Summarizing data

EDS 222

Tamma Carleton
Fall 2021

Today

Types of variables

- Categorical, numerical, ordinal, …

Probability density functions

- Definitions, the normal pdf, skew

Summary statistics

- Central tendency and spread, quantiles, outliers

Law of large numbers

- How big does my sample need to be?

Relationships between variables

Assignment #1 check-in: How's it going?

Reminder: OH today, 3418 Bren Hall, 3-4pm

Types of variables

Types of variables

Numerical variables

> Object class `numeric` in `R`
>
> - Can take on a wide range of possible values
> - Makes sense to add, subtract, multiply, etc.

- Examples:
    - Height of the tree canopy across the Amazon
    - Length of Atlantic swordfish
    - Daily average temperature

Discrete numerical variables take on only a limited set of values, often counts (e.g., population)

Continuous numerical variables: can take on infinite values within a range (e.g., arsenic concentration in groundwater)

Types of variables

Numerical variables

Source: Allison Horst

Types of variables

Categorical variables

> Object class `factor` in `R`

- Values correspond to one of a fixed number of categories

- Possible values are called **levels**

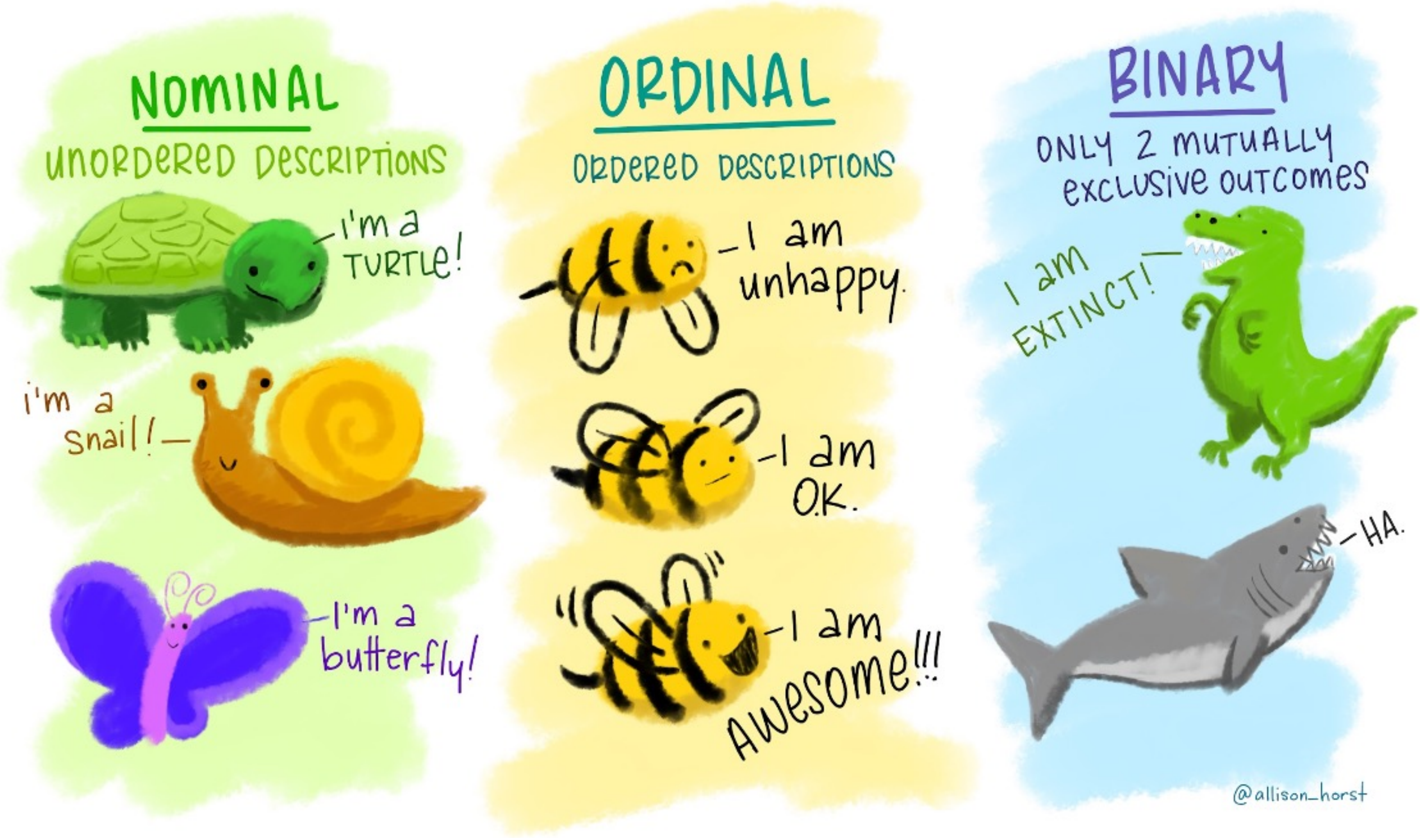
- Examples:

    - States of the US
    - Species of tree
    - Age group (e.g., <15, 15-64, 65+) (watch out! continuous numerical data can often be stored as a categorical variable!)

Nominal variables are unordered descriptions

Types of variables

Categorical variables



Source: Allison Horst

Probability density functions

Probability density functions

For *continuous* variables, the **probability density function (p.d.f.)** tells us the probability that a random variable falls within a given range of values.

Formally: The **p.d.f.** of a continuous variable X with support (i.e., range of possible values) S is an integrable function $f(x)$ satisfying:

1. $f(x)$ is positive for all x in S
2. The area under the curve $f(x)$ over the entire support S is equal to 1:

$$\int_S f(x)dx = 1$$

3. The probability that x falls between A and B is:

$$Pr(A \leq x \leq B) = \int_A^B f(x)dx$$

Why isn't this simpler?

> Q: Why can't I just interpret $f(x)$ as the probability that $X = x$?
>
> A: Because continuous variables have ∞ possible values...the probability that your variable X exactly equals x is zero!

Luckily, for **discrete variables** it *is* this simple!

For *discrete* variable x , the **probability mass function (p.m.f.)** $f(x)$ tells us the probability that $X = x$.

Formally: The **p.m.f.** of a discrete variable X with support (i.e., range of possible values) S is a function $f(x)$ satisfying:

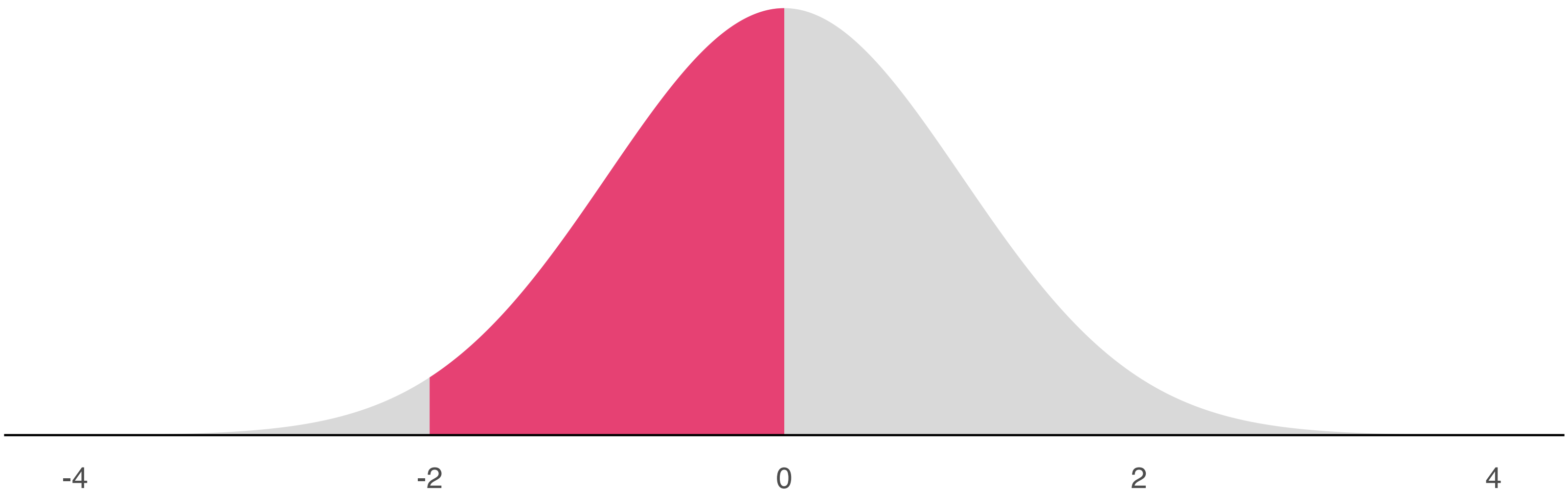
1. $P(X = x) = f(x) > 0$ for all x in support S
2. $\sum_{x \in S} f(x) = 1$
3. $P(A \leq x \leq B) = \sum_{x=A}^{x=B} f(x)$

Probability density functions (visual)

P.d.f.'s help us characterize the distribution of our population. The most common/famous ones get names (e.g., normal, Gamma, t,...)

Let's look at a **normal** distribution*

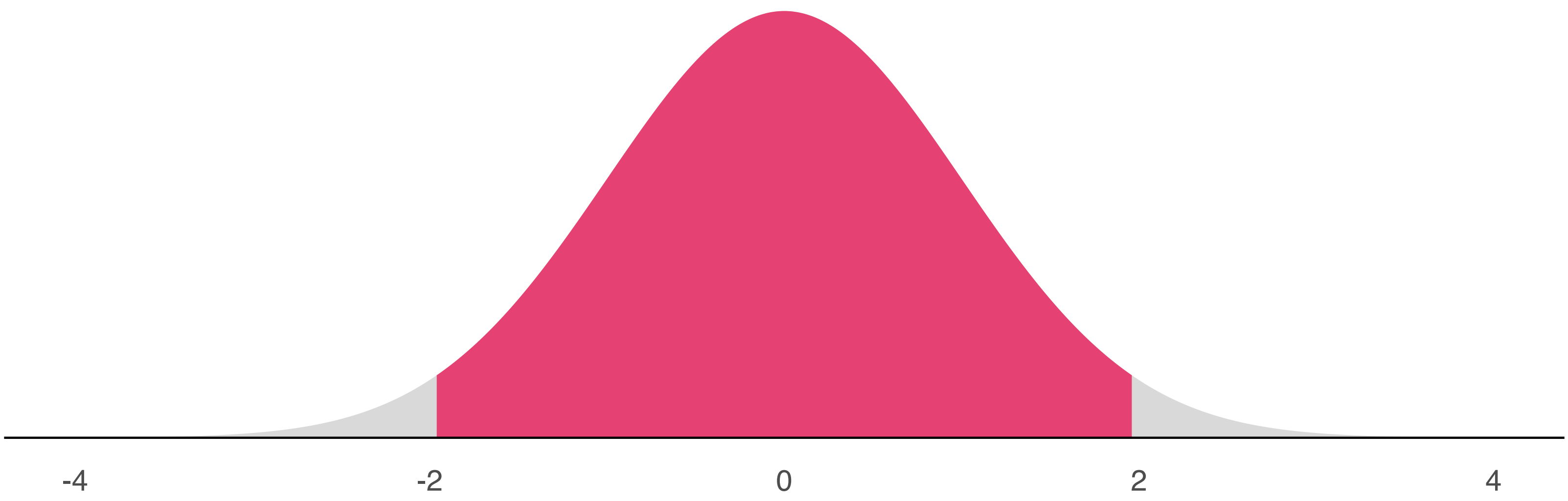
The probability this normally distributed variable takes on a value between -2 and 0 is shown in pink:



Probability density functions (visual)

Let's look at a **normal** distribution*

The probability this normally distributed variable takes on a value between -2 and 2 is shown in pink:



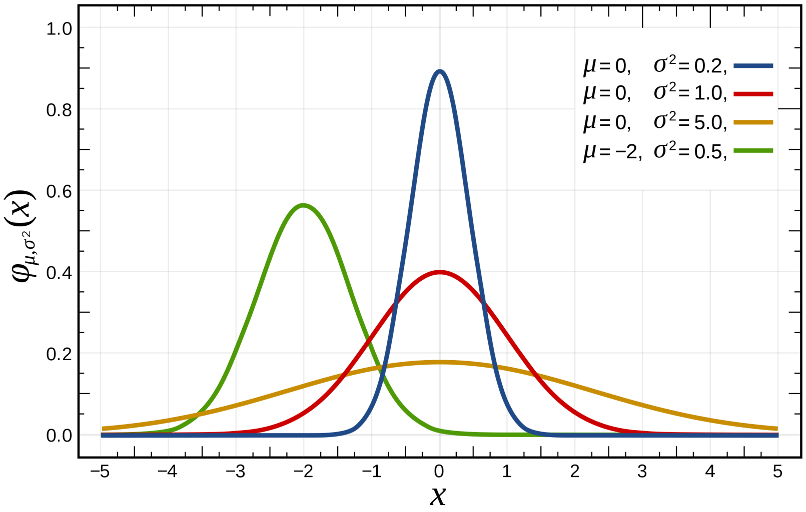
*Yep, still a "standard" normal. Details later.

The normal distribution

There are infinite different normal distributions. They all have the following p.d.f.:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

where μ is the mean (i.e., average) and σ is the standard deviation (will define soon).



Shapes of probability distributions

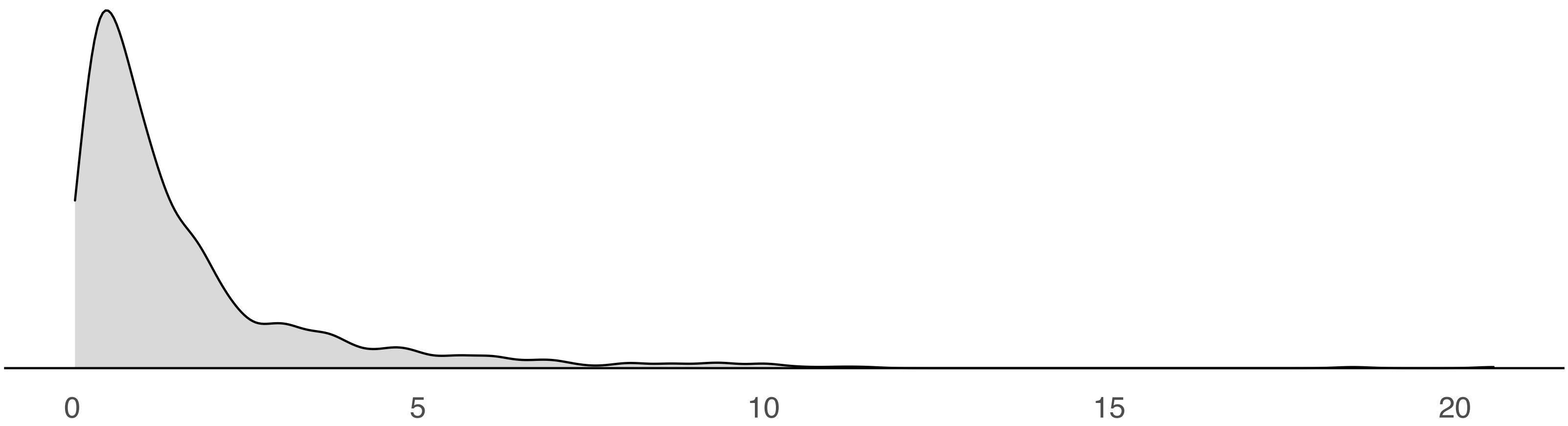
Key terms to describe p.d.f.'s:

1. A distribution can have **skew** (e.g., log-normal)
2. A distribution can have a long **right tail** or **left tail** (e.g., fat-tailed climate sensitivity distributions!)
3. A distribution can be **symmetric**
4. A distribution can be **unimodal**, **bimodal**, or **multimodal**

Shapes of probability distributions

Skew with a long right tail

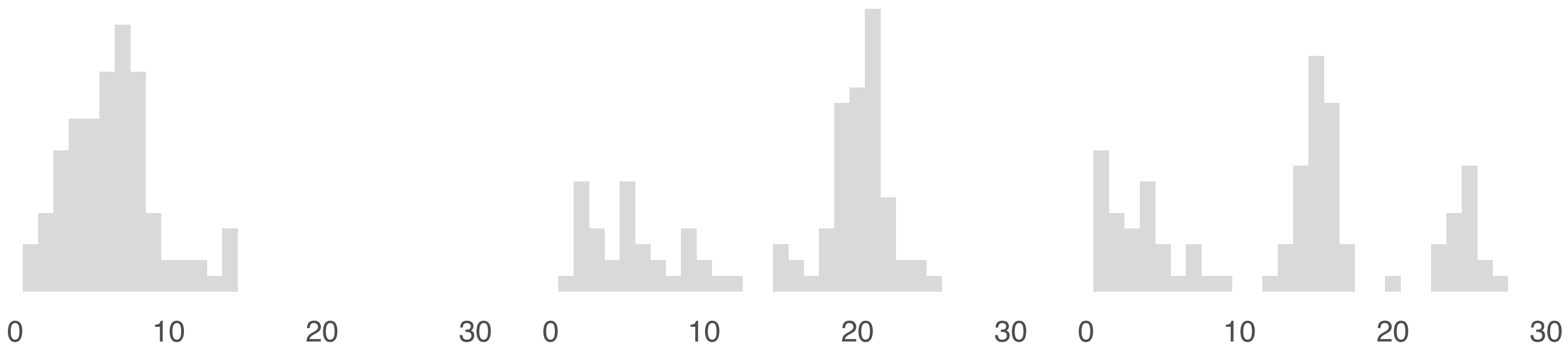
(log-normal sample distribution)



Shapes of probability distributions

Uni-, bi-, and multi-modal

(How many "peaks" do you see?)



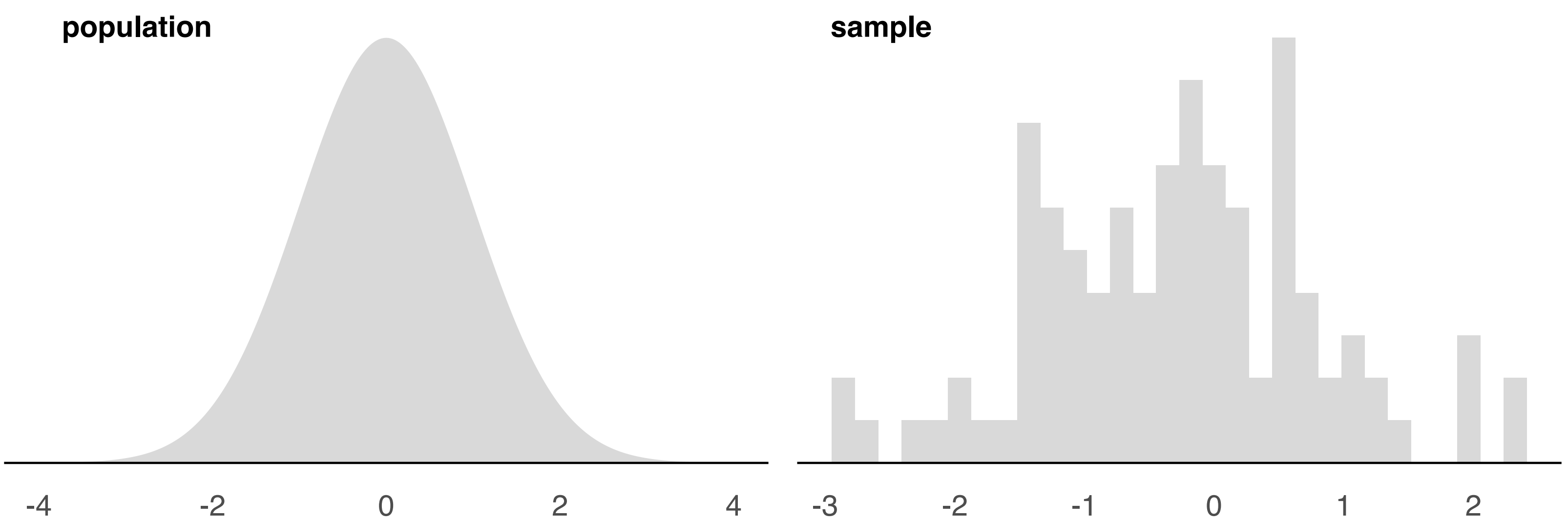
Summary statistics

Describing random variables

A probability density function describes a **population**

As we learned last week, we rarely have a **census** so we rarely can directly describe the p.d.f. itself.

Instead, we use *statistics* from a *sample* to estimate *parameters* of the *population*



Measures of central tendency

We often begin to describe a distribution using measures of **central tendency** (i.e., measures of the "middle").

Three are most common:

1. **Mean**
2. **Median**
3. **Mode**

Mean = expected value = average

In a **population**, the mean is defined as:

$$\mathbf{E}[X] = \mu = \int_x x f(x) dx$$

In our **sample**, we compute the mean as:

$$\bar{x} = \frac{1}{n} \sum_{i \in n} x_i$$

We use $\bar{x}$ as an *estimate* of the parameter of interest, μ.

Median = middle value = 50th percentile

In a **population**, the median m is defined as:

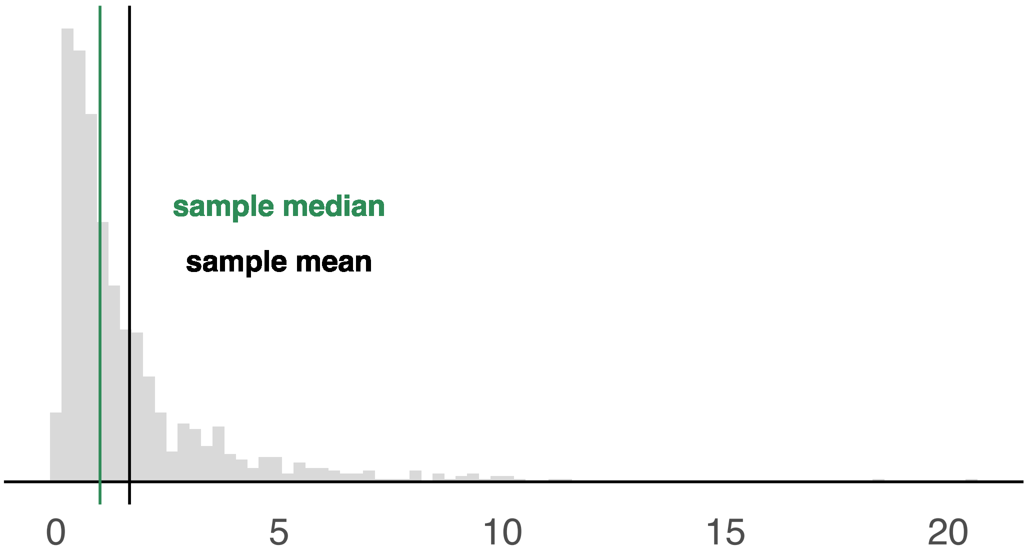
$$P(X \leq m) = \int_{-\infty}^{m} f(x)dx = \frac{1}{2} = \int_{m}^{\infty} f(x)dx = P(X \geq m)$$

In our **sample**, we compute the median as:

- n even? median = mean of the middle two values
- n odd? median = middle value

Median and mean are not always close

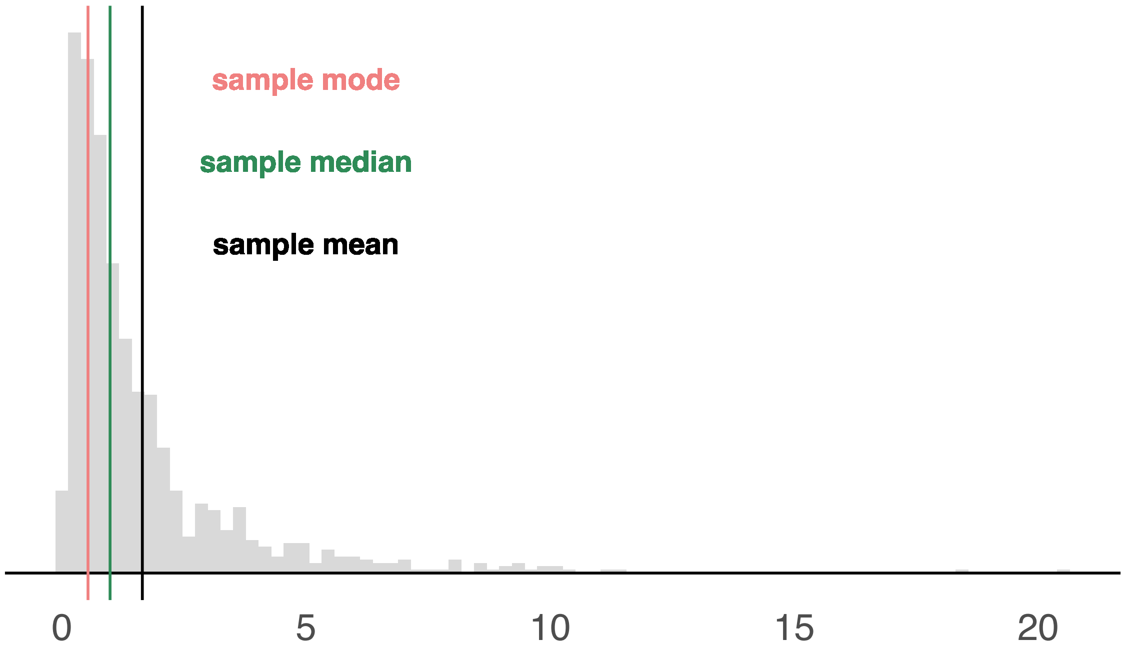
Non-normal distribution $\implies$ median and mean can diverge substantially



Mode = most frequent value

The **mode** is simply the most frequently observed value

This is much more useful for discrete data (ask yourself why!)



Measures of spread

Central tendency only gets us so far...we also need measures of **spread**.

1. **Range** (easy: min to max of your data)
2. **Variance**
3. **Standard deviation**
4. **Quantiles**

Measures of spread: Variance

Answers the question, how far are observations from the mean, on average?

In the population:

$$Var(X) = \mathrm{E}[(X-\mu)^2] = \sigma^2 = \int_{\mathrm{S}} (x-\mu)^2 f(x) dx$$

In the sample:

$$s^2 = \frac{\sum_{i \in n} (x_i - \bar{x})^2}{n-1}$$

> Q: Why do we divide by $n-1$?
>
> A: Lots of math to prove it (see here), but trust me that your calculation of s^2 will be a biased estimate of σ^2 if you do not divide by $n-1$!

Measures of spread: Standard deviation

Just the square root of the variance!

In the population:

$$SD(X) = \sqrt{\mathrm{E}[(X-\mu)^2]} = \sigma = \sqrt{\int_{\mathrm{R}} (x-\mu)^2 f(x) dx}$$

In the sample:

$$s = \sqrt{\frac{1}{n-1} \sum_{i \in n} (x_i - \bar{x})^2}$$

Units of standard deviation: units of the random variable

Some helpful rules

$$E[aX + b] = aE[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

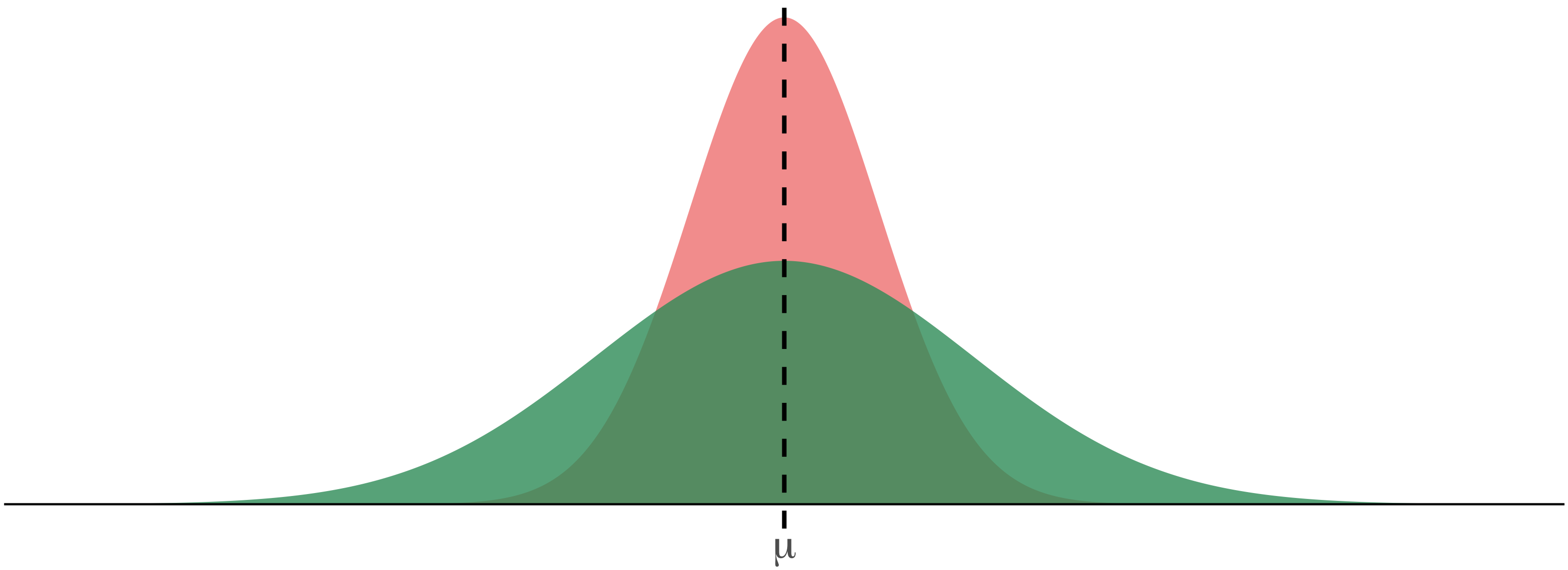
$$var(X) = E[X^2] - (E[X])^2$$

$$var(aX + b) = a^2 var(X)$$

Variance, visually

Pink: Low variance/standard deviation $\sigma = 1$

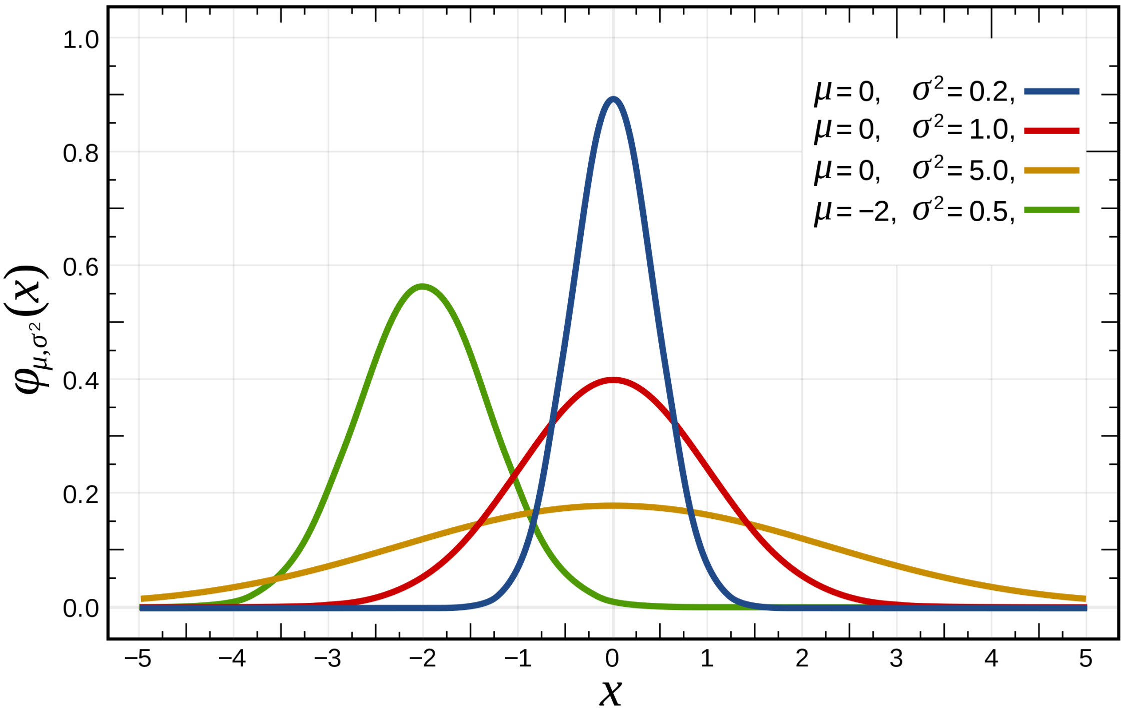
Green: High variance/standard deviation $\sigma = 2$



Variance, visually

Back to the normal distributions

- Changes in the *mean* shift the distribution right to left
- Changes in the *standard deviation* stretch the distribution out (or shrink it in)



Measures of spread: Quantiles

Quantiles are cut points of a probability distribution

In our sample, quantiles are cut points of our sample data

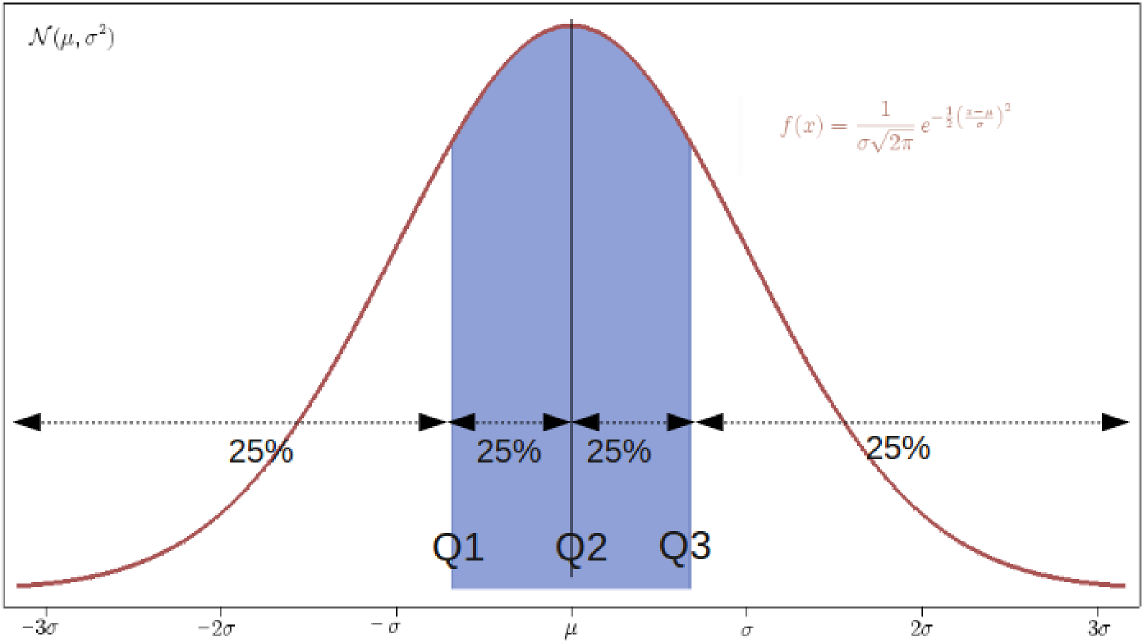
How do we compute them?

- We order our data from lowest to highest
- For the q-quantile, we divide these ordered data into q equal sized subsamples
- The value at the edge of the kth subsample is the kth q-quantile
  - This tells you the value below which $\frac{k}{q}$ of the data lie

> Question: How many q-quantiles are there for any given q?
>
> Answer: There are $q-1$ of the q-quantiles

Example: The normal distribution

Common quantiles have names you have head of, such as *quartiles* for $q = 4$:



Quartiles of the normal distribution

Common quantiles and interpretation

Common quantiles have names you have heard of:

- $q = 2$ **Median** tells us the value for which 50% of our sample sits *below* (and 50% above). This is quantile 0.5 (or 50% quantile)
- $q = 3$ **Terciles**: tell us the values for which 33.33% (1st tercile) and 66.66% (2nd tercile) of our sample sits *below*
- $q = 4$ **Quartiles**: tell us the values for which 25% (1st quartile), 50% (2nd quartile), and 75% (3rd quartile) of our sample sits *below*
- $q = 10$ **Deciles**: tell us the values for which 10% (1st decile), ..., 50% (5th decile), ..., and 90% (9th decile) of our sample sits *below*

q The k*th* q-quantile tells us the value for which $\frac{k}{q} \times 100\%$ of our sample sits *below*

This sounds a lot like percentiles...

Percentiles are simply quantiles for q=100!

We hear about percentiles in daily life more often, and in practice people often use "percentiles" language for the more general term "quantiles".

Examples of percentiles:

- At 5'3", my height is the 40th percentile of the U.S. adult female height distribution $\rightarrow$ 40% of American female adults are shorter than me
- At 24.5 lbs, my son's weight is the 81st percentile of U.S. male babies of 13 months old $\rightarrow$ 81% of American male 13 month olds are lighter than my son
- Your GRE score is in the 90th percentile of tests taken this year $\rightarrow$ 90% of GRE scores this year were lower than your score

> Exercise: Draw approximately where you think the 1st, 10th, 20th, 50th, 80th, 90th and 99th percentiles would be on a normal

Quantile-Quantile (Q-Q) Plots

Histograms plot the frequency of our data within bins

- `geom_histogram()` with `ggplot2` in `R`

Q-Q plots plot the quantiles of our data *against* quantiles of some theoretical distribution

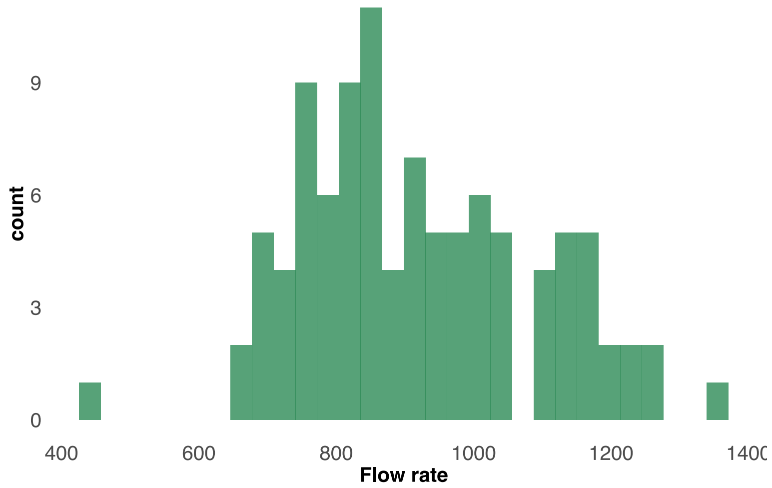
- `geom_qq()` with `ggplot2` in `R`

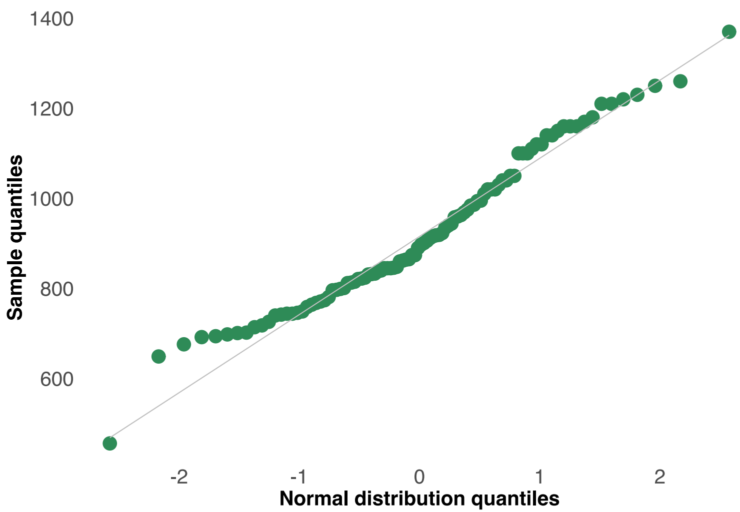
> This is helpful if we want to ask things like, are my data approximately normally distributed?

Straight line on a Q-Q plot indicates sample and theoretical distributions match

Q-Q plot: Example

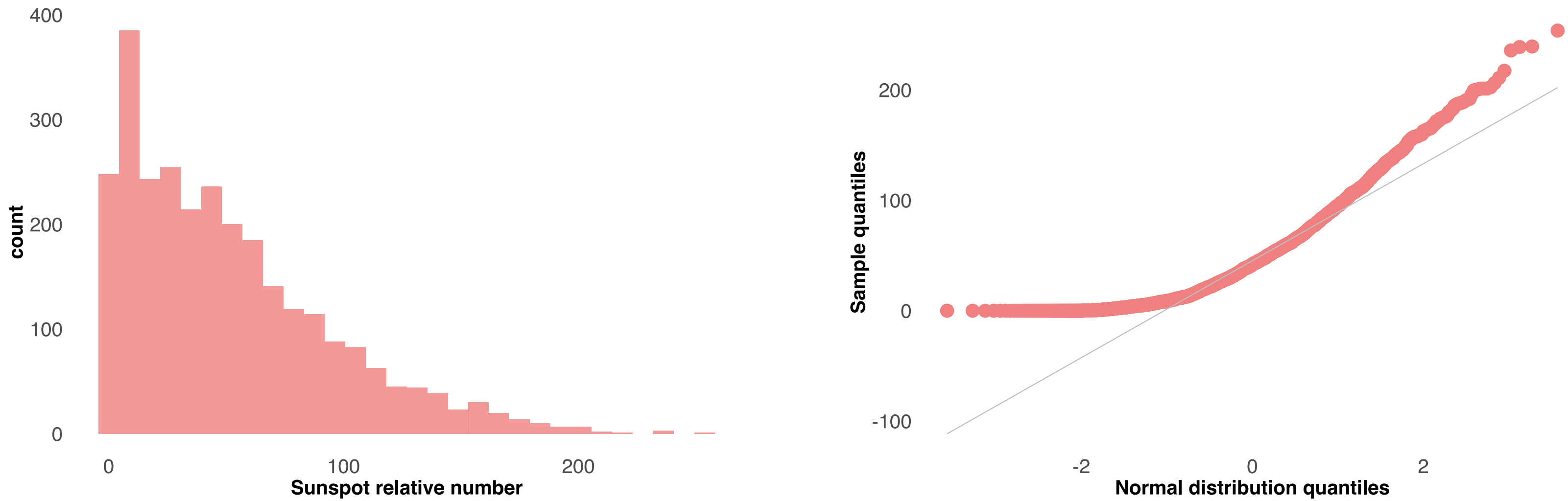
Annual flow of the river Nile at Aswan, 1871-1970, in 10^8 m^3





Q-Q plot: Example

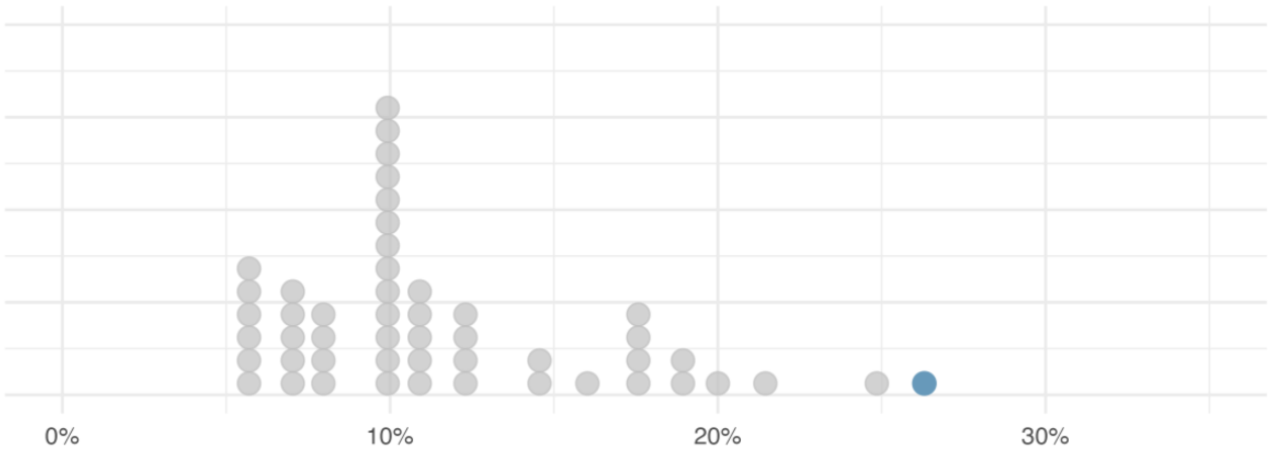
Monthly mean relative sunspot numbers, 1749-1983



> We will continually return to the normal distribution. Always a good idea to check whether your data look normally distributed or not!

Which statistics are robust to outliers?

- Consider a sample of loans from a bank, each with an associated interest rate.
  - $\bar{x} = 11.57$
  - $s = 5.05$
- The highest value in the data is somewhat of an outlier, $x_{max} = 26.3$.



Source: IMS, Ch. 5.6

Which statistics are robust to outliers?

- Consider a sample of loans from a bank, each with an associated interest rate.
  - $\bar{x} = 11.57$
  - $s = 5.05$
- The highest value in the data is somewhat of an outlier, $x_{max} = 26.3$.
- How do summary statistics change if we modify this outlier?

	Robust		Not robust	
Scenario	**Median**	**IQR**	**Mean**	**SD**
Original data	9.93	5.75	11.6	5.05
Move 26.3% to 15%	9.93	5.75	11.3	4.61
Move 26.3% to 35%	9.93	5.75	11.7	5.68

Table 5.4: A comparison of how the median, IQR, mean, and standard deviation change as the value of an extereme observation from the original interest data changes.

Law of large numbers

Big data

You probably have intuition that a larger sample is better than a smaller one...but why?

Goal: Use sample statistics to estimate population parameters.

Suppose we have a **random** sample of some size n. How well does $\bar{x}$ approximate μ?

Law of large numbers:

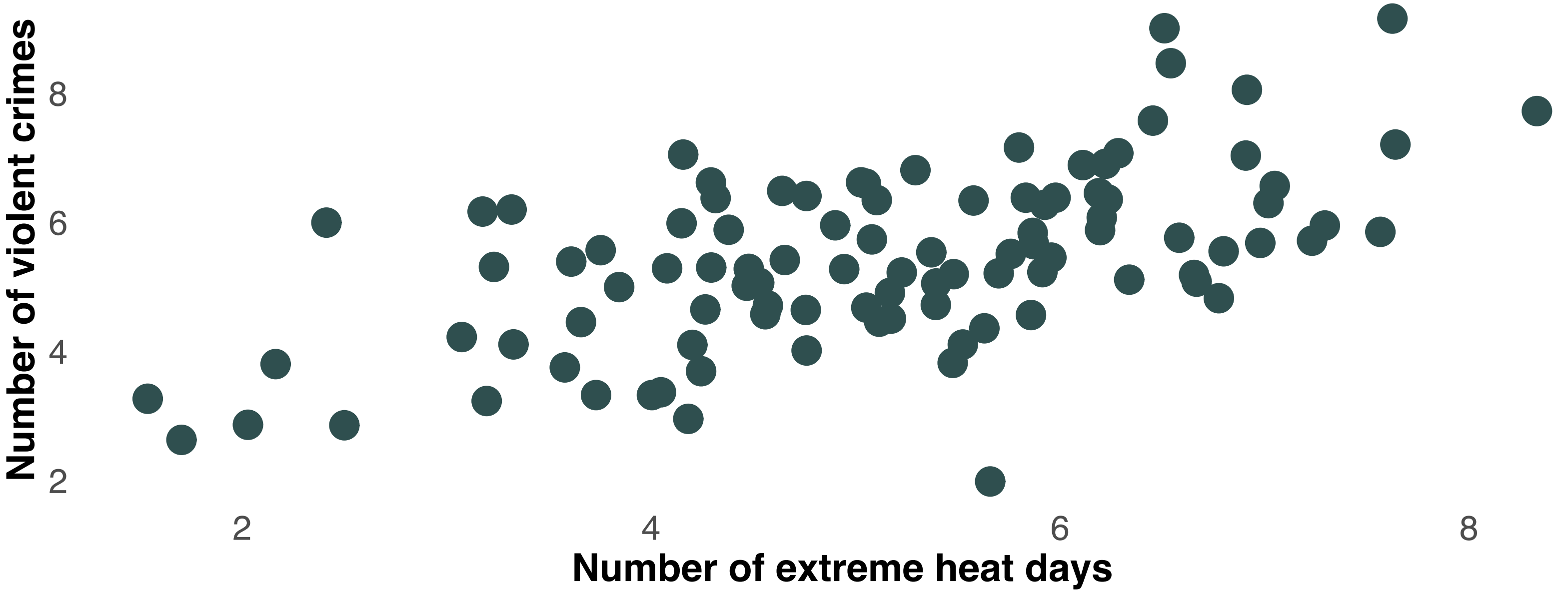
$$\bar{x} \to \mu \text{ as } n \to \infty$$

Relationships between variables

Two random variables

Often we are interested in the *relationship* between two (or more) random variables.

E.g., heat waves and heart attacks, nitrogen fertilizer and water pollution



*Note: these are simulated data. But the violence-temperature link is real! See here for a summary of research.

Two random variables

What metrics can we use to characterize the *relationship* between two variables?

There are lots. But let's start with...

1. Covariance

2. Correlation

Covariance

Variance indicates how dispersed a distribution is (average squared deviation from the mean)

Covariance is a measure of the *joint* distribution of two variables

- Higher values of X correspond to higher values of $Y \rightarrow$ **positive** covariance
- Higher values of X correspond to lower values of $Y \rightarrow$ **negative** covariance

In the population:

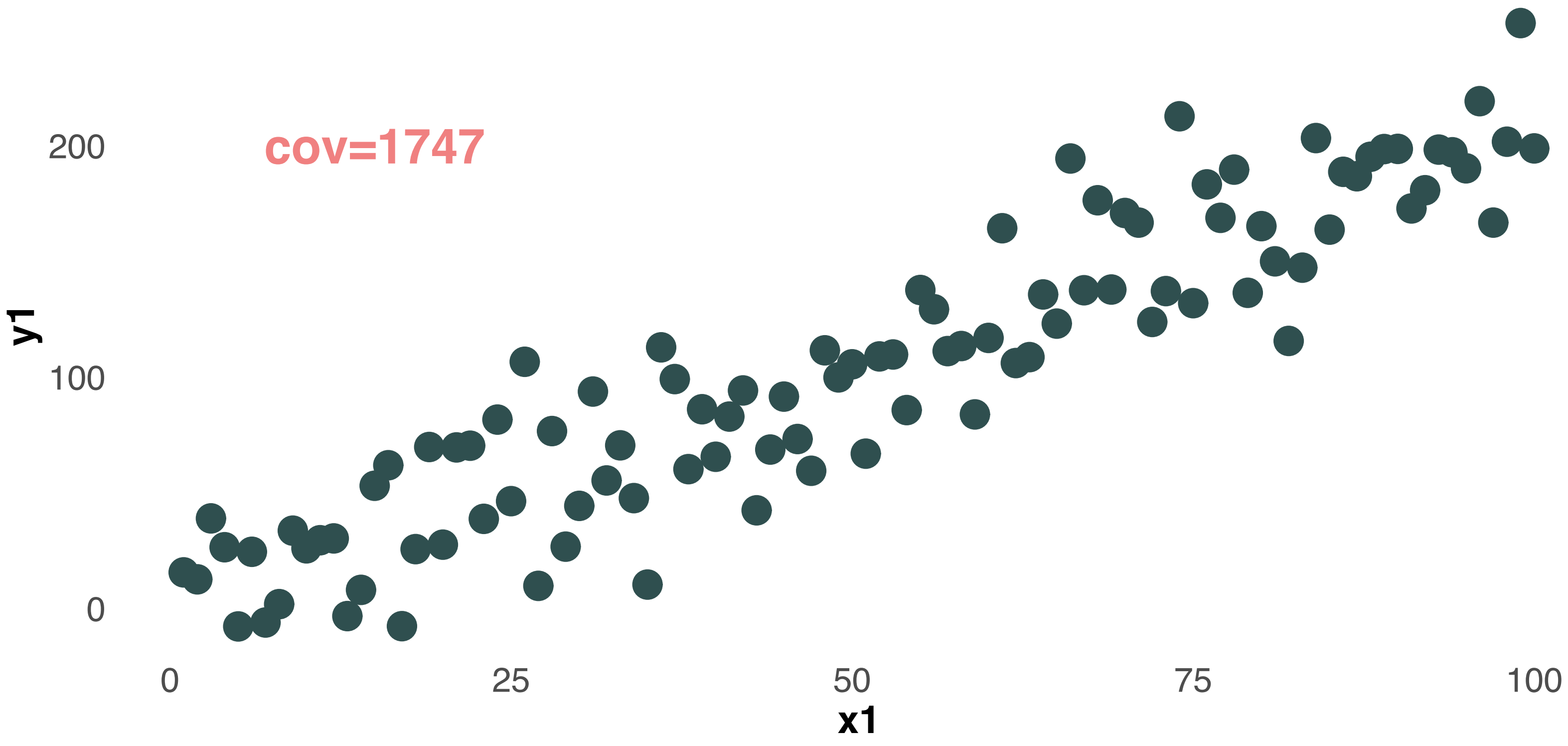
$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x \mu_y$$

In the sample:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

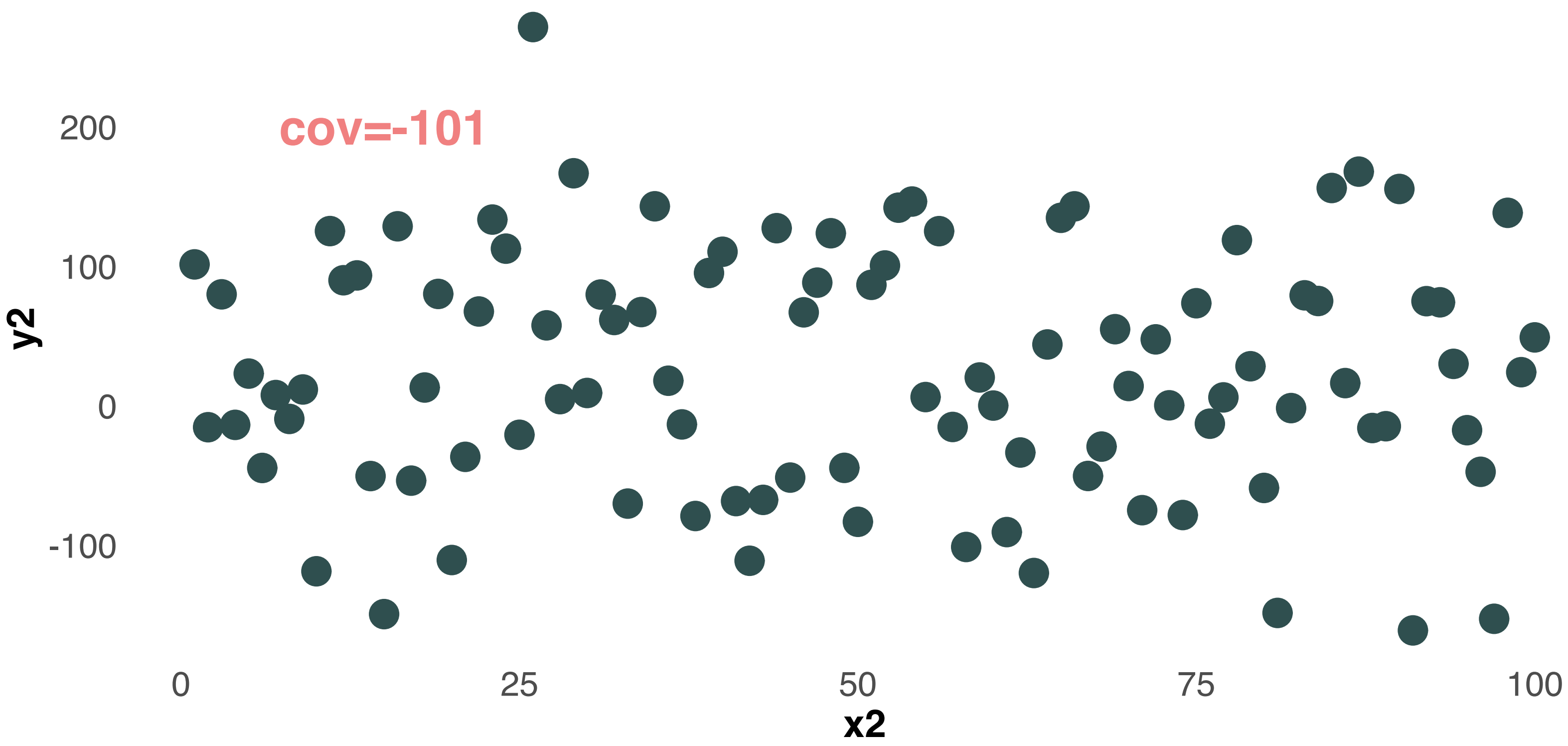
Covariance

Example: positive covariance



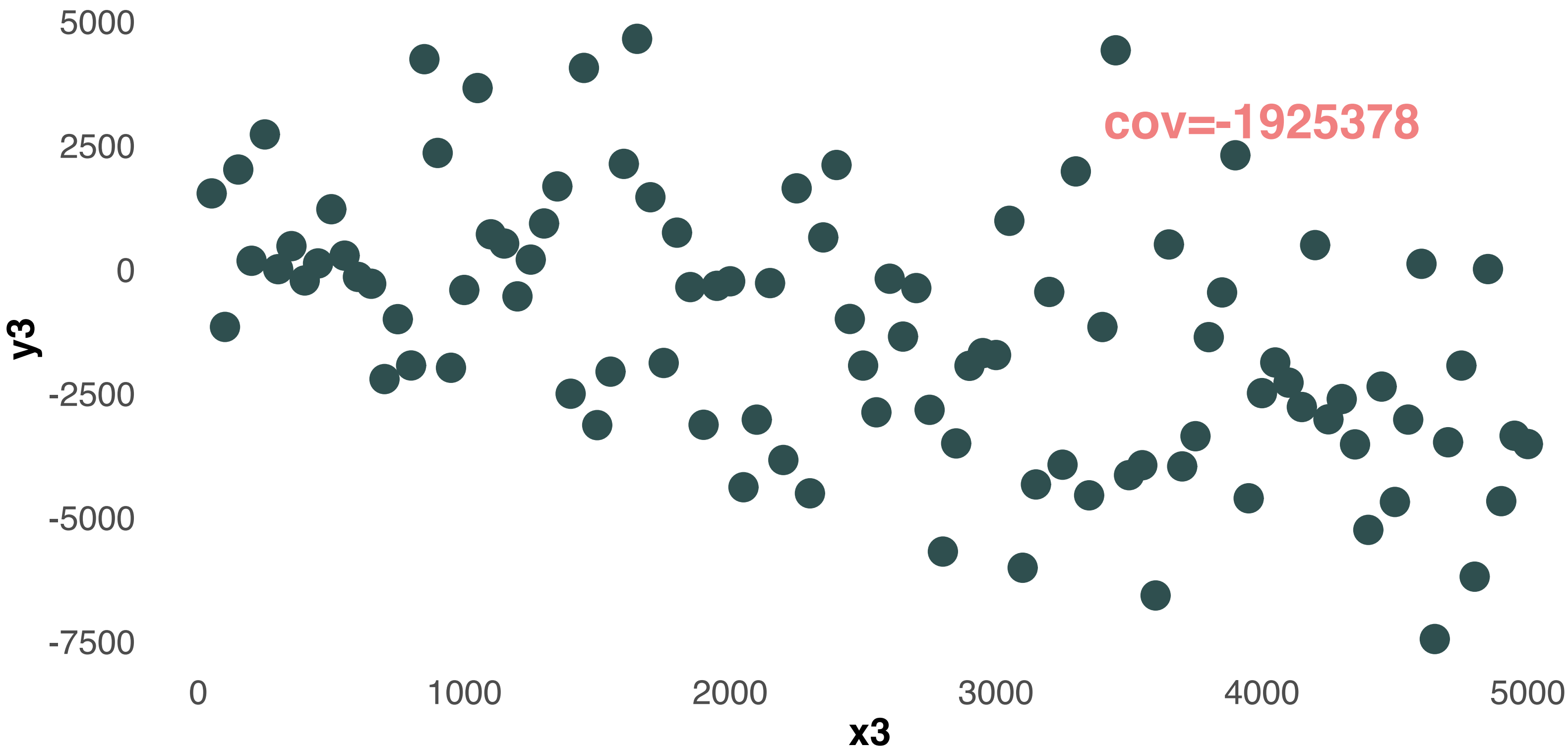
Covariance

Example: zero covariance



Covariance

Example: Negative covariance



How do I interpret these units?! Hard to compare across these three graphs...

Correlation

Correlation allows us to normalize covariance into interpretable units

The sign still tells us about the nature of the (linear) relationship between two variables:

- **positive** covariance $\rightarrow$ **positive** correlation (and vice versa)

But now, the magnitude is interpretable:

- Ranges from -1 to 1, with magnitude indicating *strength* of the relationship

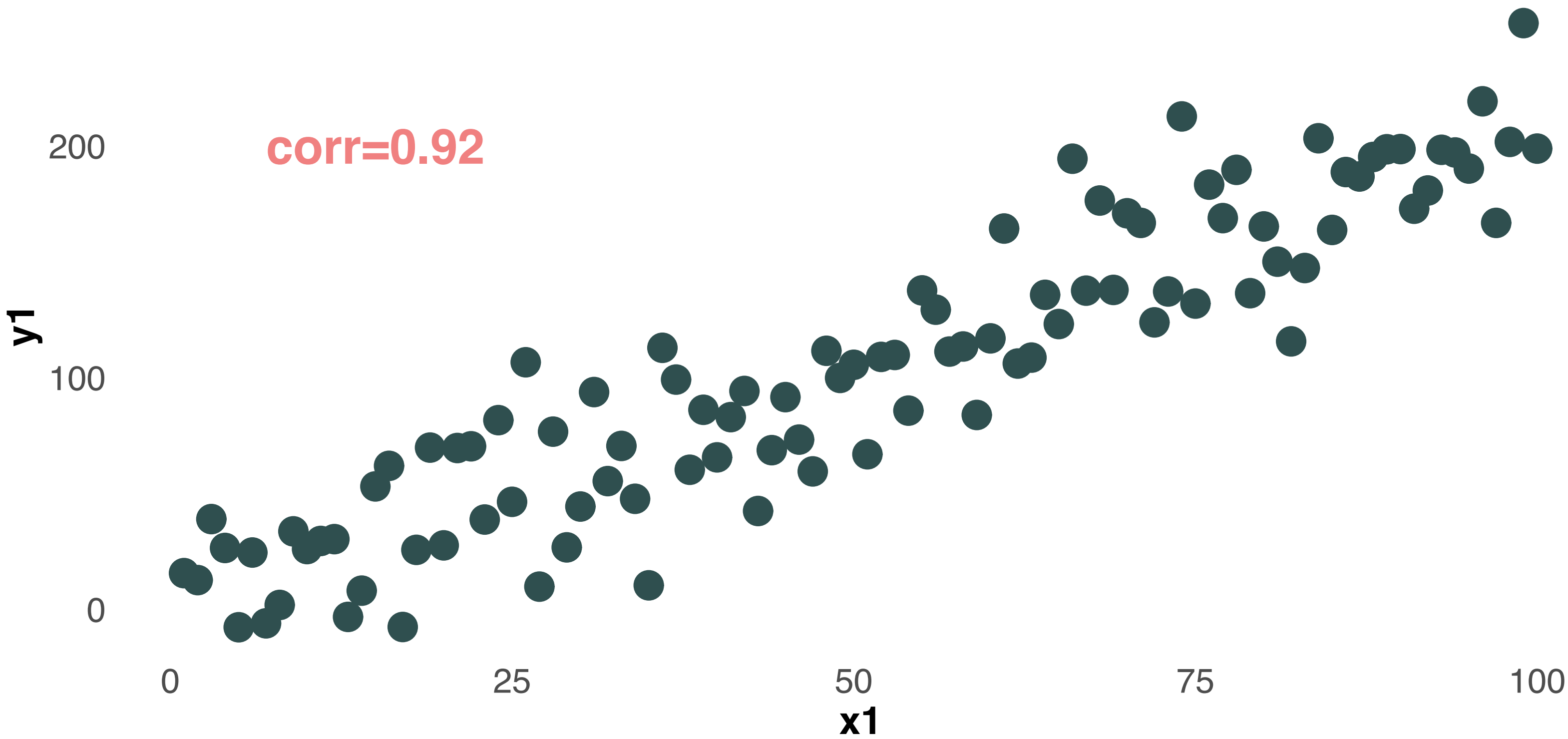
  In the population:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

  In the sample:

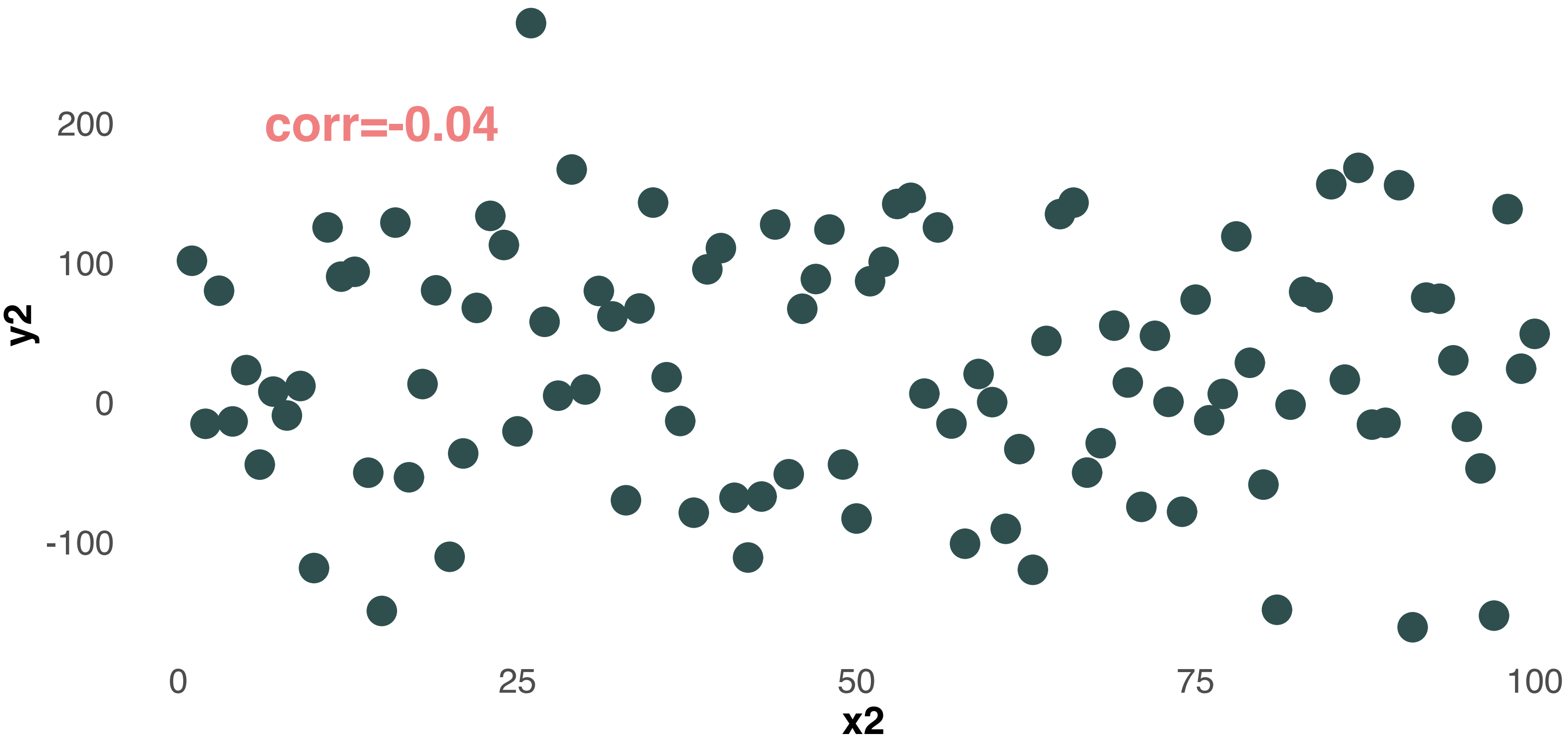
Correlation

Example: Strong positive correlation



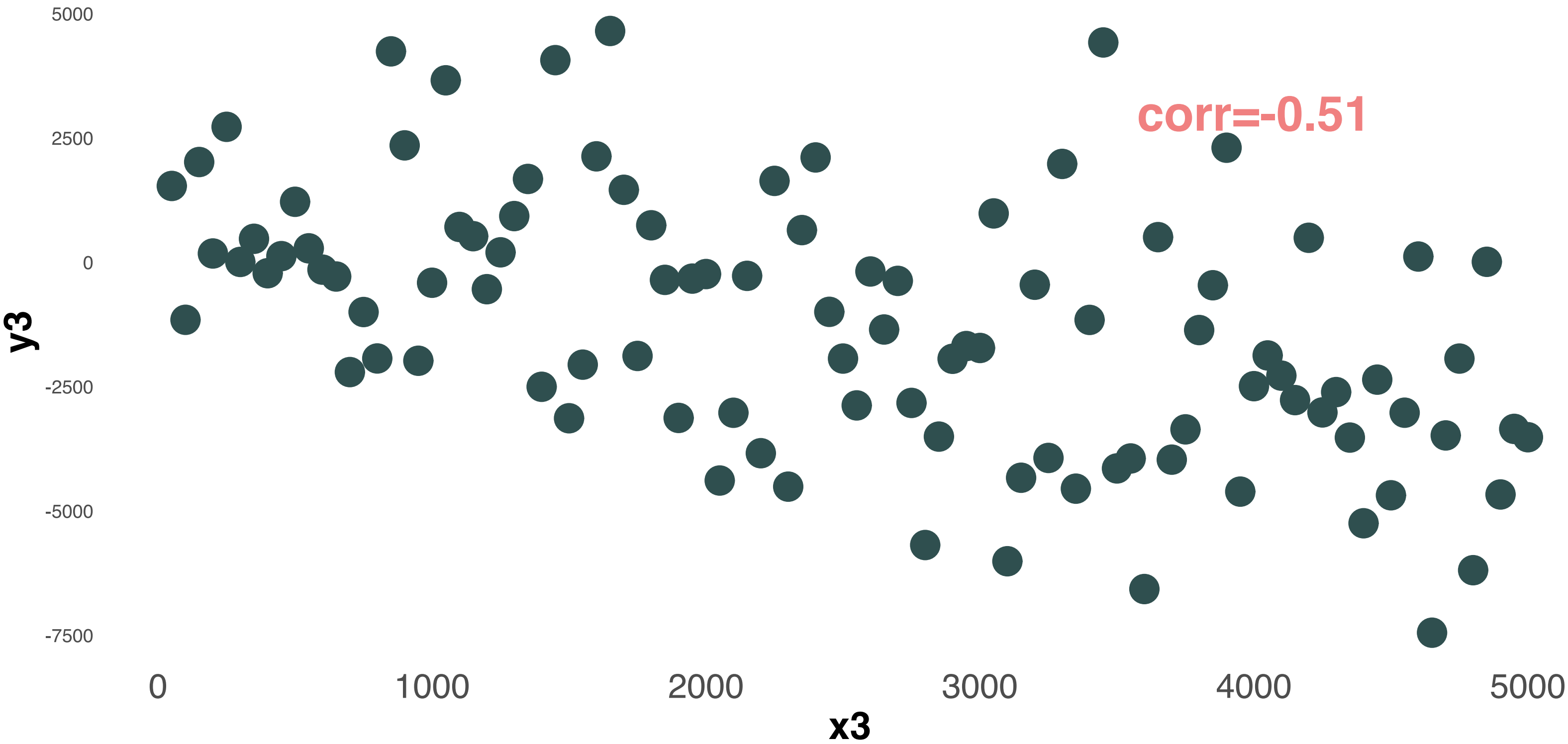
Correlation

Example: zero correlation



Correlation

Example: Moderate negative correlation



Next up

Summary statistics in `R` (Thursday lab, Assignment 02)

Linear regression (Next Tuesday)

Slides created via the R package **xaringan**.

Some slide components were borrowed from Ed Rubin's awesome course materials.