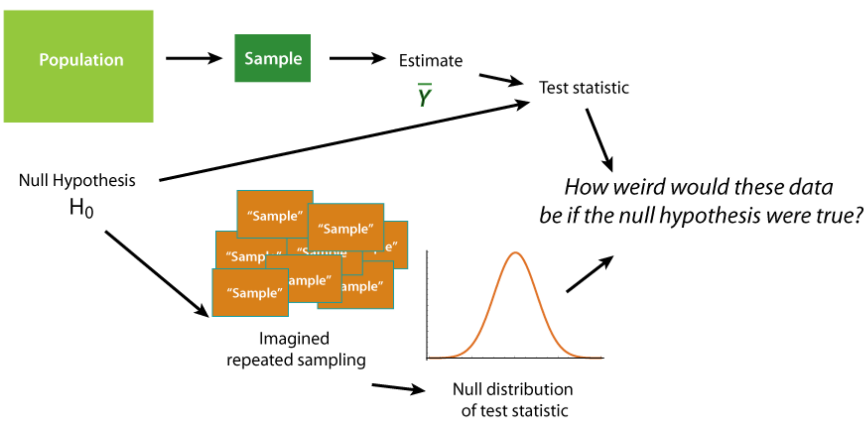
Hypothesis testing

Chapter 6

Hypothesis testing asks how unusual it is to get data that differ from the null hypothesis.

If the data would be quite unlikely under H_0, we reject H_0.



Hypotheses are about populations, but are tested with data from samples

Hypothesis testing usually assumes that sampling is random.

Null hypothesis: a specific statement about a population parameter made for the purposes of argument.

Alternate hypothesis: represents all other possible parameter values except that stated in the null hypothesis.

The *null hypothesis* is usually the simplest statement, whereas the *alternative hypothesis* is usually the statement of greatest interest.

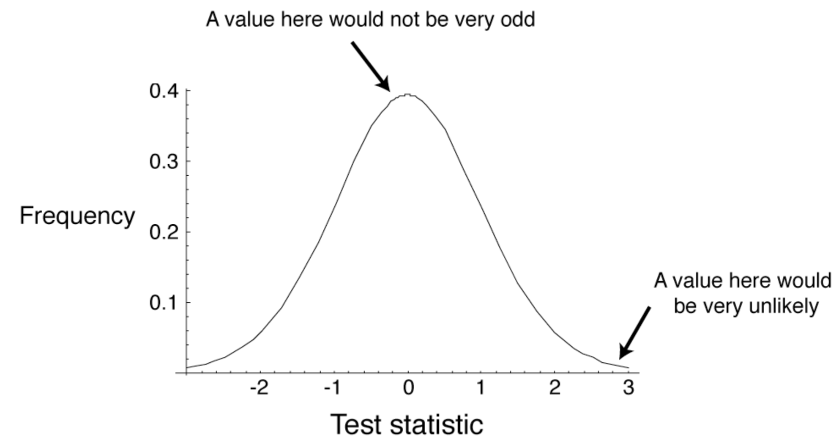
A good null hypothesis would be interesting if proven wrong.

A null hypothesis is specific;
an alternate hypothesis is not.

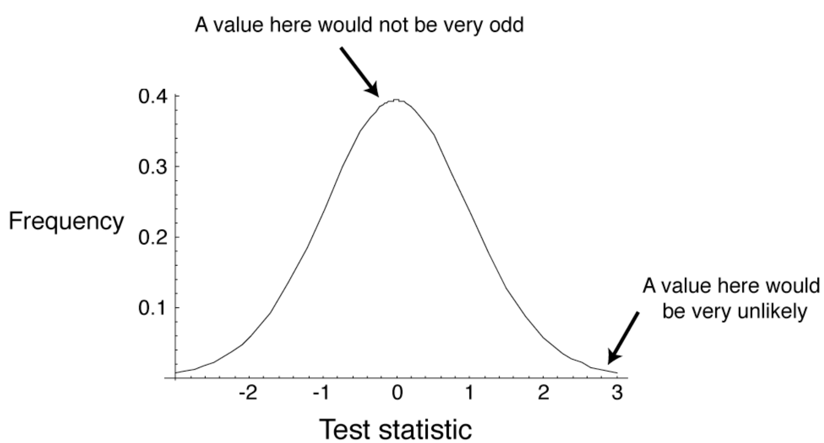
Test Statistic

A number calculated to represent the match between a set of data and the null hypothesis

Can be compared to a general distribution to infer probability

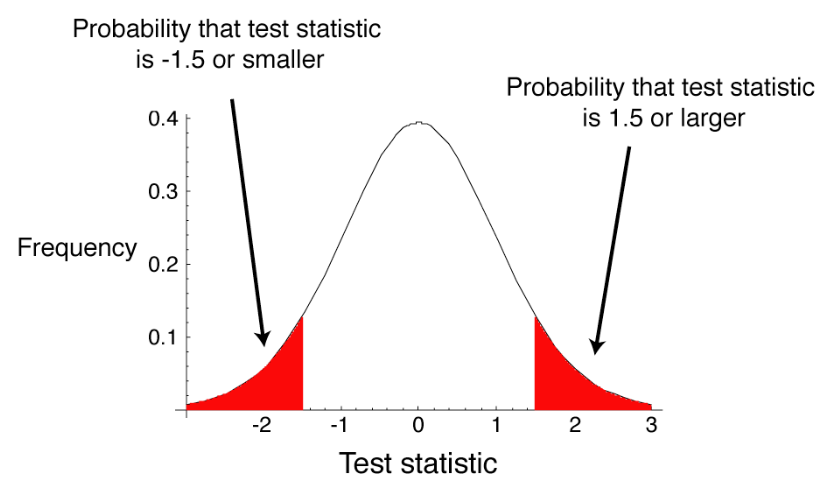


Possible outcomes from samples under null hypothesis



A test statistic summarizes the match between the data and the null hypothesis

P-value



A *P*-value is the probability of getting the data, or something as or more unusual, if the null hypothesis were true.

How to find *P*-values

Simulation

Parametric tests

Permutation

Hypothesis testing: an example

Does a red shirt help win wrestling?

The experiment and the results

Animals use red as a sign of aggression

Does red influence the outcome of wrestling, taekwondo, and boxing?

- 16 of 20 rounds had more red-shirted than blue-shirted winners in these sports in the 2004 Olympics
- Shirt color was randomly assigned

Hill, RA, and RA Burton 2005. Red enhances human performance in contests Nature 435:293.

Stating the hypotheses

H_0: Red- and blue-shirted athletes are equally likely to win (*proportion* = 0.5).

H_A: Red- and blue-shirted athletes are not equally likely to win (*proportion* ≠ 0.5).

Estimating the value

16 of 20 is a proportion of *proportion* = 0.8

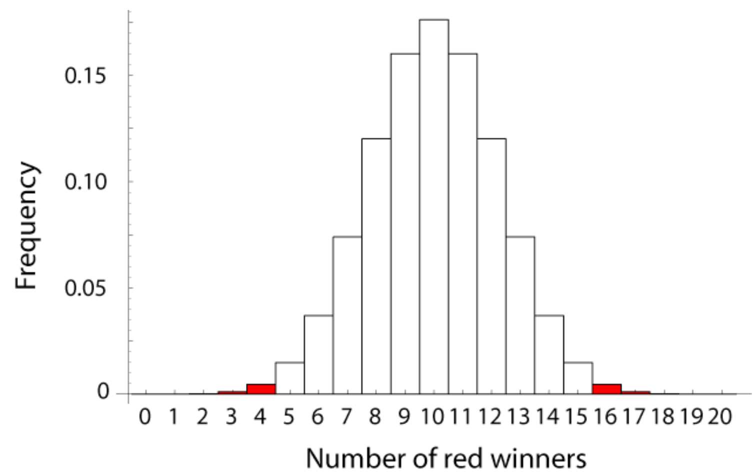
This is a discrepancy of 0.3 from the proportion proposed by the null hypothesis, *proportion* = 0.5

Is this discrepancy by chance alone?:

Estimating the probability of such an extreme result

The *null distribution* for a test statistic is the probability distribution of alternative outcomes when a random sample is taken from a population corresponding to the null expectation.

The null distribution of the *sample* number of red wins



Calculating the P-value from the null distribution

The P-value is calculated as

$P = 2 \times [\Pr(16) + \Pr(17) + \Pr(18) + \Pr(19) + \Pr(20)] = 0.012.$

α is often 0.05

Statistical significance

The significance level, α, is a probability used as a criterion for rejecting the null hypothesis.

If the P-value for a test is less than or equal to α, then the null hypothesis is rejected.

Significance for the red shirt example

$P = 0.012$

$P < \alpha$, so we can reject the null hypothesis

Athletes in red shirts were more likely to win.

Larger samples give more information

A larger sample will tend to give and estimate with a smaller confidence interval

A larger sample will give more power to reject a false null hypothesis

Sample R code for doing this simulation (Note: This is not the most efficient code for this!)

```
binarySample = function(n, prob){
  results = rep(NA,n)
  for(i in 1:n){
    if(runif(1) < prob) results[i] = "red"
  else
    results[i] = "blue"
  }
  length(which(results=="red"))
}

numreps=10000
resultsDF = data.frame(numberRedWins =
    replicate(numreps, binarySample(20,.5)))
```

Hypothesis testing: another example

Do dogs resemble their owners?

Common wisdom holds that dogs resemble their owners. Is this true?

41 dog owners approached in parks; photos taken of dog and owner separately

Photo of owner and dog, along with another photo of dog, shown to students to match

Roy, M.M., & Christenfeld, N.J.S. (2004). Do dogs resemble their owners? *Psychological Science*, **15**, 361–363

Hypotheses

H_0: The proportion of correct matches is *proportion* = 0.5.

H_A: The proportion of correct matches is different from *proportion* = 0.5.

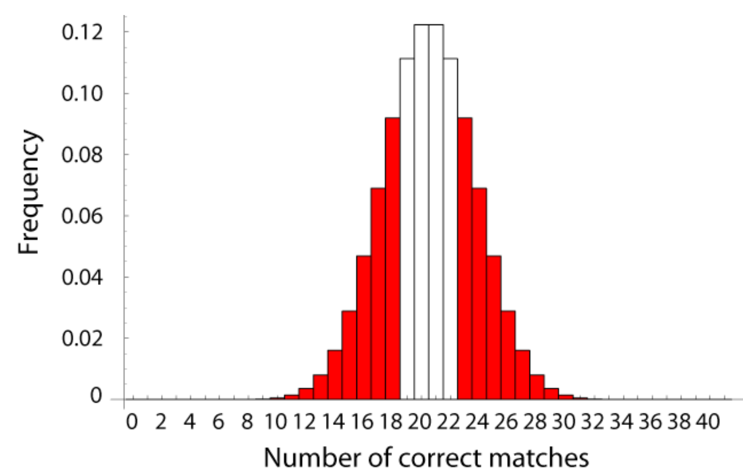
Data

Of 41 matches, 23 were correct and 18 were incorrect.

Estimating the proportion

$$sample\ proportion = \frac{23}{41} = 0.56$$

Null distribution for dog/owner resemblance



The *P*-value:

$P = 0.53$

We do not reject the null hypothesis that dogs do not resemble their owners.

Jargon

Significance level

The acceptable probability of rejecting a true null hypothesis

Called α

For many purposes, $\alpha = 0.05$ is acceptable. α is somewhat arbitrarily chosen by researchers.

Type I error

Rejecting a true null hypothesis

Probability of Type I error is α (the significance level)

Type II error

Not rejecting a false null hypothesis

The probability of a Type II error is β.

The smaller β, the more *power* a test has.

Power

The ability of a test to reject a false null hypothesis

Power $= 1 - \beta$

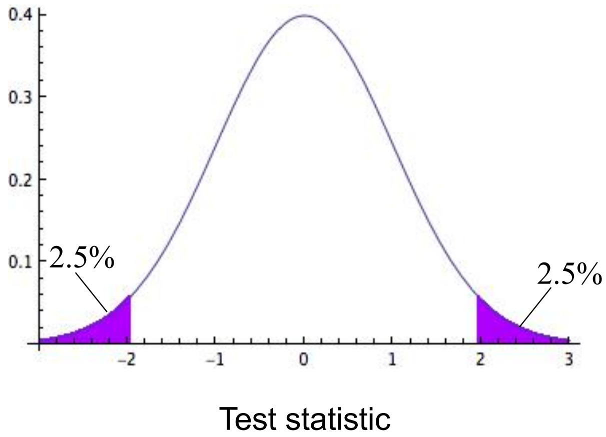
Power increases with more information (i.e. with larger sample size)

One- and two-tailed tests

Most tests are *two-tailed tests.*

This means that a deviation in either direction would reject the null hypothesis.

Normally α is divided into $\alpha/2$ on one side and $\alpha/2$ on the other.



One-tailed tests

Only used when the other tail is nonsensical

For example, comparing grades on a multiple choice test to that expected by random guessing

Critical value

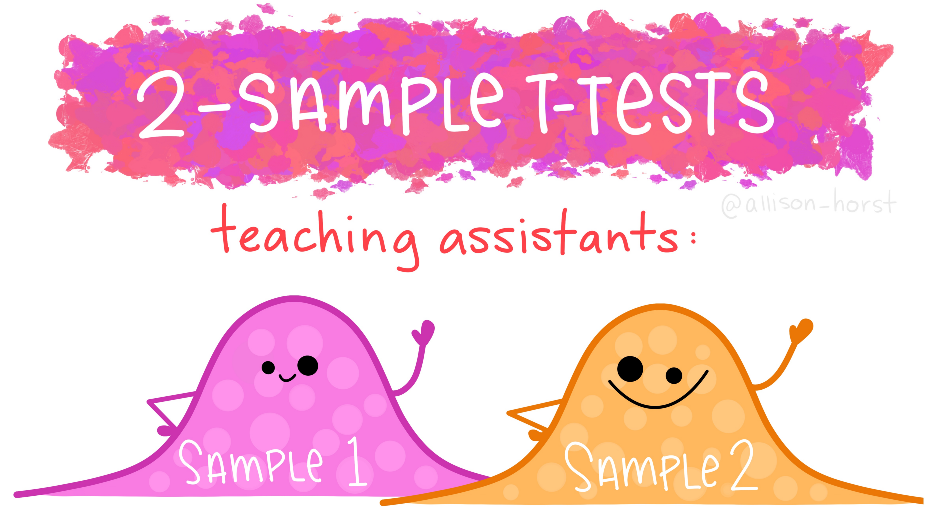
The value of a test statistic beyond which the null hypothesis can be rejected

We never "accept the null hypothesis"

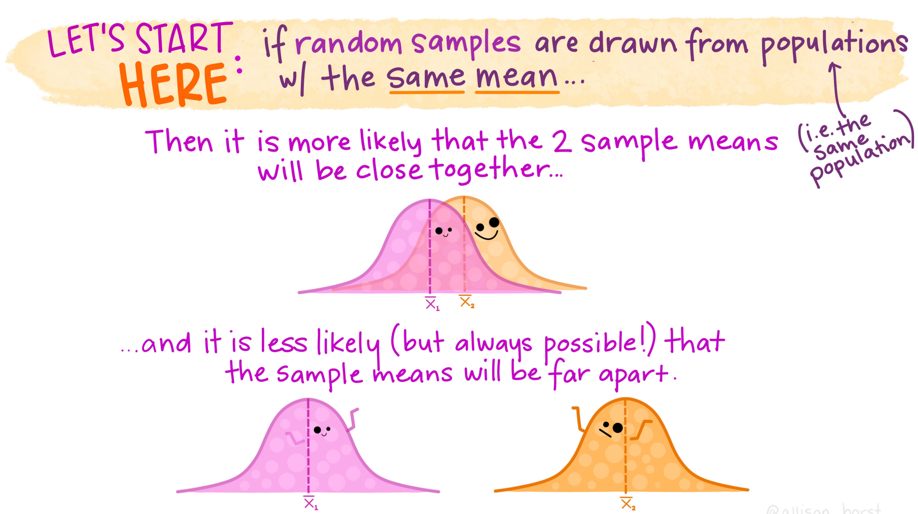
“Statistically significant”

$P < \alpha$

We can “reject the null hypothesis”

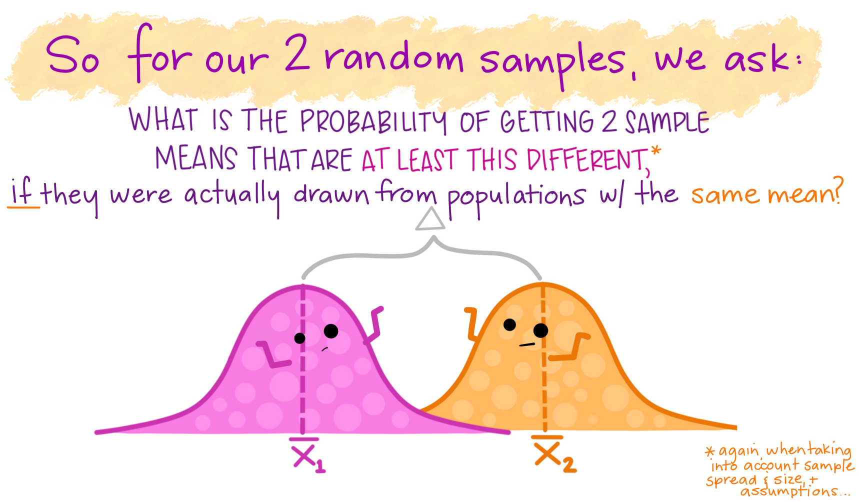


Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations

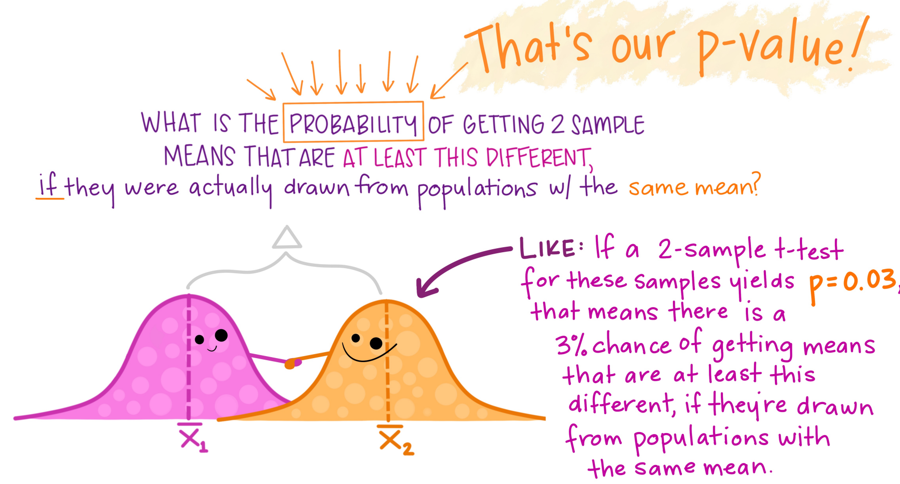


Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations

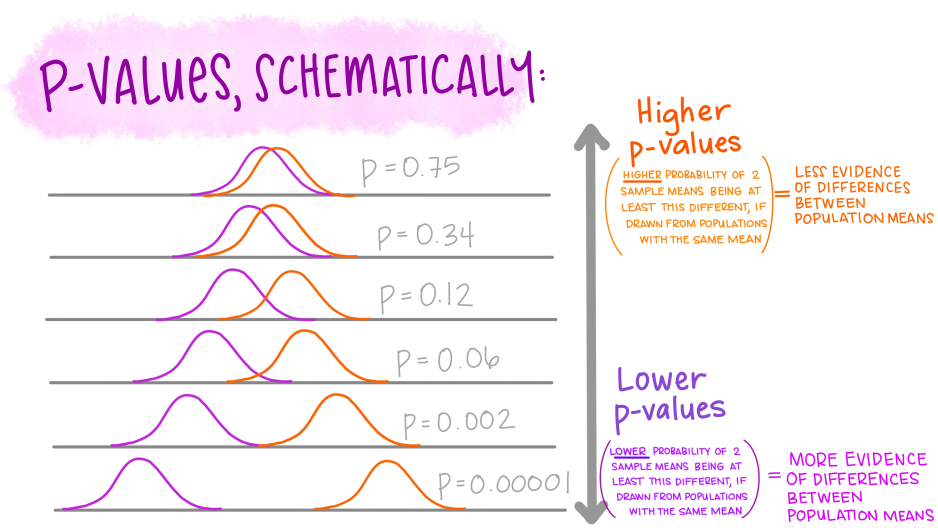
Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations



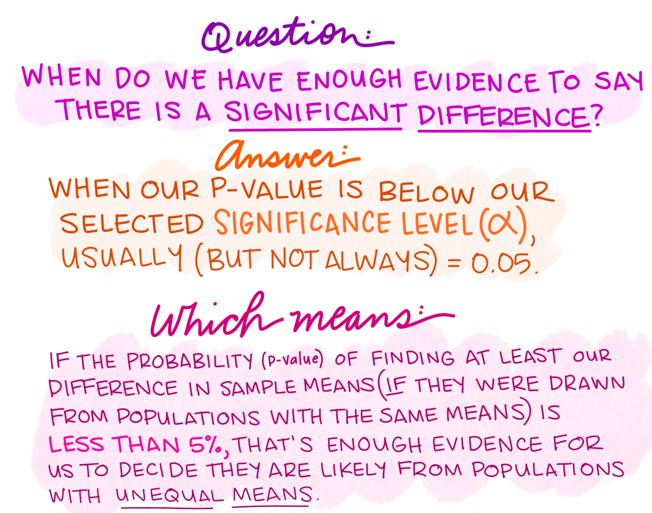
Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations



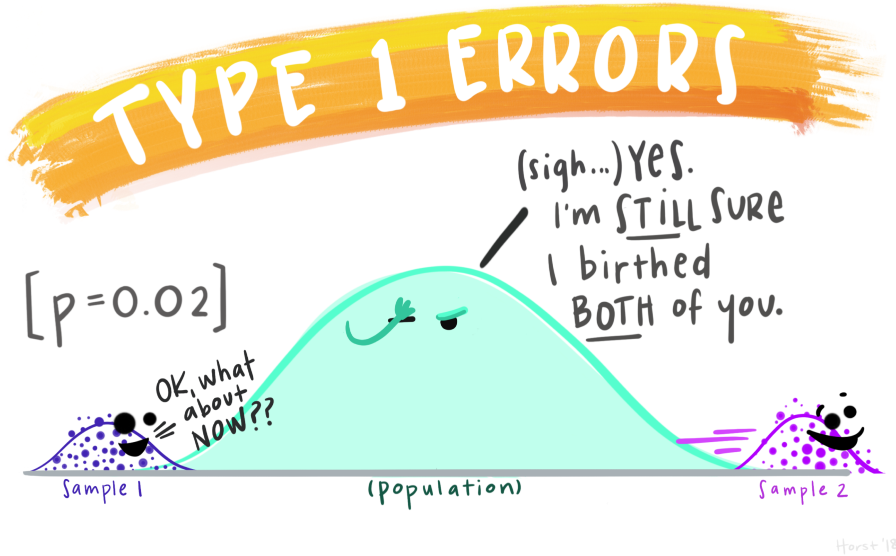
Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations



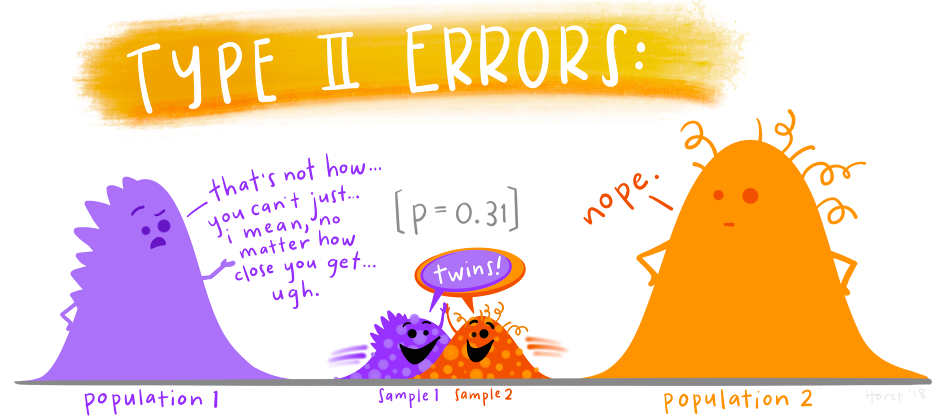
Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations



Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations



Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations

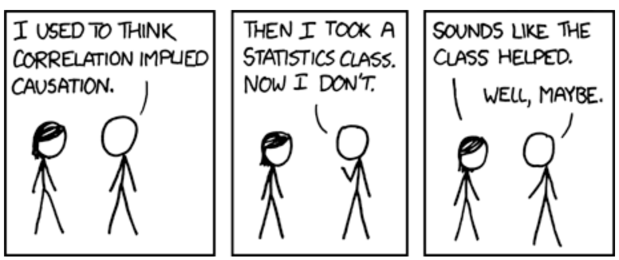


Artwork by @allison_horst https://github.com/allisonhorst/stats-illustrations

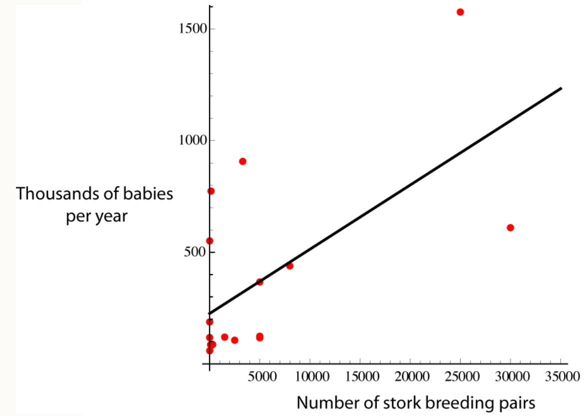
Statistical significance ≠ Biological importance

	Important	Unimportant
Significant	Polio vaccine reduces incidence of polio	Things you don't care about, *or* already well known things:  BRIEFS Study Shows Frequent Sex Enhances Pregnancy Chances By The Associated Press BOSTON — A study that researchers say gives the best estimate ever of nature's window of female fertility comes to a not-so-startling conclusion: The best way to make babies is to have lots of sex. Every other day is better than once a week. Daily is best. The discovery, however, is not ... ples don't want to use other forms of birth control. Researchers say there are six days in every menstrual month when a woman can get pregnant. Until now, estimates of women's fertility ranged from two days in a menstrual cycle to 10 or more. The new study found that conception is possible if a woman has intercourse on the five days before ovulation as well as on the day ...
Insignificant	Small study shows a possible effect, leading to larger study which finds significance. *or* Large study showing no effect of drug that was thought to be beneficial.	Studies with small sample size and high *P*-value *or* Things you don't care about

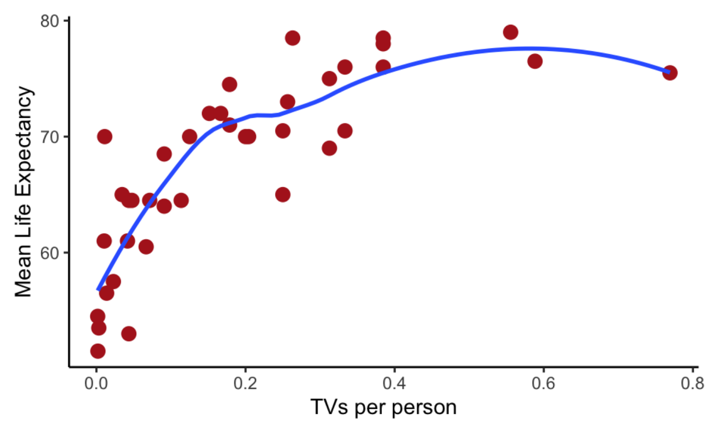
Correlation does not automatically imply causation



Correlation does not automatically imply causation



Life expectancy by country:



Confounding variable

An unmeasured variable that may be the cause of both *X* and *Y*

Observations vs. Experiments