LAW 1: Modern Statistical Inference

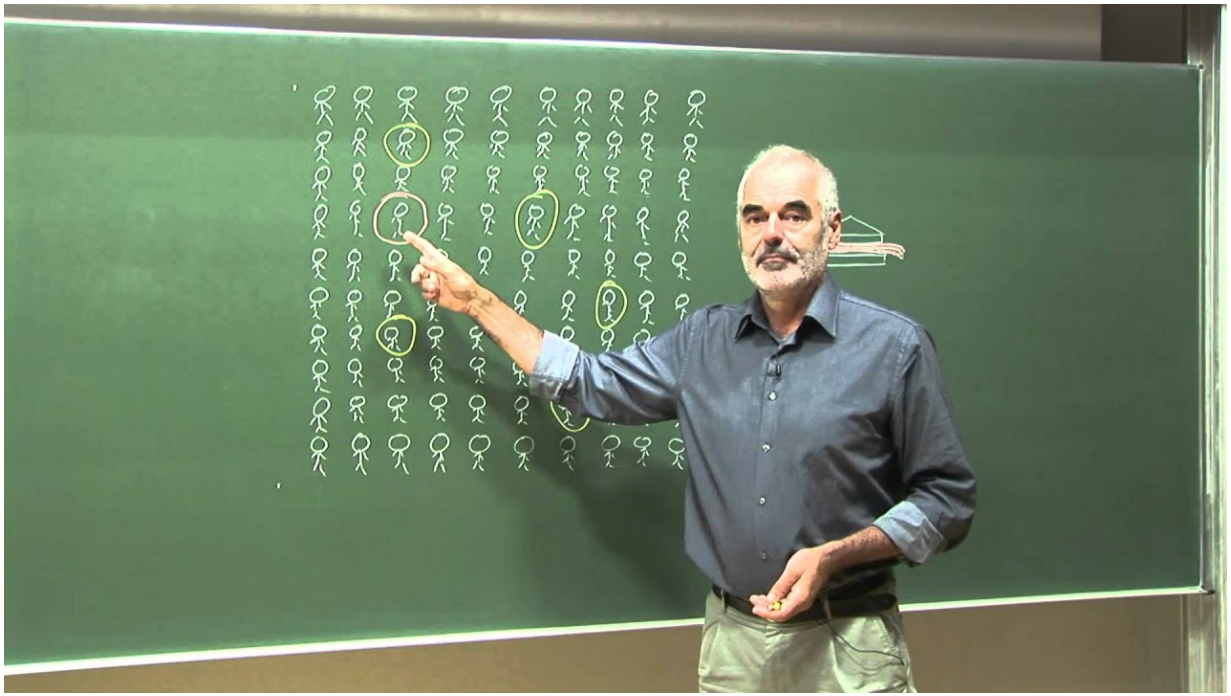
Dr Ian Hunt[1]

[1]Manager, Statistical Consulting Service
Data Science and AI Platform
Monash University
ihunt@bunhill.co.uk; ian.hunt@monash.edu; statisticalconsulting@monash.edu

October 14, 2020

1. Foundations of statistics
2. Running useful hypothesis tests
3. Understanding p-values, size, power and confidence intervals
4. Statistics for research and critical thinking

Why are we here?



- David Speigelhalter began as a statistician and guru mathematician. Now he is Professor of Risk and Evidence Communication at the University of Cambridge.

"In general, I don't feel statistical evidence is handled well by courts. They like either incontrovertible numerical 'facts', or overall expert opinions. But statisticians deal with a delicate combination of data and judgement that often gives rise to ***'rough' numbers****, and these don't seem to fit well with the legal process."* (Spiegelhalter quoted within Flanagan, 2015)

Lecture outcomes

1. Understand the **foundations** of statistics and sampling variance.
2. Run **useful hypothesis tests**.
3. **Understand** p-values, test size, test power and confidence intervals.
4. Use "statistical analysis" for **research and critical thinking**.

Lecture outcomes

- I will make reference to several auxiliary papers and books.
- In particular, I strongly encourage you to purchase and read Spiegelhalter (2019).

- An additional resource to which I refer is Hunt (2020), which is my set of notes for a longer introduction to statistics (this course is not currently being offered at Monash).

Data

- I use the following data set for the examples.
- X = 1.214, 1.480, 1.088, 1.444, 0.637, 1.123, 0.136, 1.490, 0.636, -0.294
- Y = -0.492, 2.843, 2.500, -4.017, -5.082, 1.001, 0.212, -2.490, 1.997, 1.542, 3.198, 2.505, 0.881, 0.311, 1.445

Table of Contents

A sample is born

- Let a **sample** be a set of n observations that you lump together. This is the data of sample size n (where "lumping" simply means to group together in a logical set).
- A sample is an observed **sub-set** of a wider unobserved **population**.
- **Hypothetically**, you could have observed a different sample from the population.

Don't be a philistine

- *"Anyone who disdains the hypothetical is a philistine."* (Good, 1983, page 29)
- I. J. Good was Bayesian pioneer in the mid 20th century. He spent many decades arguing, in good humour, for his practical but ingenuous approaches to real world problems. Good worked closely with Alan Turing at Bletchley Park during World War II.
- He was also an enigma and possibly a spy.

Distributions

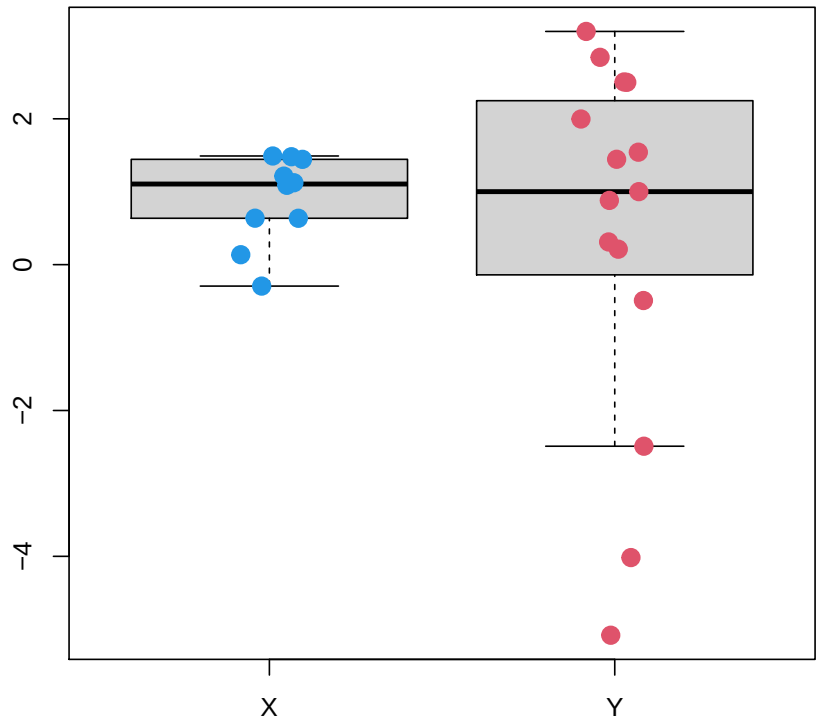
- Populations have distributions.
  - A sample's distribution approximates the population's distribution.
- Think of a histogram.
- Think of ordering the data from smallest to biggest and then plotting this versus its **quantiles**.

Statistical summaries rule

- Means, medians, standard deviations, skewness, kurtosis ...
- We will focus on the **mean** and **standard deviation**.
- A mean is the sample average of the data.
- Standard deviation is the square root of the **variance**.
  - Variance is a "measure of spread" of the data.
  - A sample variance is the average squared difference to the mean.

Use tables and charts to summarise

	X	Y
n	10.00	15.00
mean	0.90	0.42
sd	0.61	2.49
var	0.37	6.21
min	-0.29	-5.08
max	1.49	3.20
median	1.11	1.00
quantile .25	0.64	-0.14
quantile .5	1.11	1.00
quantile .75	1.39	2.25



The mean is random

- The mean is a "**random variable**", just like each data observation.
  - Because hypothetically, you could have had a different sample and thus different mean.
- What is the standard deviation of the mean, then?
  - A simple function of the standard deviation of the data and n.
- For simplicity, I will assume in these notes that samples consist of **independent and identically distributed** ("iid") data.

Sampling variance

- Let the "**sampling variance**" of any estimated measure be its variance.
- Sampling variance gauges how accurately we have estimated something.
  - But we have to estimate sampling variance from the data, as well ....
- Sampling variance should go *down* as *n* goes *up*.

Estimating sampling variance of the mean

Three points of view

1. Take new samples from the actual population.
2. Re-sample (or **"bootstrap"**)the data that you do have.
   - Pretending that this is a new sample.
3. Use mathematical **theory**.
   - First, make **assumptions**, like "the Central Limit Theorem applies".
   - Then **plug-in** estimates of missing pieces, based on your sample.

1. Take new samples

- Take new samples from the actual population.
- For each new sample you re-calculate the mean.
- The distribution of the new means represents the means you could have got originally, but didn't.
- Sampling variance is the variance of the new means!
- *Too expensive and, in any case, illogical!*
  - *You would just use the super-set of observations that you have collected.*

2. Re-sampling or bootstrapping

- Bootstrapping a sample consists of creating new data sets of the same length by re-sampling the original data, with replacement.
- For each bootstrap sample you re-calculate the mean.
- The distribution of the bootstrap means represents the means you could have got but didn't.
  - Requires assumptions e.g. the data is "representative" of the population.
- *Sampling variance is the variance of the bootstrap means!*

3. Theoretical approach

- The Central Limit Theorem says that the sample mean of a set of (*iid*) data, randomly sampled from a population, tends to have a normal distribution.
  - Regardless (with a few exceptions) of the shape or distribution of the population.
- This relies the assumption that n is "big enough."
  - This varies with how non-normal or funky the population is.
- *The sampling variance of a mean is approximately the variance of the data divided by n.*

Difference in means?

- The difference between the means of two samples, say a "control sample" and a "treatment sample", is also a mean.
- So the sampling variance of the difference in means between samples can be calculated by the same methods used in single samples.
- Bootstrapping applies in *any case*.
  - But the theoretical approach only applies to averages.

Sampling variance of X example

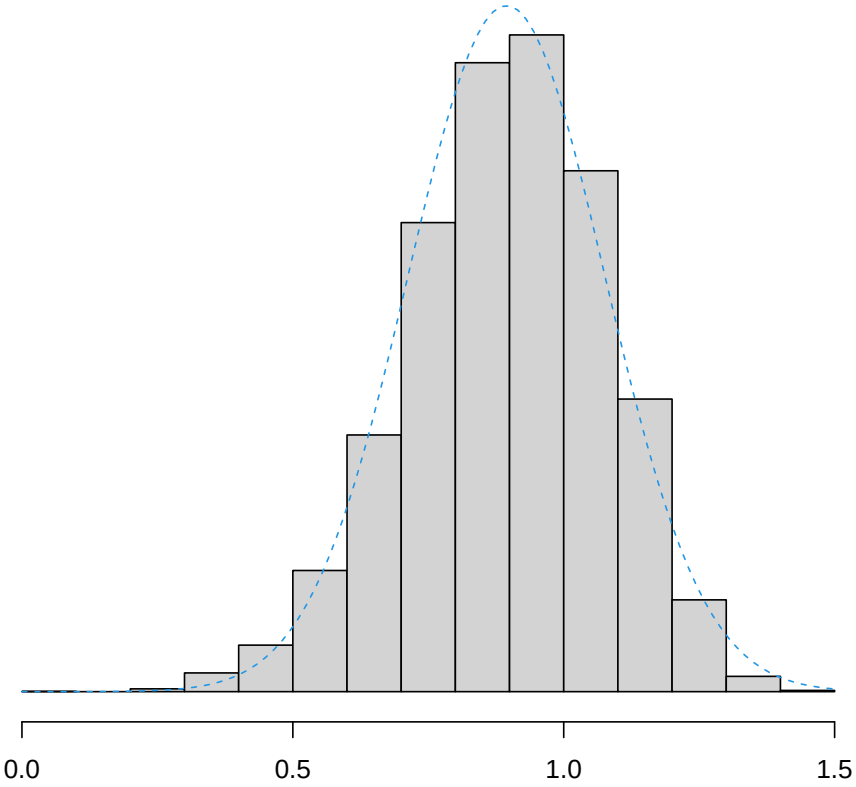


Figure 1: Histogram of 10000 bootstrap re-sampled means of X (each new sample has n=10). The curve is a normal distribution with the same mean and variance as the histogram data.

- The means are approximately normally distributed: **the Central Limit Theorem in action!**
- See Hesterberg (2015) for more.

Sampling variance of X example

	bootstrap samples	theoretical
n	10000	
mean	0.894	
sd	0.182	0.192
var	0.033	0.037
quantile .025	0.518	
quantile .975	1.218	

Table 1: Statistical summary of the bootstrap samples for X. The theoretical values assume the central limit theorem applies.

For X, the sampling variance of the bootstrap samples is approximately the same as that implied by the theoretical approach.

Key words and reading

- Sampling variance
  - Spiegelhalter (2019, chapter 3 and chapter 7)
  - Hunt (2020, section 1.2)
- Bootstrapping
  - Spiegelhalter (2019, chapter 7)
  - Hunt (2020, section 3.3)
  - Hesterberg (2015)
- Central Limit Theorem
  - Hunt (2020, section 4.3)

Quiz

Question: Which statement[s] are false?

A. Bootstrapping can gauge the sampling variance of any estimate, not just means.
B. Summary statistics are always more useful than charts.
C. The Central Limit theorem doesn't apply to skewed. data
D. Bootstrapping requires no assumptions.

Answer: B, C and D.

Table of Contents

Philosophy of Science

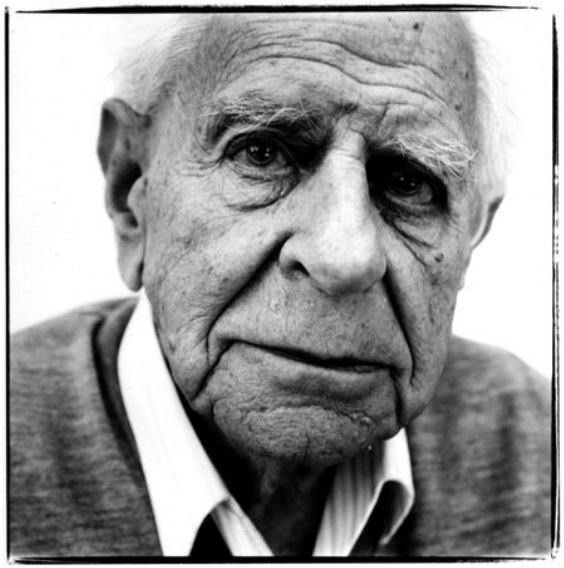


Figure 2: Karl Popper.

- Karl Popper's theory about scientific methodology is that we advance knowledge by "conjecturing" theories, which we subsequently try to "refute".
- It is too hard to *refute* things in real science.
- **But we do form hypotheses and test them!**

Four Popper principles

1. Be **explicit** about your premises for models and hypotheses.
2. Do as many *meaningful* tests as you can (within reason).
   - Tests must have the power to identify evidence against false hypotheses.
3. Present all your key statistical results.
   - And do this in a format that is useful for others i.e. **share**.
4. Be **creative** and **critical**.

Useful "null hypotheses" ...

- The **true mean** of X is zero.
- The **true mean** of X is the same as the **true mean** of Y.
- The **distribution** of X is the same as the distribution of Y.
- The **proportion** of "negatives" is the same for X and Y.

A prototype hypothesis test

1. Take a random sample of size n from the population of interest.
2. Calculate the mean and sample standard deviation
3. Add assumptions:
   - The Central Limit Theorem is tenable (or holds approximately).
   - The true mean is equal to zero [or something else meaningful] (this is the "null hypothesis").

Abstract notation

Abstract notation

- The data: x_i for $i = 1$ to n, let each x_i be *iid* with true mean μ and sd σ.
- Sample mean: $\hat{\mu} = \sum x_i / n$.
- Sample variance: $\hat{\sigma}^2 = \sum (x_i - \hat{\mu})^2 / (n - 1)$.
- Variance of the mean: $\hat{\sigma}^2 / n$.
- Null hypothesis: $H_0 : \mu = \mu_0$
  - i.e. "the true mean μ is equal to μ_0".

It's all about assumptions

- With the Central Limit Theorem and assuming the null hypothesis is true we know that:

$$\hat{\mu} \overset{d}{\sim} N(\mu_0, \frac{\sigma^2}{n})$$

- This says that the sample mean "tends in distribution" (as n increases) to a normal random variable with the same mean as the true mean of x_i (which the null hypothesis assumes is μ_0) and variance σ^2/n.

Abstract notation ... with assumptions ...

Don't run: it is not so bad

- The only problem now is that we do not know σ^2. **But we can estimate it with $\hat{\sigma}^2$!**

- Then we can use a so-called t-test statistic, which will be t-distributed with $n-1$ degrees of freedom, rather than normally distributed:

$$t = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \tag{1}$$

- And that is basically it. No screams.
  - Equation (1) is strictly only t-distributed if the underlying data are normally distributed. We will side-step this wrinkle, which is only of concern for "small" samples.

A useful hypothesis test 1/2

- Assume something substantive about the distribution of the data, in the form of a null hypothesis.
- Combine the logical implications of the null hypothesis with other assumptions (especially the Central Limit Theorem).
- Aim for a **test statistic** that has a known distribution (the "**null distribution**") when all the assumptions are true.

A useful hypothesis test 2/2

- Compare the test statistic with the *quantiles* of the null distribution.
- Assume that if the null hypothesis is **false** then the test statistic value is more likely to be in the **outer extremes** of the null distribution.
- If the test statistic lies in an **extreme** part of the null distribution then conclude there is empirical evidence against the null hypothesis.

Definition of extreme?

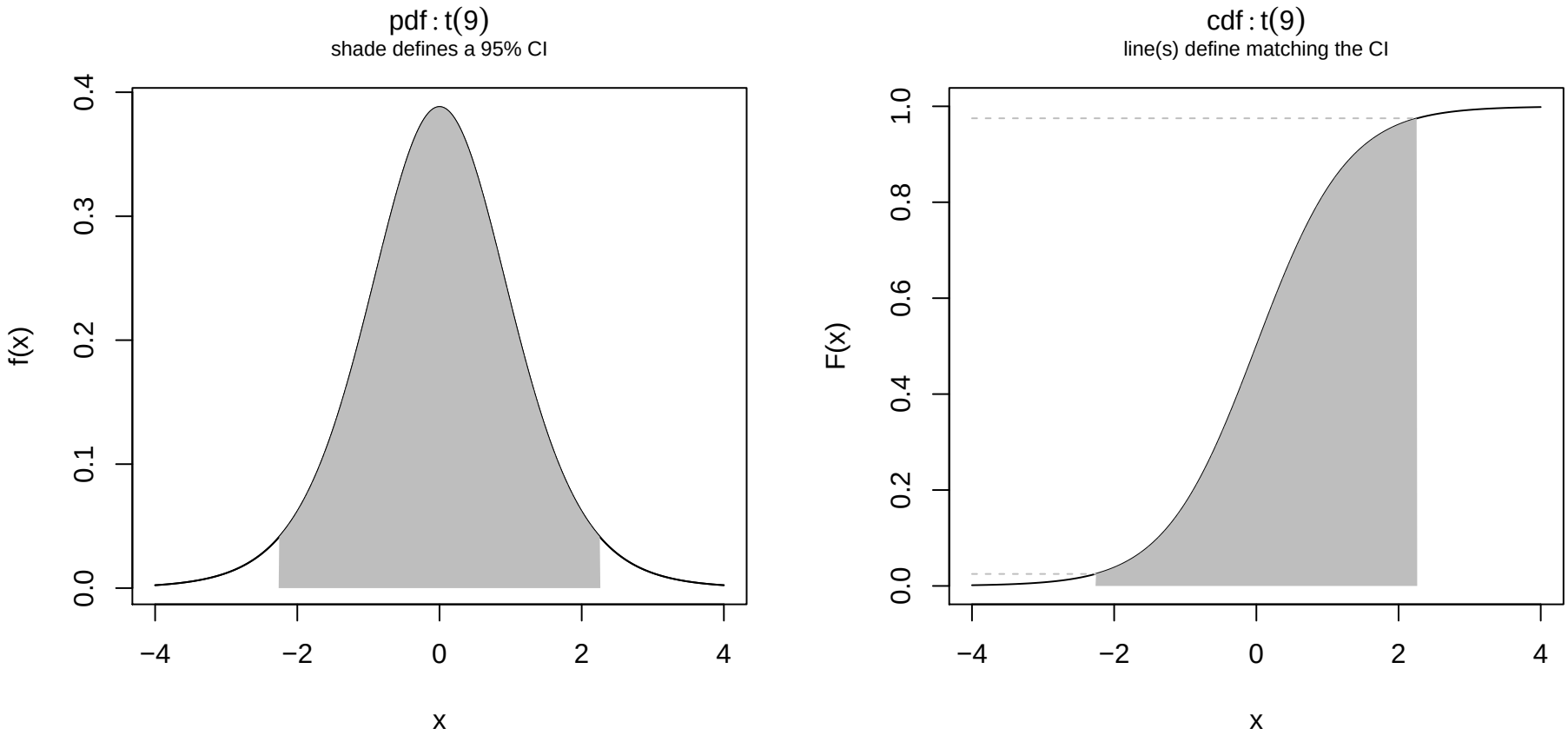


Figure 3: A t-distribution with, 9 or $(n-1)$, degress of freedom. The shaded area (95%) defines the cut-offs for what counts as 'extreme'. See Hunt (2020, section 4.4) for more about probability density functions (left-hand-side of chart) and cumulative density functions (right-hand-side of chart).

What makes a test statistic extreme?

Maths (t-statistic)

$$t = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

- Big numerator.
- Small denominator.

Example: true mean is zero?

- Assume that the true mean of the data is equal to zero and calculate a t-statistic
- Then compare the t-statistic to the quantiles of the t-distribution
- If the t-statistic is extreme then conclude there is evidence against the null hypothesis
  - This approximately requires a t-statistic > 2 or < -2 (for n of 30 or more).

Example: true mean is zero?

	X	Y
n	10.00	15.00
sample mean	0.90	0.42
hypothesised mean	0.00	0.00
mean sd	0.19	0.64
t-stat	4.67	0.66
p-value	0.00	0.52
95% CI lower	0.46	-0.96
95% CI upper	1.33	1.80
t(n-1) quantile .975	2.26	2.14

Table 2: Summary and test statistics for X and Y.

- H_0^x: "true mean of $X = 0$"
- H_0^y: "true mean of $Y = 0$"

Challenge

Would you reject H_0^x and H_0^y?

*Example: true means of X and Y are the same?

- Same idea as the one-sample test but if variances are different then use a so-called "Welch" adjusted t-test.
- H_0: "true mean of X = true mean of Y"

*A non-parametric test

- A "Wilcoxon rank-sum test" (or Mann-Whitney) tests the null hypothesis that two distributions are the same.
  - If the distributions have the same variance and are symmetric then this is very similar to testing for a difference in means.
- Both samples are pooled and ranked by value.
  - The test then assesses the observed "sum of the ranks" for one of the samples against what would be expected by chance if the null hypothesis was true.

If you understand the t-test then you understand them all ...

- Conventional hypothesis tests comprise of the following three things.
  1. An "estimated value" which is in some sense is an average.
  2. A "true value" (as assumed by a null hypothesis.)
  3. The sampling variation (or its square root) of the estimated value.
- These things combine so that a known probability distribution applies.
  - The use of averages justifies invoking the Central Limit Theorem.
- This is how the t-test works (check).
- Other tests that more-or-less follow this recipe (with different combination functions) include Chi-squared tests for contingency tables, F-tests for regression models, ANOVA, Wald tests for model parameter restrictions and most "non-parametric" tests.

Key words and reading

- Null hypothesis.
- Test statistic.
- Null distribution.
- Reading
  - Spiegelhalter (2019, chapter 10).
  - Hunt (2020, chapter 3 and section 4.4)

Quiz

Question: What tends to increase the absolute value of a test statistic?

A. Bigger sample size.
B. Lower variance or noise within the underlying data.
C. Larger difference between the true mean of the data and its hypothesised mean.
D. All of the above.

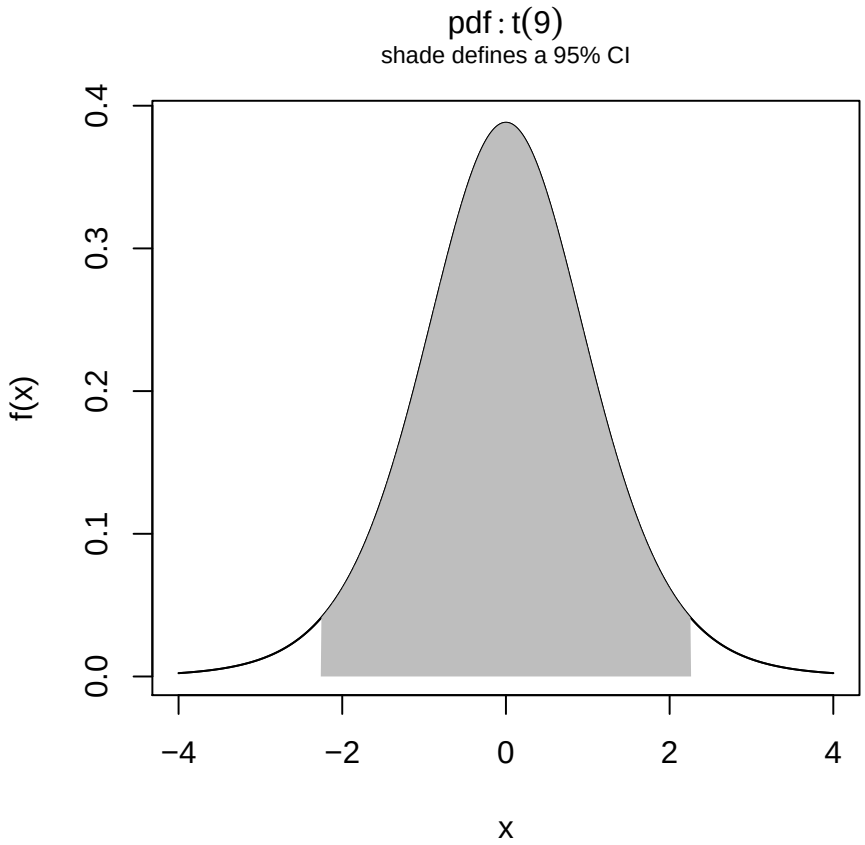
Answer: D.

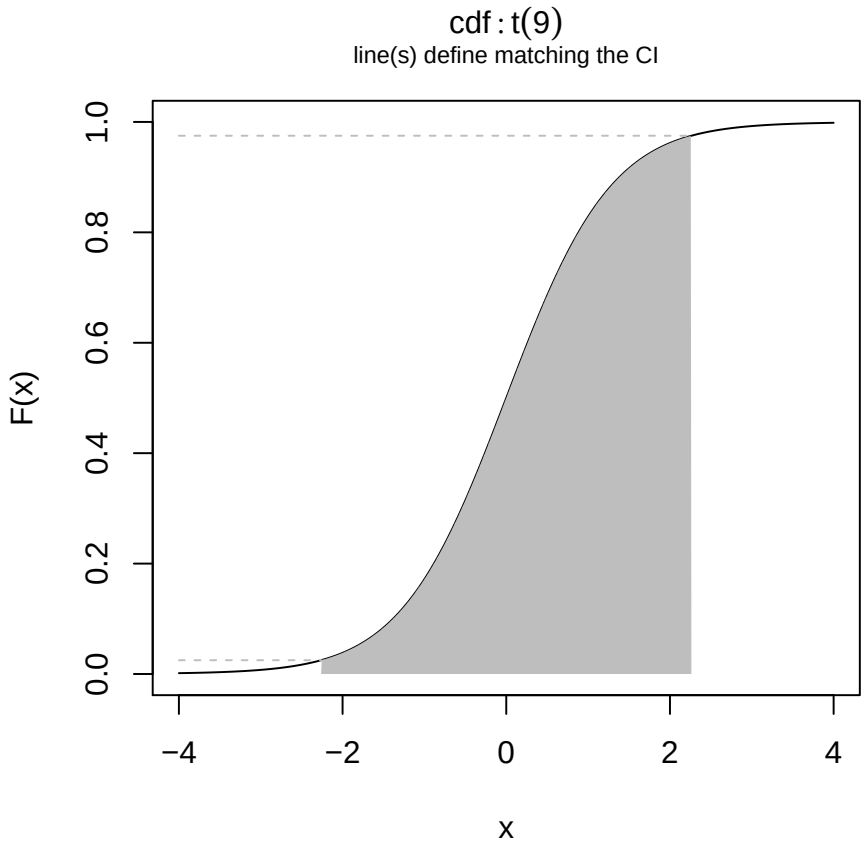
Table of Contents

p-values

- A p-value is a measure of discrepancy between data and a null hypothesis
- For a conventional test, like a t-test, this can be expressed as the probability of observing a test statistic "at least as extreme" as what was observed if the null hypothesis was true
- **It is not the probability of the null hypothesis being true!**

"Look-up" a p-value





p-value: look up the cumulative density function ("cdf") for the absolute value of the test statistic, then multiply one minus this value by two.

Example: true mean is zero?

	X	Y
n	10.00	15.00
sample mean	0.90	0.42
hypothesised mean	0.00	0.00
mean sd	0.19	0.64
t-stat	4.67	0.66
p-value	0.00	0.52
95% CI lower	0.46	-0.96
95% CI upper	1.33	1.80
t(n-1) quantile .975	2.26	2.14

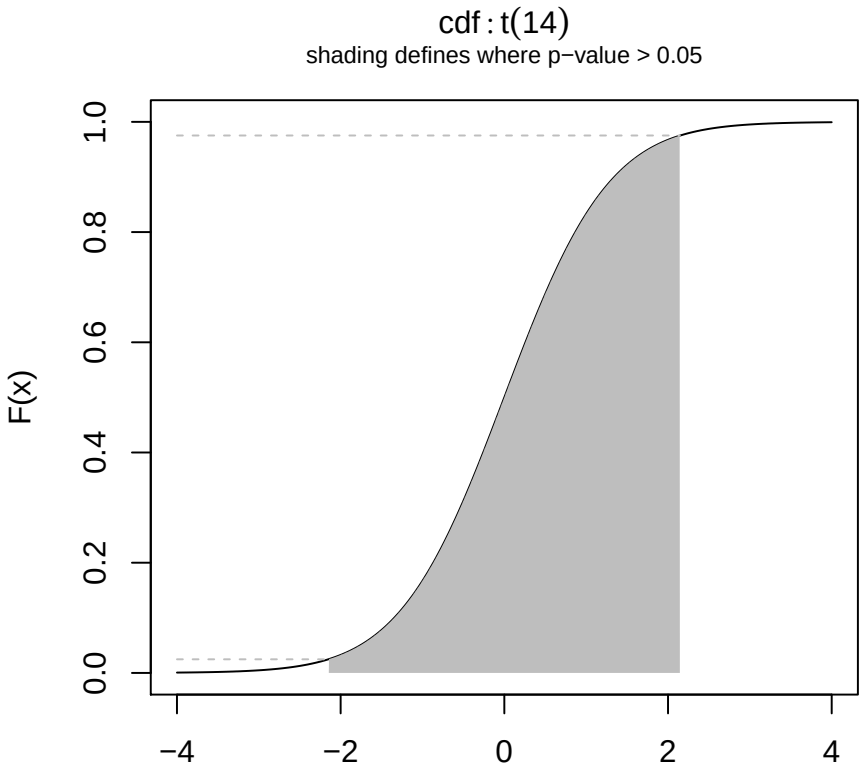


Figure 4: The p-value for the test on Y is 0.521. This is $2(1 - 0.74)$.

Example: true mean is zero?

	X	Y
n	10.00	15.00
sample mean	0.90	0.42
hypothesised mean	0.00	0.00
mean sd	0.19	0.64
t-stat	4.67	0.66
p-value	0.00	0.52
95% CI lower	0.46	-0.96
95% CI upper	1.33	1.80
t(n-1) quantile .975	2.26	2.14

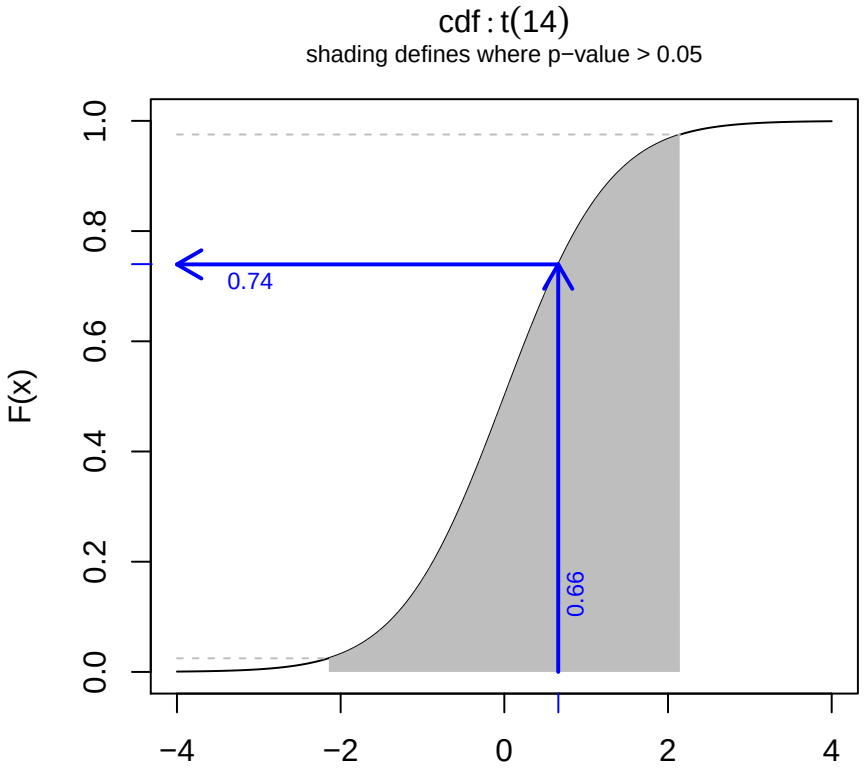


Figure 5: The p-value for the test on Y is 0.521. This is $2(1 - 0.74)$.

Decision rules

- A conventional "decision rule" is to "reject the null" if the p-value is less than some cut-off, usually .05.
  - Which entails the t-statistic lies in the "extreme region".
  - Which entails the 95% confidence interval excludes the hypothesised value.
- A hypothesis rejection is sometimes called (not very usefully) a finding of "statistical significance".

Decision table and errors

- Assume that if the test statistic is extreme enough, then your decision is to "Reject H_0".
  - This rejection is relative to an alternative hypothesis H_a, which contradicts the null in some meaningful fashion.

- It is difficult to know what this really means. Nonetheless, you leave open the possibility of **error**, relative to the truth.

	H_0 **is True**	H_a **is True**
Accept H_0	✓	Type II error
Reject H_0	Type I error	✓

Size and power

- **Size**: let the "size" of a test be the probability that an extreme test statistic occurs when the null hypothesis is true.
  - This is like a "false-positive rate".
- **Power**: let the "power" of a test be the probability of an extreme test statistic when the alternative hypothesis is true.
  - In other words, power is the test statistic's sensitivity to violations of the null.
- *In your research you want high power and low size.*

Size and power in terms of errors

- Let $\mathbb{P}(\mathrm{A}|\mathrm{B})$ mean the probability of A, conditional upon B being true.
- We can say that size is $\mathbb{P}(\text{Type I error}|H_0)$.
- And that power is $\mathbb{P}(\text{Type II error}|H_a)$.

	H_0 **is True**	H_a **is True**
Accept H_0	✓	Type II error
Reject H_0	Type I error	✓

Confidence intervals

- An xx% confidence interval for any measure is an interval such that if you repeatedly made such intervals then xx% of the time the true value of the measure would be within the bounds of the interval.
  - Make an 0.80 confidence interval for the "Student T" distribution at: `https://seeing-theory.brown.edu/frequentist-inference/index.html#section2`
- For any particular measure (you actually only do it once!), its true value will be within its confidence interval with either probability 1 or probability 0.
  - This is analogous to the problem that p-values are not the probability of the null hypothesis being true.

Example confidence intervals

	X	Y
n	10.00	15.00
sample mean	0.90	0.42
hypothesised mean	0.00	0.00
mean sd	0.19	0.64
t-stat	4.67	0.66
p-value	0.00	0.52
95% CI lower	0.46	-0.96
95% CI upper	1.33	1.80
t(n-1) quantile .975	2.26	2.14

Table 3: Summary and test statistics for X and Y.

- A 95% confidence interval for a mean is approximately the mean +/- two times the square root of the sampling variance.
  - More precisely, "two times" is really the .975 quantile for t(n-1).

Challenge

What is the relationship between p-values and confidence intervals?

Challenge:p-values vs. confidence intervals

	X	Y
n	10.00	15.00
sample mean	0.90	0.42
hypothesised mean	0.00	0.00
mean sd	0.19	0.64
t-stat	4.67	0.66
p-value	0.00	0.52
95% CI lower	0.46	-0.96
95% CI upper	1.33	1.80
t(n-1) quantile .975	2.26	2.14

Table 4: Summary and test statistics for X and Y.

- As a rule-of-thumb *"if and only if"* a 0.95 interval excludes the hypothesised value in the test then the p-value is less than 0.05.
- And if a 0.95 interval is exactly on the hypothesised value then the p-value is exactly 0.05.
- These statements assume a symmetric confidence interval and a two-sided test. See more in Hunt (2020, chapter 3).

* Example confidence interval from bootstrapping

	bootstrap X	bootstrap Y
n	10000	10000
mean	0.894	0.419
sd	0.182	0.624
var	0.033	0.389
quantile .025	0.518	-0.860
quantile .975	1.218	1.546

Table 5: Bootstrap sample summaries and implied confidence intervals.

- A different 95% confidence interval can be found from the **quantiles of the bootstrap distribution**.
- Results for small samples can be slightly different to relying on the Central Limit Theorem.
- Other bootstrap options exist (Hesterberg, 2015).

* The probability of a hypothesis

- Gauging the probability of a hypothesis requires **background knowledge** and **prior belief** to be expressed as a probability distribution.
- Bayes' rule (rule/formula/theorem) can be used to combine prior belief with what the data say about a hypothesis.
  - This relies on conditional probability judgements.
- This is not what "classical" hypothesis testing (the examples we are doing) does.
  - There is no *explicit* accounting of prior belief in the hypothesis tests we have examined.

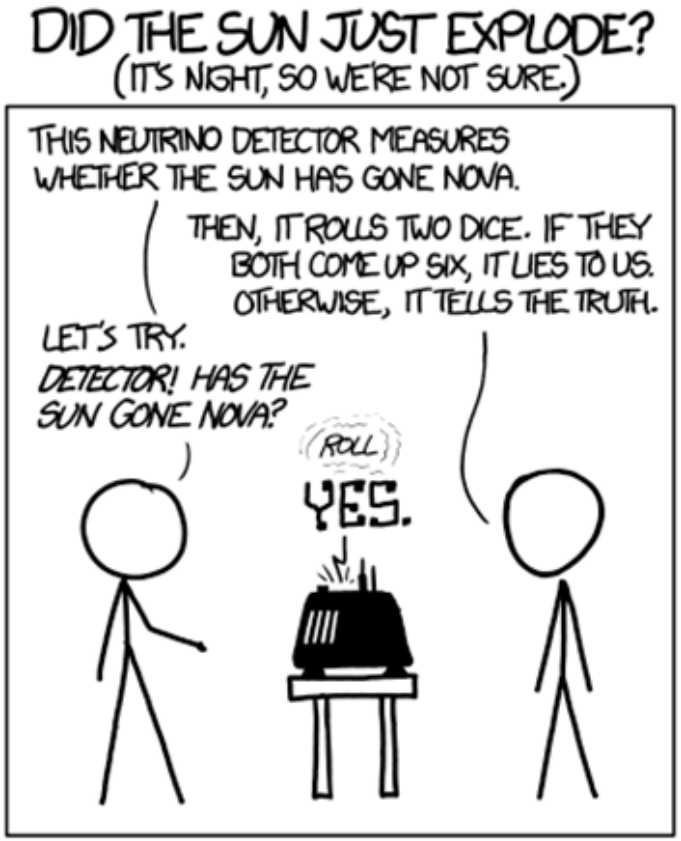
Example for size and power

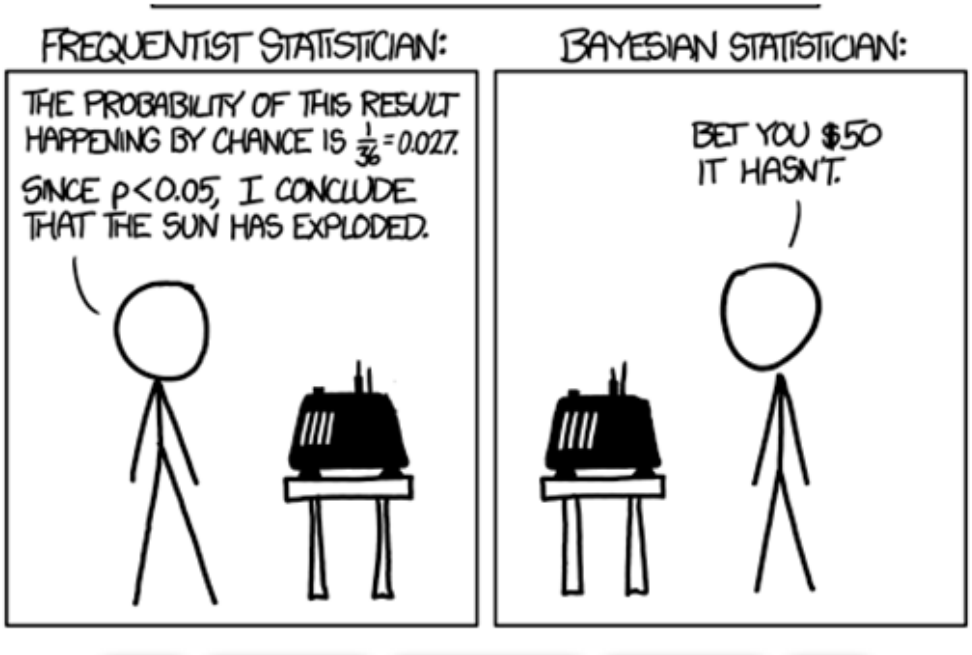
Challenge

Try to work out the **size** and **power** of the test implied by the following **classical** statistical test scenario.

Why is the Bayesian clever?

Example for size and power





Example for size and power — discussion

- H_0: "the sun didn't just explode."
- H_a: "H_0 is false, so we are doomed. Doomed."
- Size $= 1/36 = 0.027$
- Power $= 1 - 1/36 = 0.973$
- The Bayesian used her **background knowledge and science!**

Example for size and power — discussion

- The *power* of a test is a function of sample size, the variance of the data and the "extent of the null hypothesis violation".
  - So power is a number (a probability) only when each of these three pieces is fixed.

- For a t-test (look at the test statistic function):
  - Power increases with n.
  - Power decreases with sample variance (the "noise", if you like).
  - Power increases the further that the true mean is from what is hypothesised (the "signal", if you like).

Maths (t-statistic)

$$t = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

Key words and reading

- Concepts
  - p-value
  - Cumulative density function (and extreme test statistic)
  - Decision rule
  - Size
  - Power
  - Confidence interval
- Reading
  - Spiegelhalter (2019, chapters 7, 8, 9 and 10).
  - Hunt (2020, chapter 3 and section 4.4).

Quiz

Question: Which statements are fallacies?

A. The probability of a hypothesis is of no use for decision theory.
B. A p-value is a measure of discrepancy between a null hypothesis and the data.
C. An 95% confidence interval means a 95% chance the true value lies within it.
D. A p-value is the probability of the null hypothesis being false.

Answer: A, C and D

Table of Contents

Critical thinking (reading)

- Spiegelhalter (2019, chapters 10 and 14)
- Schwen and Rueschenbaum (2018).
- Amrhein et al. (2019) .

Lumping and splitting

- Statistics is about lumping and splitting
  - Classifications of "similar" cases are key.
  - Finding the right dimensions for your research and its write-up is difficult.
- For example: do jelly beans cause acne?
  - Eater dimension: volume eaten, age, gender, country, general diet ...
  - Jelly bean dimension: size, shape, colour ...
  - Outcome dimension: acne count, region size, affect, treatment required ...

Dimensions and big data

- You can think of a data set as a two dimensional matrix (rows and columns).
  - Let each row correspond to a basic observational unit like a patient or transaction or law case.
  - Let the columns be specific measurements about the row-level unit - for example, a patient's age, treatment type and recovery time; or a transaction's date, type and amount; or a law case's year, court, jurisdiction and outcome.
- The term "big data" usually refers to a "vast number of rows".

Dimensions and big data

- But *most* academic research is not about big data: there are typically too many columns relative to rows.
  - For example, DNA expressions for 7 people or *attributes* of legal cases in Australia.
- This means lumping of columns is often required and the amount of splitting in terms of row units is limited.
- *Searching for the right lumping and splitting categories is both art and science.*

Dimensions and big data

- But *most* academic research is not about big data: there are typically too many columns relative to rows.
  - For example, DNA expressions for 7 people or *attributes* of legal cases in Australia.
- This means lumping of columns is often required and the amount of splitting in terms of row units is limited.
- *Searching for the right lumping and splitting categories is both art and science.*

Rationality

"Ignorance: He that judges without informing himself to the utmost that he is capable, cannot acquit himself of judging amiss."
(Locke, 1690, chapter XXI, paragraph 69)

We should take at least two things from Locke.

1. Searching across data sets, in terms of lumping and splitting, is a fundamental requirement of being rational.
2. We need to go beyond statistical hypothesis tests and p-values — logical steps include examining how results change across different points-of-view, scenarios and "models".

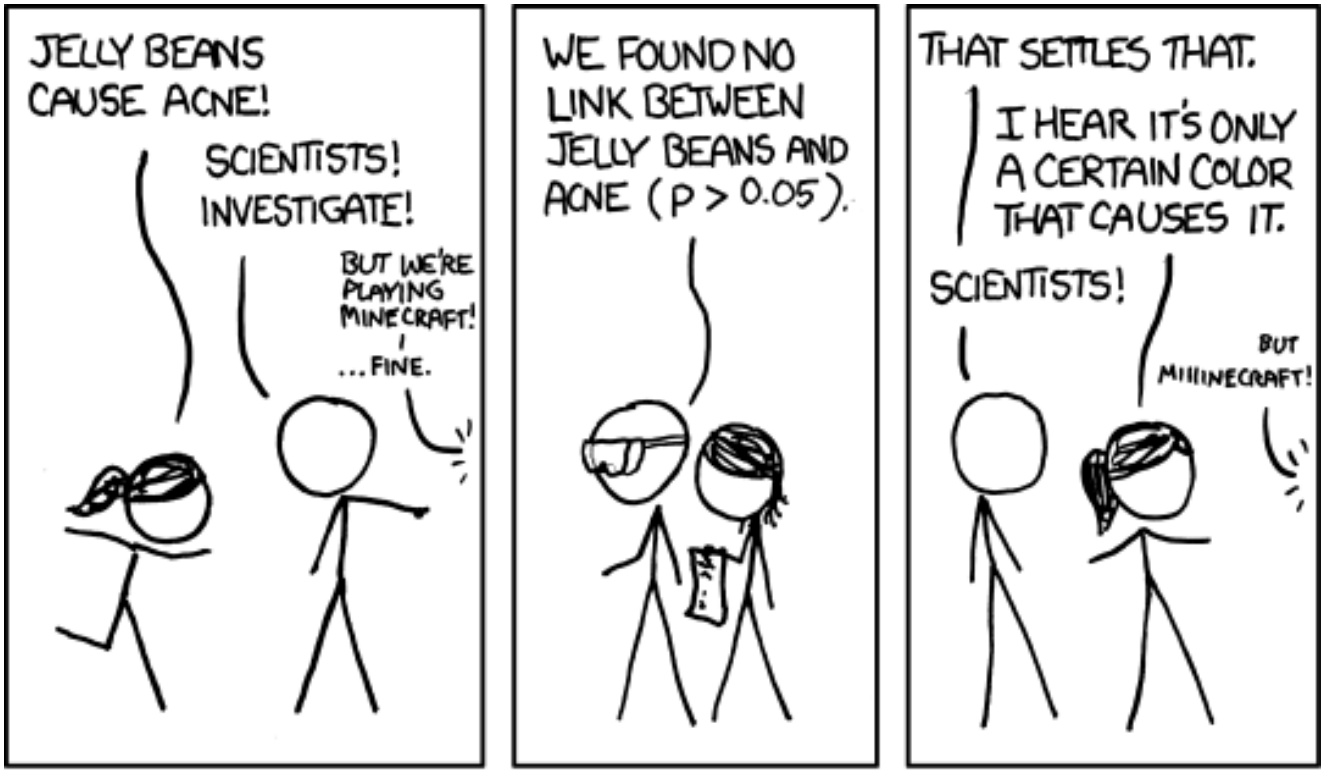
Example of lumping splitting

Challenge

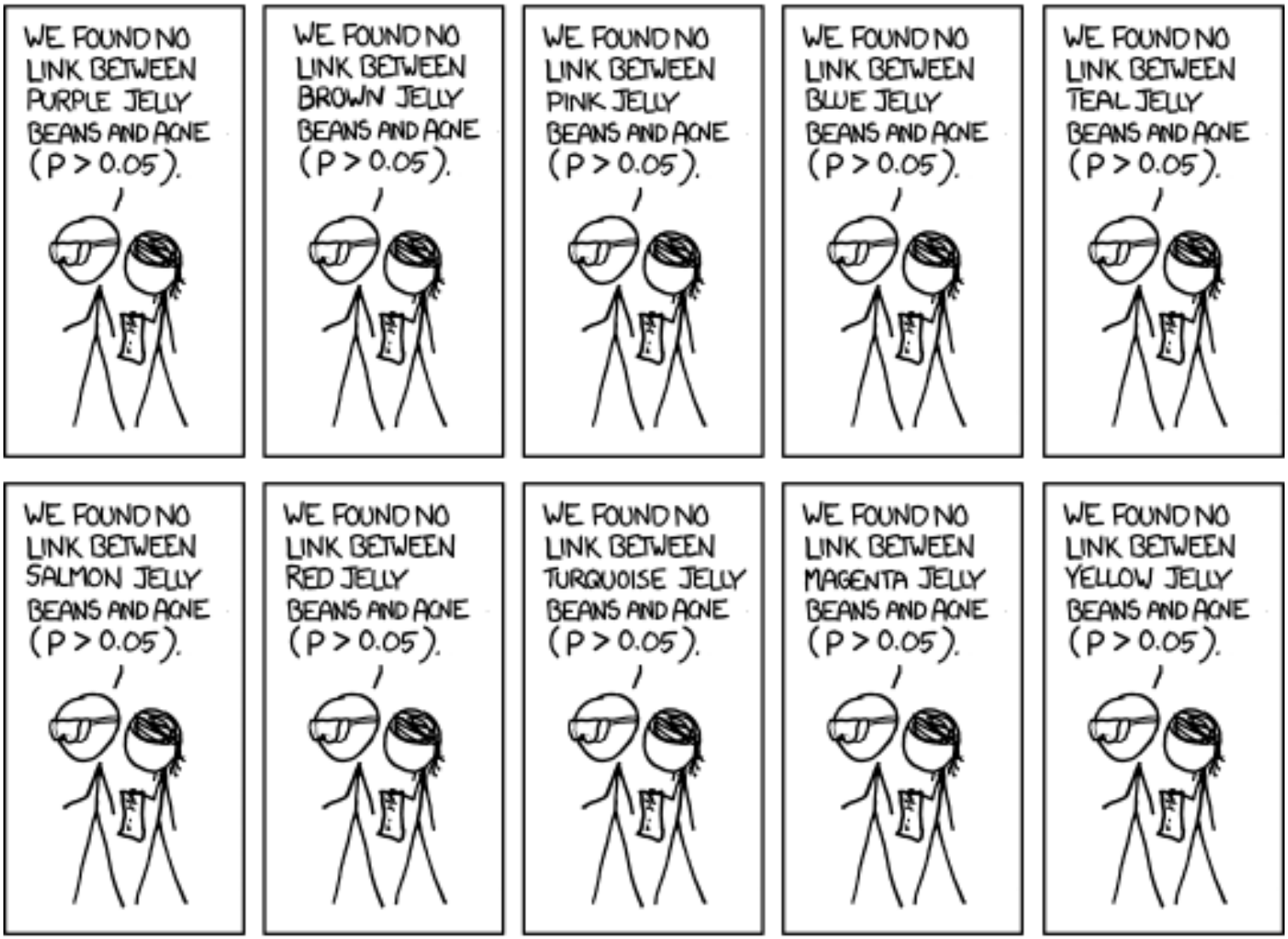
Try to work out what is going wrong (and right) with the following example of "scientific research".

Each of following four slides is from `xkcd.com/882/`.

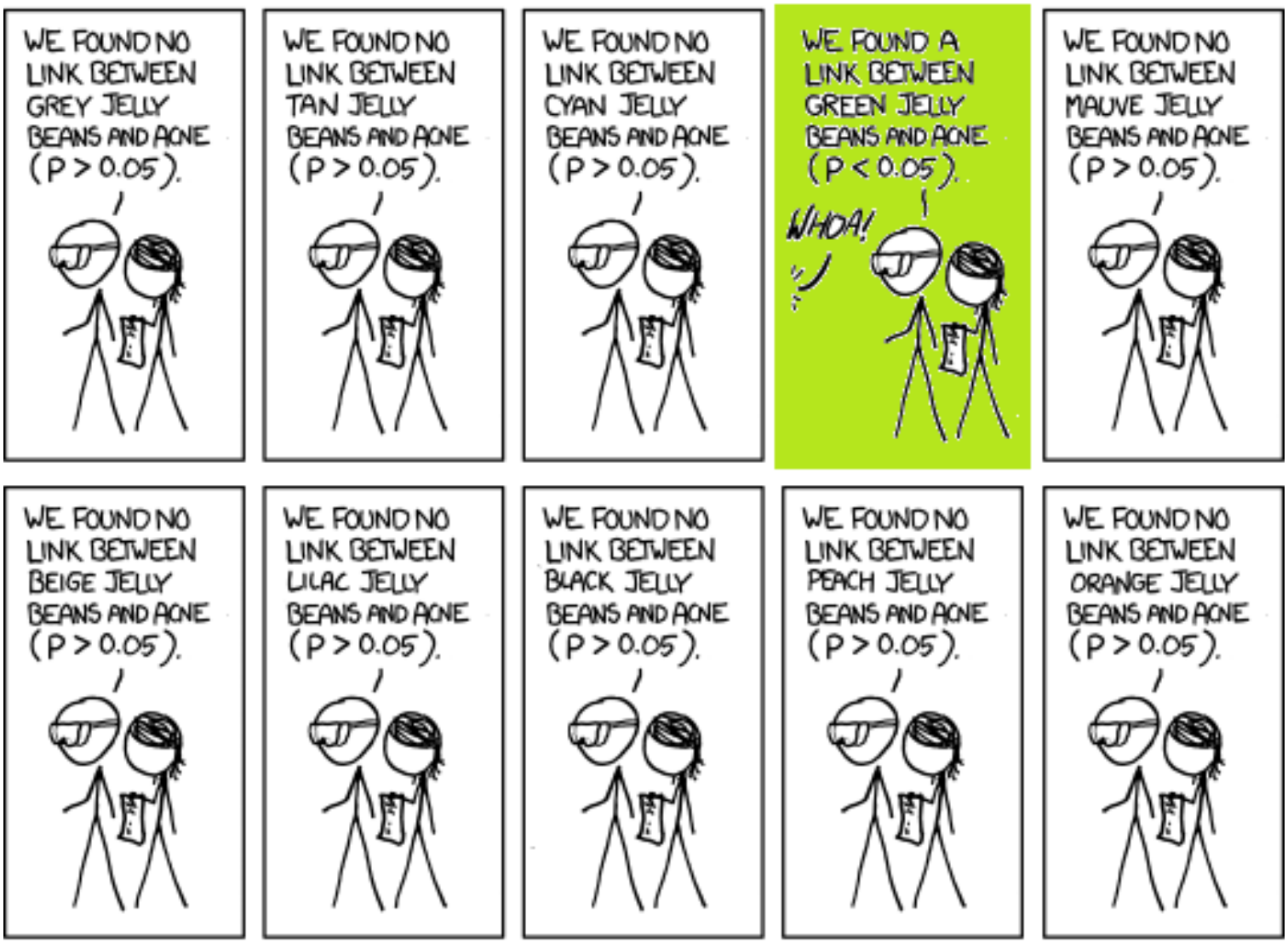
Example of lumping splitting (a)



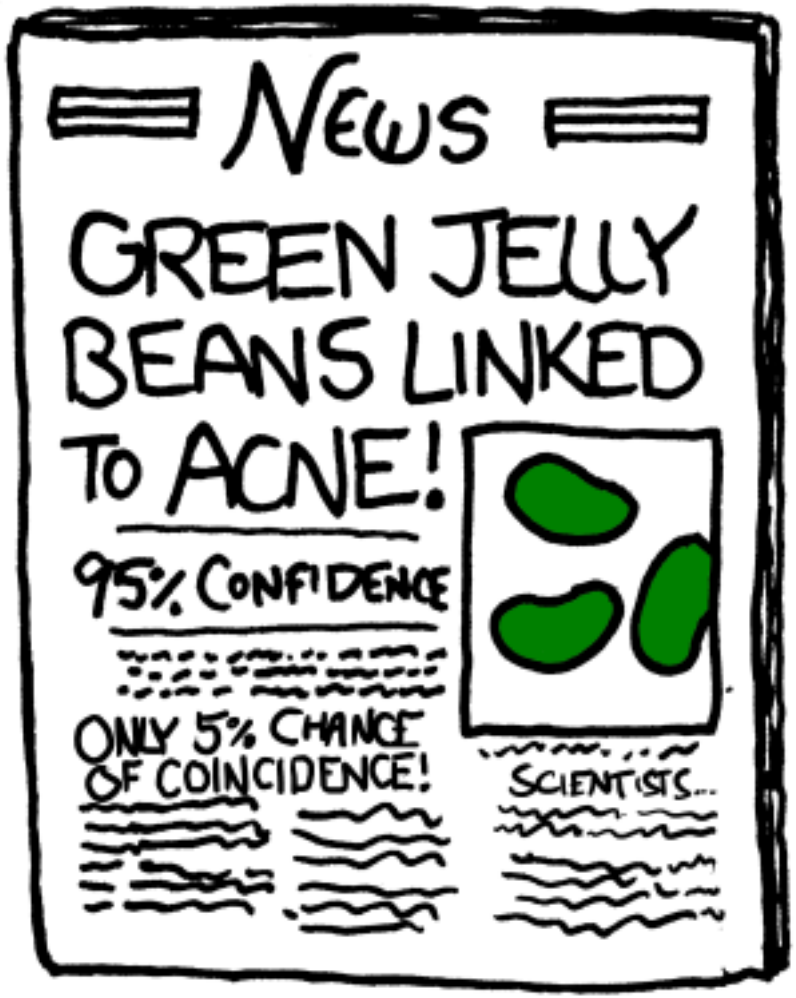
Example of lumping splitting (b)



Example of lumping splitting (c)



Example of lumping splitting (d)



Green jelly beans discussion

- When searching, the probability of false discoveries goes up
  - The chance of observing some p-values less than .05 increases (and approaches probability one), **even when every null hypothesis is true**.
- This is the problem of **"multiple comparisons"**.
  - Which is synonymous with **"multiple hypothesis testing"** or **"multiplicity"**.
- *But if you do not search enough you may not find what you really want.*

Adjustments to limit false-positives 1/2

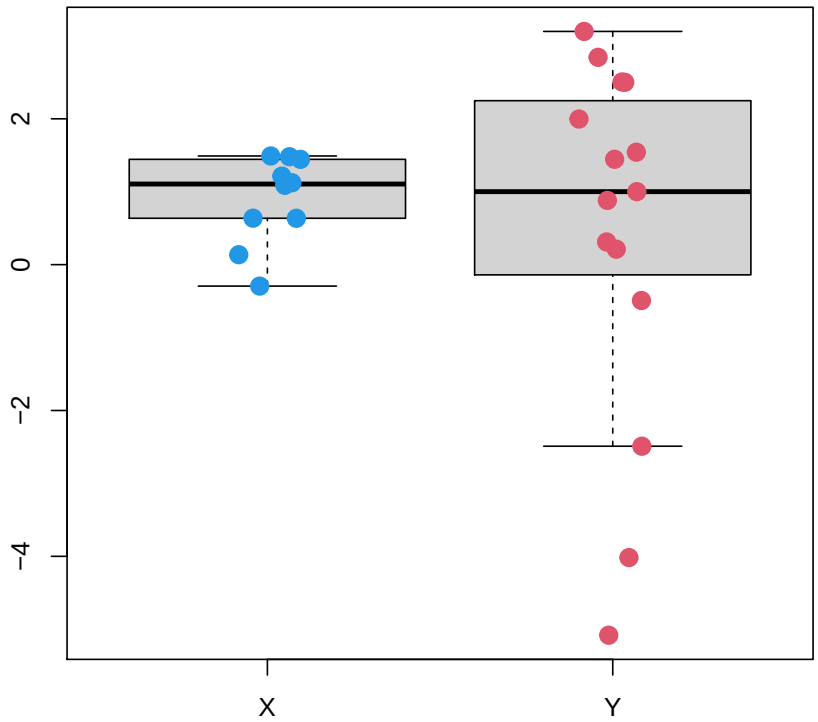
- *"One way around this problem [of multiple comparisons] is to demand a very low p-value at which significance is declared, and the simplest method, known as the **Bonferroni** correction, is to use the threshold of .05/n, where n is the number of tests done."* (Spiegelhalter, 2019, page 28).
- *"Another way to avoid false-positives is to demand **replication** of the original study ... "* ibid.
- Also see "False Discovery Rate" or FDR analysis in Hunt (2020, section 4.2).

Adjustments to limit false-positives 2/2

- Making p-value cut-offs harsher deals with the *logical* problem of too many false-positives.
- But a *practical* problem remains: power is greatly reduced for tests with harsher p-value cut-offs, so it can be difficult to detect evidence against false null hypotheses.
  - How many false hypotheses did you expect in the first place?
- One minimalist option is to report raw p-values *plus* the nature and total number of all the tests you ran.

Use tables and charts to summarise

	X	Y
n	10.00	15.00
mean	0.90	0.42
sd	0.61	2.49
var	0.37	6.21
min	-0.29	-5.08
max	1.49	3.20
median	1.11	1.00
quantile .25	0.64	-0.14
quantile .5	1.11	1.00
quantile .75	1.39	2.25



But label your charts and tables!

	X	Y
n	10.00	15.00
mean	0.90	0.42
sd	0.61	2.49
var	0.37	6.21
min	-0.29	-5.08
max	1.49	3.20
median	1.11	1.00
quantile .25	0.64	-0.14
quantile .5	1.11	1.00
quantile .75	1.39	2.25

Table 6: Summary statistics for X and Y. The first row, n, denotes sample size. Quantiles are the same as percentiles (e.g. the .5 quantile is the median).

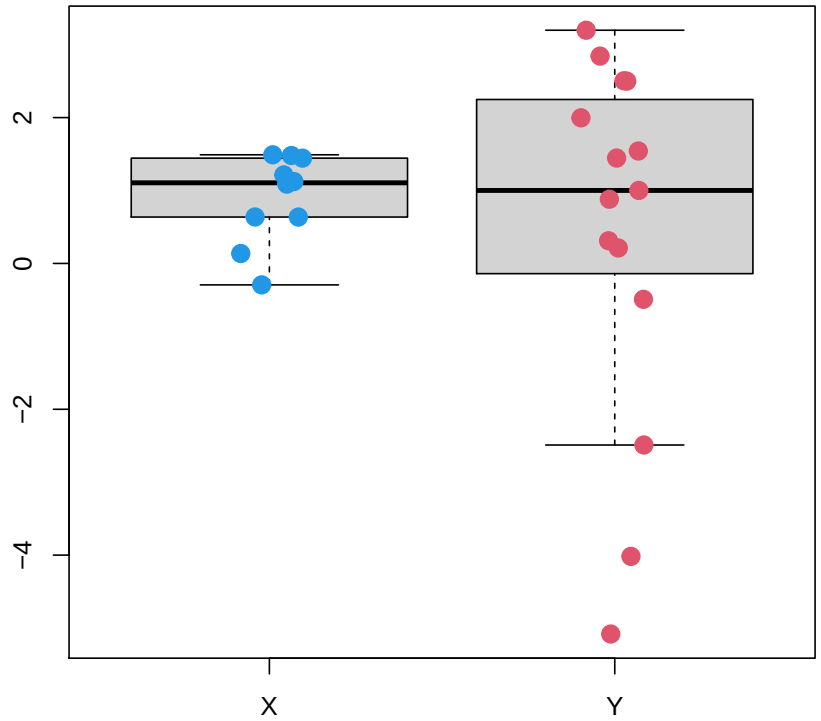


Figure 6: X and Y data. The central box covers the inter-quartile range (the middle 50% of observations) and marks the median (the thick line). Outliers are at least 1.5 times the length of inter-quartile range (which is delimited by dotted lines).

Be honest and uphold integrity

- Be explicit and precise about your hypotheses.
- Be prepared to share your data.
- Be honest about how many hypothesis tests you have done.
  - Including all data splitting and lumping!
- Acknowledge the key weaknesses in your conclusions.

Background knowledge and prior belief

- The hardest part of research is blending your prior belief and background knowledge ("the science", if you like) with empirical observations.
  - A strictly Bayesian approach requires you to blend explicit probability distributions – one for prior belief and the other for the likelihood of the data.
- Most researchers use a hybrid or ad-hoc approach that uses confidence intervals, data summaries and hypothesis tests.
  - Interesting and rigorous explanations of how the science links with empirical results is crucial.

Key words and reading

- Lumping and splitting.
- Multiple comparisons problem
  - Hunt (2020, sections 2.2, 4.1 and 4.2).
- Bonferroni correction.
  - Hunt (2020, section 4.2 and 4.1.3).
- Explicit Bayesian approaches.
  - Spiegelhalter (2019, chapter 11).

More important reading

- Claiming "discoveries" with statistics: Spiegelhalter (2019, chapters 10 and 14).
- Getting charts right: Cumming et al. (2007)
  - See David Vaux at `www.youtube.com/watch?v=BDArx6eecow` !
- Moving to a world beyond "$p < 0.05$": Wasserstein et al. (2019).

Quiz

Question: Which statements are incorrect?

A. The logical problem with multiple comparisons can be dissolved by adjusting a tests p-value cut-off (though the practical problem of lower power remains).
B. Do not split your data in too many ways because the multiple comparison problem makes interesting discoveries impossible.
C. The captions on charts and tables should be enough to enable a reader to understand the data presented within them.
D. p-values are more informative than confidence intervals.

Answer: B and D

Lecture outcomes

1. Understand the **foundations** of statistics and sampling variance.
2. Run **useful hypothesis tests**.
3. **Understand** p-values, test size, test power and confidence intervals.
4. Use "statistical analysis" for **research and critical thinking**.

Overall Summary

- Statistics is about learning from data, including estimating things and basic data summaries.
- Anything estimated from a sample, like a mean, has inherent error (sampling variance).
- Despite sampling variance, it is possible to test substantive hypotheses using data.
- Tests need to have controlled false-positive error rates (size) and the ability to produce evidence against false hypotheses (power).
- The key to good research is linking background knowledge, science and prior belief with empirical observations.

Bibliography I

Amrhein, V., S. Greenland, and B. McShane (2019). Scientists rise up against statistical significance. *Nature*, 305–307.

Cumming, G., F. Fidler, and D. L. Vaux (2007). Error bars in experimental biology. *The Journal of cell biology 177*(1), 7–11.

Flanagan, O. (2015). Statistics and the law: a recent history of an uneasy relationship. *Statslife (Royal Statistical Society)*.

Good, I. J. (1983). *Good Thinking. The Foundations of Probability Theory and Its Applications*. University of Minnesota Press.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician 69*(4), 371–386.

Hunt, I. (2020). *Introduction to Statistcs*. Ian Hunt (Monash University).

Locke, J. (1690). *An Essay Concerning Human Understanding*.

Bibliography II

Schwen, L. O. and S. Rueschenbaum (2018). Ten quick tips for getting the most scientific value out of numerical data. *PLoS computational biology 14*(10), e1006141.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Penguin UK.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p< 0.05".