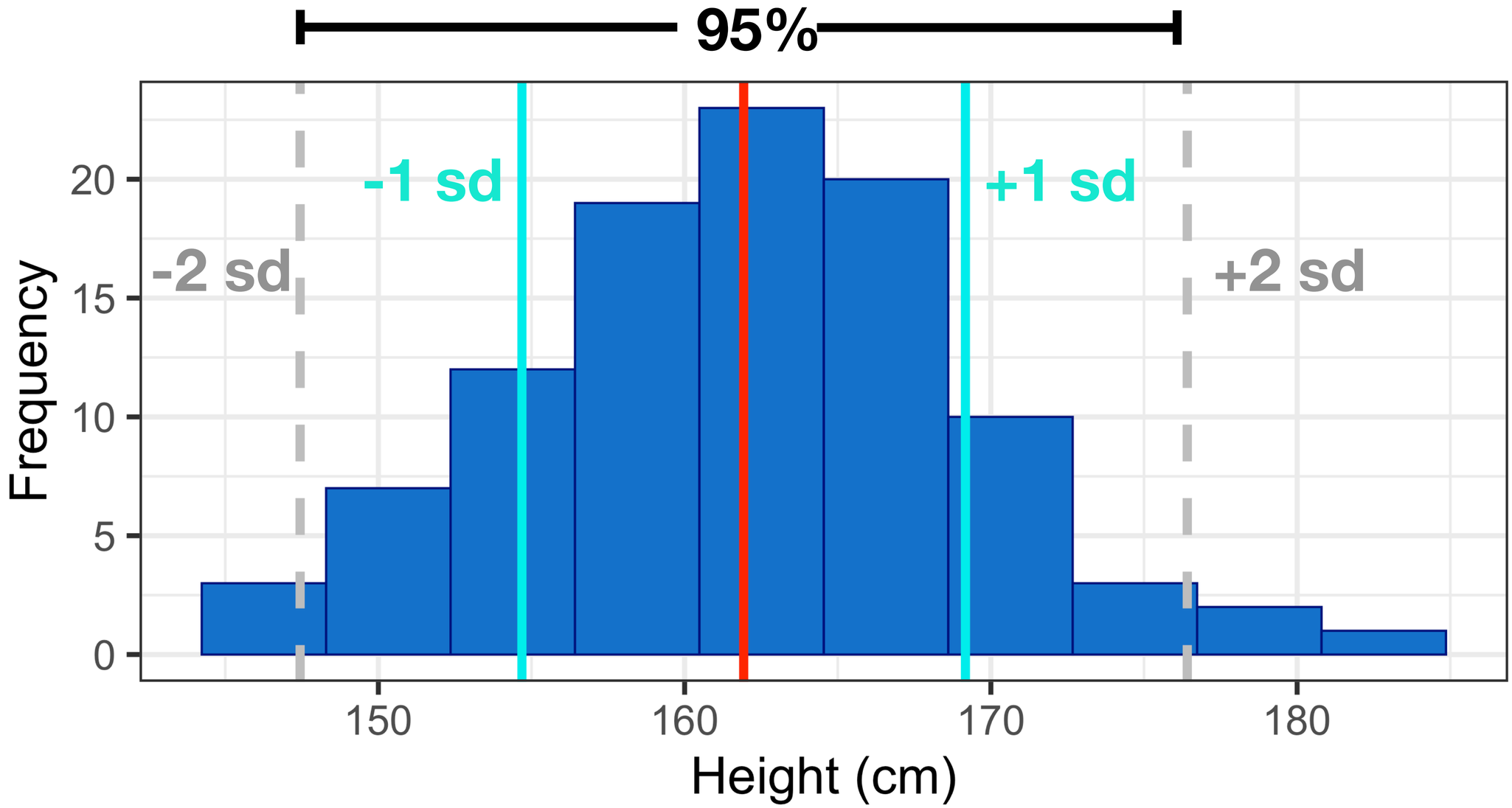
Has to be +20 from mean…

Sum of deviations is always zero!

Measures of dispersion: **standard deviation**



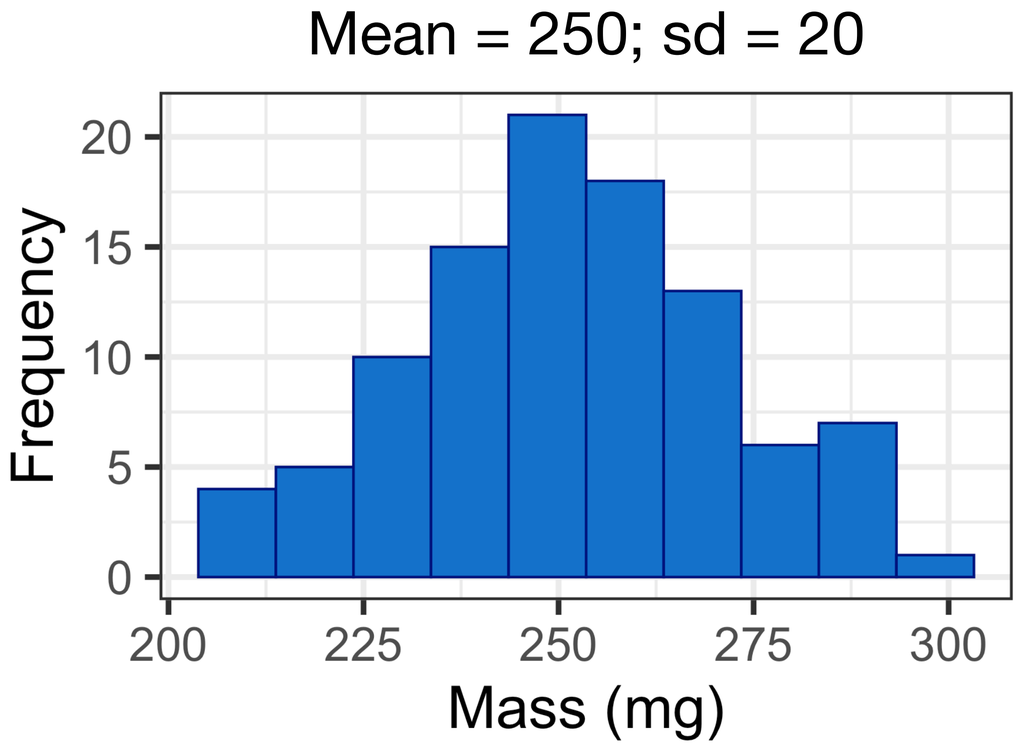
Measures of dispersion: **standard deviation**



Linear transformation of variables

Suppose you collected data in milligrams but need to convert to grams.

```
new_data = (1/1000) * old_data
```

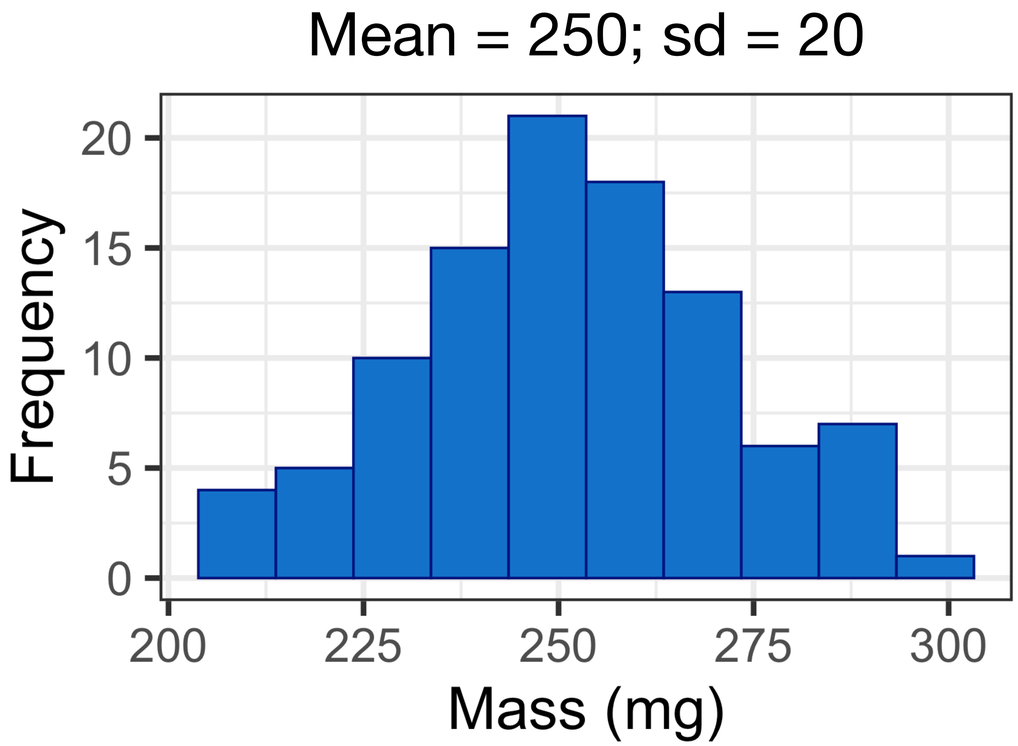


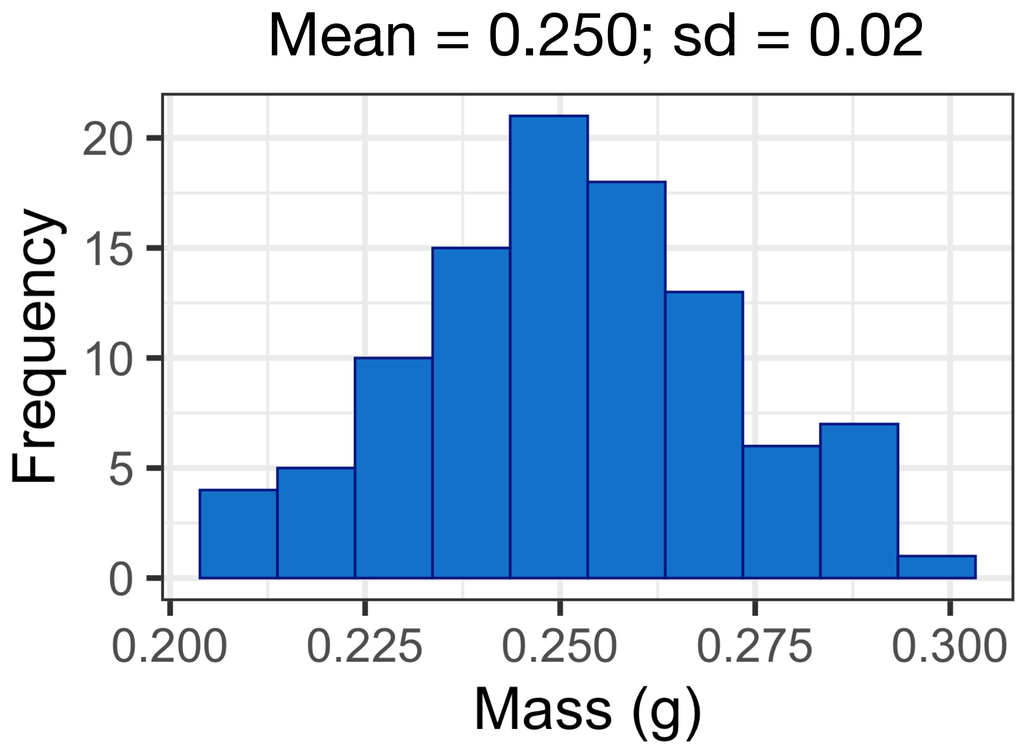
Linear transformation of variables

Suppose you collected data in milligrams but need to convert to grams.

```
new_data = (1/1000) * old_data
```

Mean & SD scale

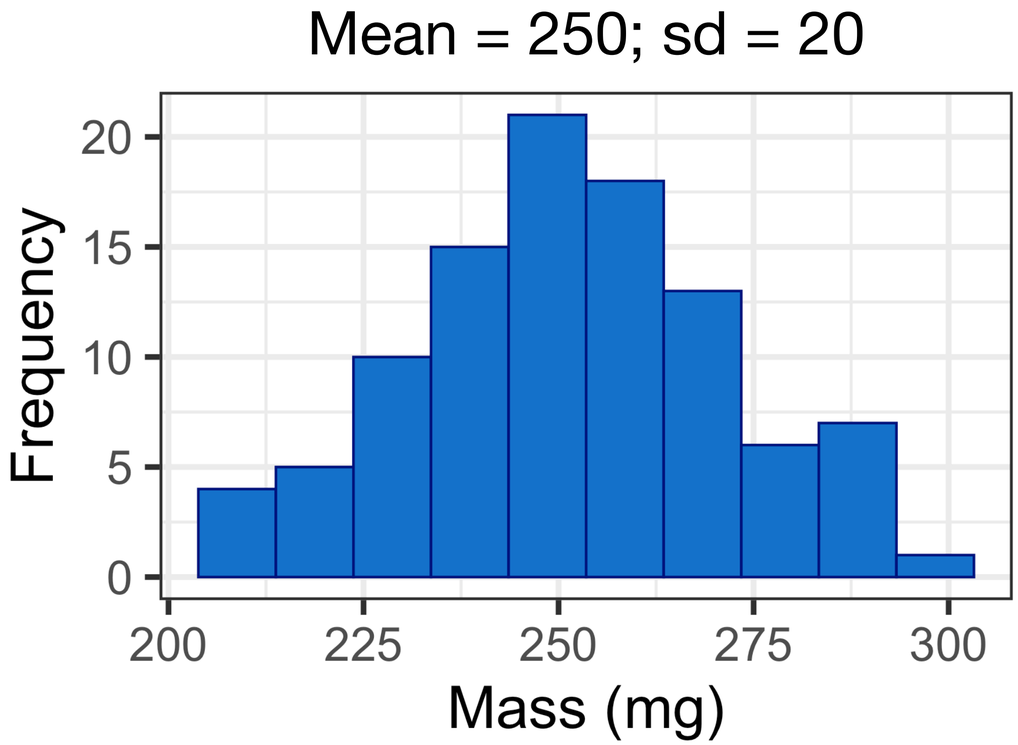




Linear transformation of variables

Suppose your scale is off by 50 mg…

```
new_data = old_data + 50
```

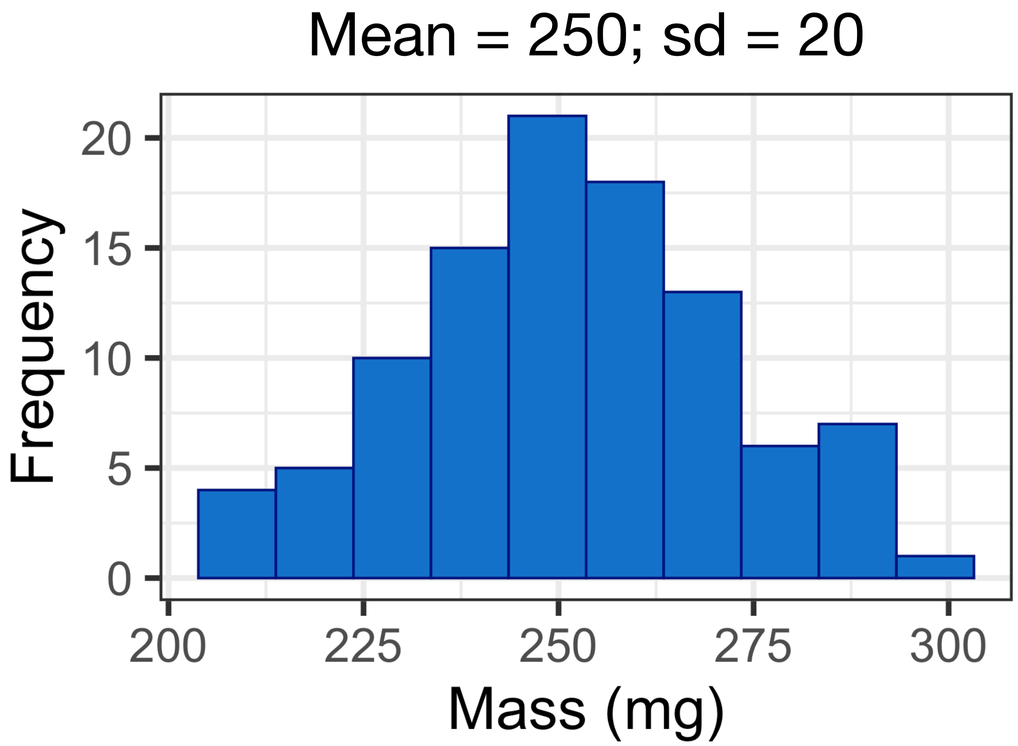


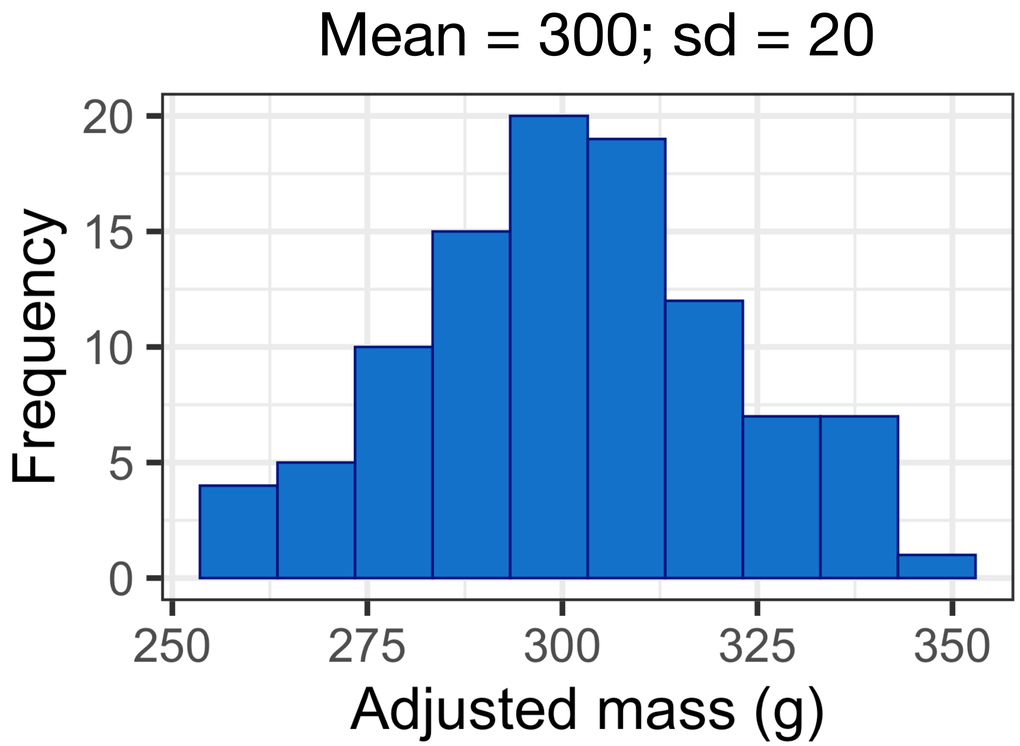
Linear transformation of variables

Suppose your scale is off by 50 mg…

```
new_data = old_data + 50
```

Mean scales, SD not

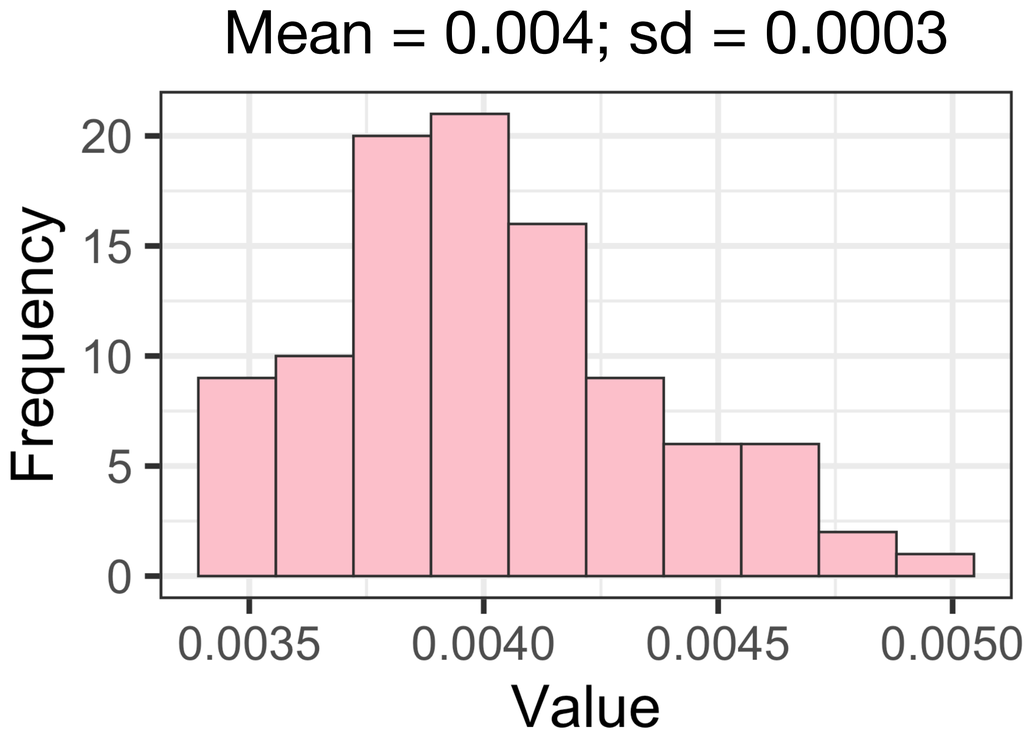




Non-linear transformation of variables

Suppose your data has a strong right skew…

```
new_data = old_data + 50
```



Non-linear transformation of variables

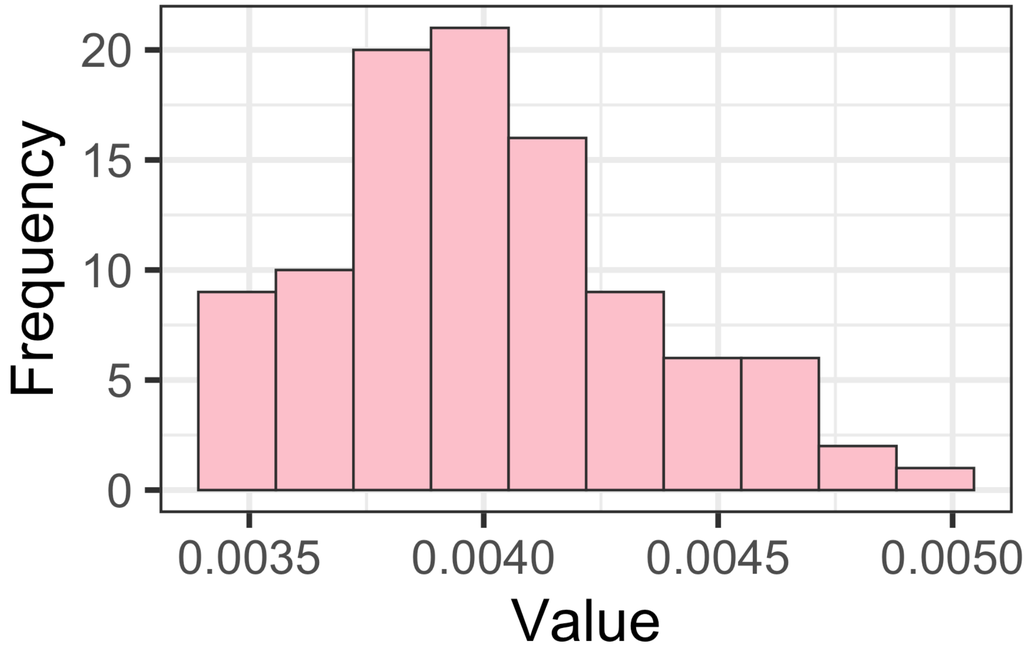
- $Y' = \sqrt{Y}$
- $Y' = \log(Y)$
- $Y' = 1/Y$
- $Y' = Y^2$

Suppose your data has a strong right skew…

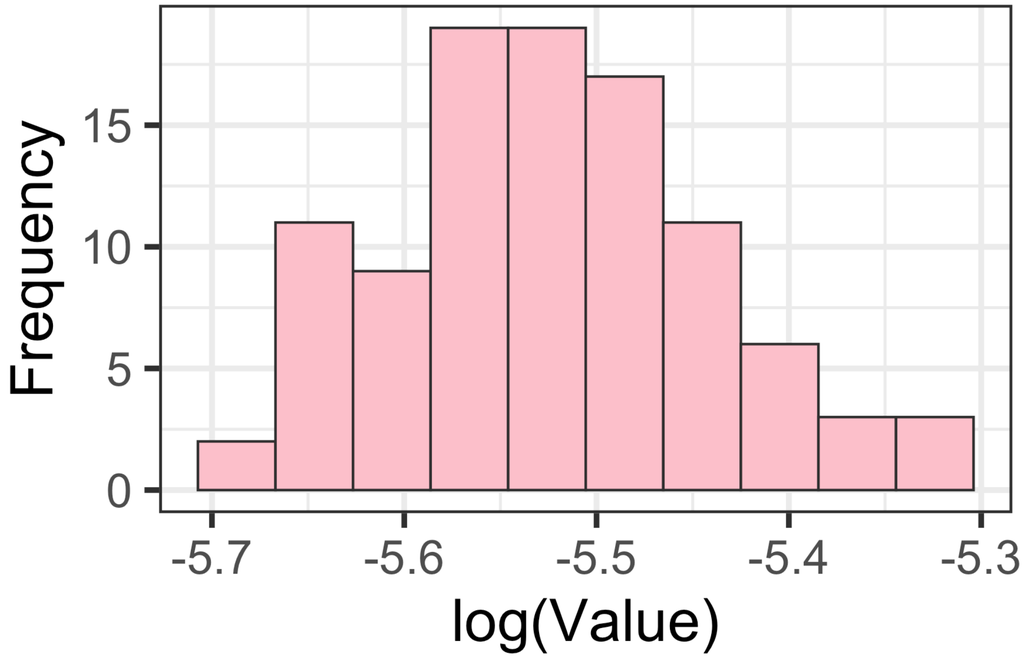
```
new_data = old_data + 50
```

Mean and SD scale non-linearly

Mean = 0.004; sd = 0.0003



Mean = -5.5; sd = 0.08



Data Visualization with ggplot2 : : CHEAT SHEET

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and geoms—visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.

Complete the template below to build a graph.

```
ggplot (data = <DATA>) +                                   required
<GEOM_FUNCTION> (mapping = aes( <MAPPINGS> ),
 stat = <STAT> , position = <POSITION> ) +                 Not required, sensible defaults supplied
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

ggplot(data = mpg, **aes**(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

aesthetic mappings | data | geom

qplot(x = cty, y = hwy, data = mpg, geom = "point") Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend=long+1,curvature=z)) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend="butt", linejoin="round", linemitre=1)
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group))
x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept=0, slope=1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend=lat+1, xend=long+1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian")
x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot()
x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

d <- ggplot(mpg, aes(fl))

d + geom_bar()
x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2)
x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size

e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , continuous y

f <- ggplot(mpg, aes(class, hwy))

f + geom_col(), x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir = "center"), x, y, alpha, color, fill, group

f + geom_violin(scale = "area"), x, y, alpha, color, fill, group, linetype, size, weight

discrete x , discrete y

g <- ggplot(diamonds, aes(cut, color))

g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

i <- ggplot(economics, aes(date, unemploy))

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))

j + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar(), x, ymax, ymin, alpha, color, group, linetype, size, width (also **geom_errorbarh()**)

j + geom_linerange()
x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))

k + geom_map(aes(map_id = state), map = map) **+ expand_limits**(x = map$long, y = map$lat), map_id, alpha, color, fill, linetype, size

THREE VARIABLES

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))l <- ggplot(seals, aes(long, lat))

l + geom_contour(aes(z = z))
x, y, z, alpha, colour, group, linetype, size, weight

l + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE)
x, y, alpha, fill

l + geom_tile(aes(fill = z)), x, y, alpha, color, fill, linetype, size, width

Ice cream analysis

The data were collected on 200 high school students and are scores on various tests, including a video game and a puzzle. The outcome measure in this analysis is the student's favorite flavor of ice cream – vanilla, chocolate or strawberry- from which we are going to see what relationships exists with video game scores (**video**), puzzle scores (**puzzle**) and gender (**female**).

Variable name	Variable	Data type
Id	Identity of the student	Nominal
female	Gender (0: Male, 1:Female)	Binary
ice_cream	Favorite Flavor (1: Vanilla, 2: Chocolate, 3: Strawberry)	Nominal
video	Score on the video game	Scale/Continuous
puzzle	Score on the puzzle	Scale/Continuous