Session 5: Hypothesis Testing

Kostis Christodoulou
London Business School

Contents

- Review of Sessions 3-4
- Hypothesis testing: ruling out chance
- Testing for differences in populations
- Hypothesis testing using **infer**

Inference

	Population Parameter	Sample statistic/ point estimate
Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Difference between proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Difference between means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Intercept	β_0	$\hat{\beta}_0$
Slope	β_1	$\hat{\beta}_1$
Standard deviation	σ	s

Use sample statistics to ***infer***, or make conclusions, about the underlying population parameters

Inferential Statistics Overview

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** (sample mean, sample median) as **point estimates** for the unknown population parameters of interest.
- **Sample statistics** vary from sample to sample. If we take two samples, say those to the left or those to the right of the class, we will not get exactly the same sample statistics
- Quantifying how sample statistics vary provides a way to estimate **the margin of error** associated with our point estimate.

What do you want to do?

- Estimation -> Confidence intervals
- Decision -> Hypothesis test

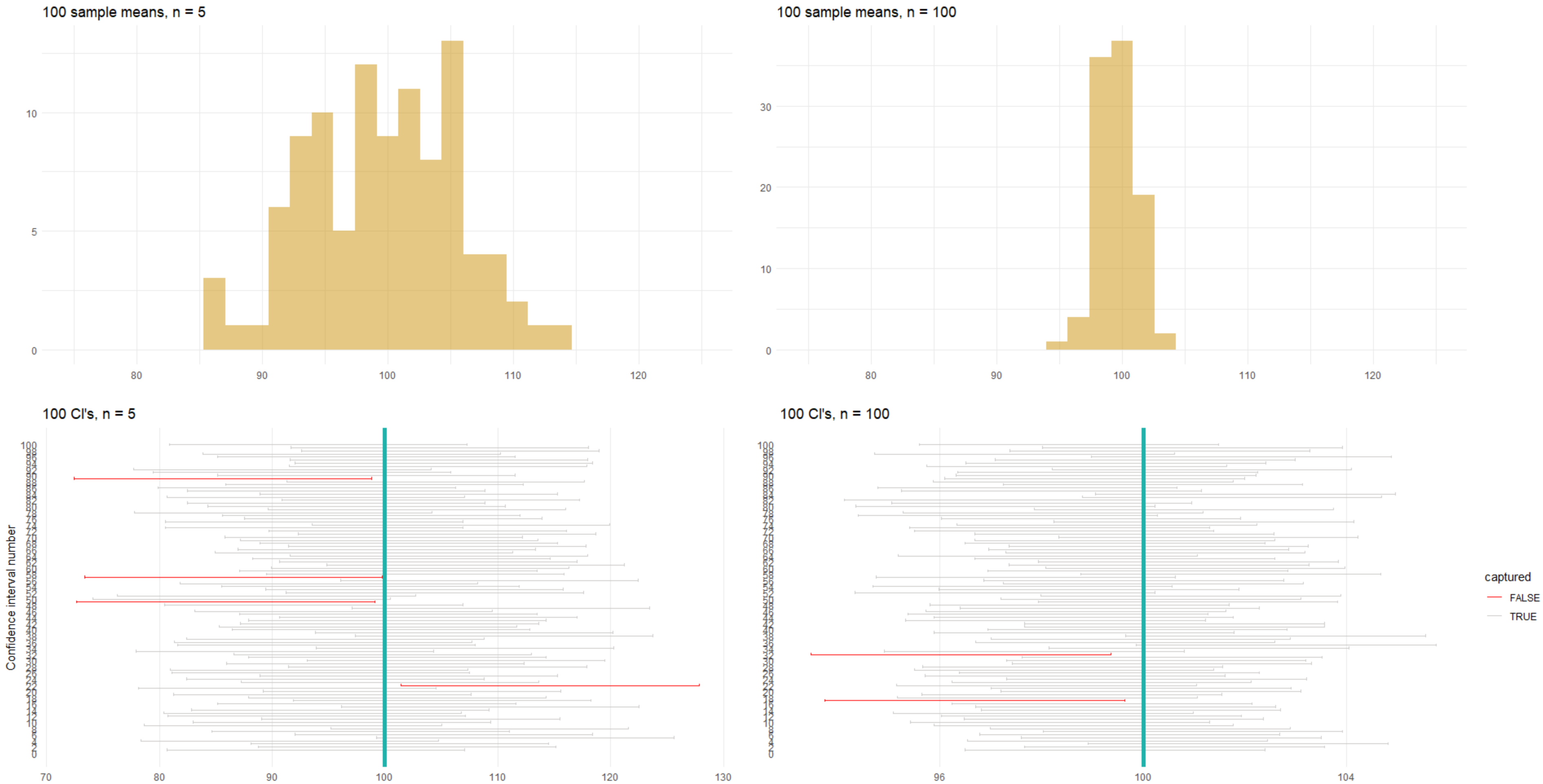
First step: Ask the following questions

- What is the (research) question you want to answer?
- How many and what types of variables?
- What happens to your sample statistic/point estimate as you increase the size of the sample?

Approximation formulas for Inference

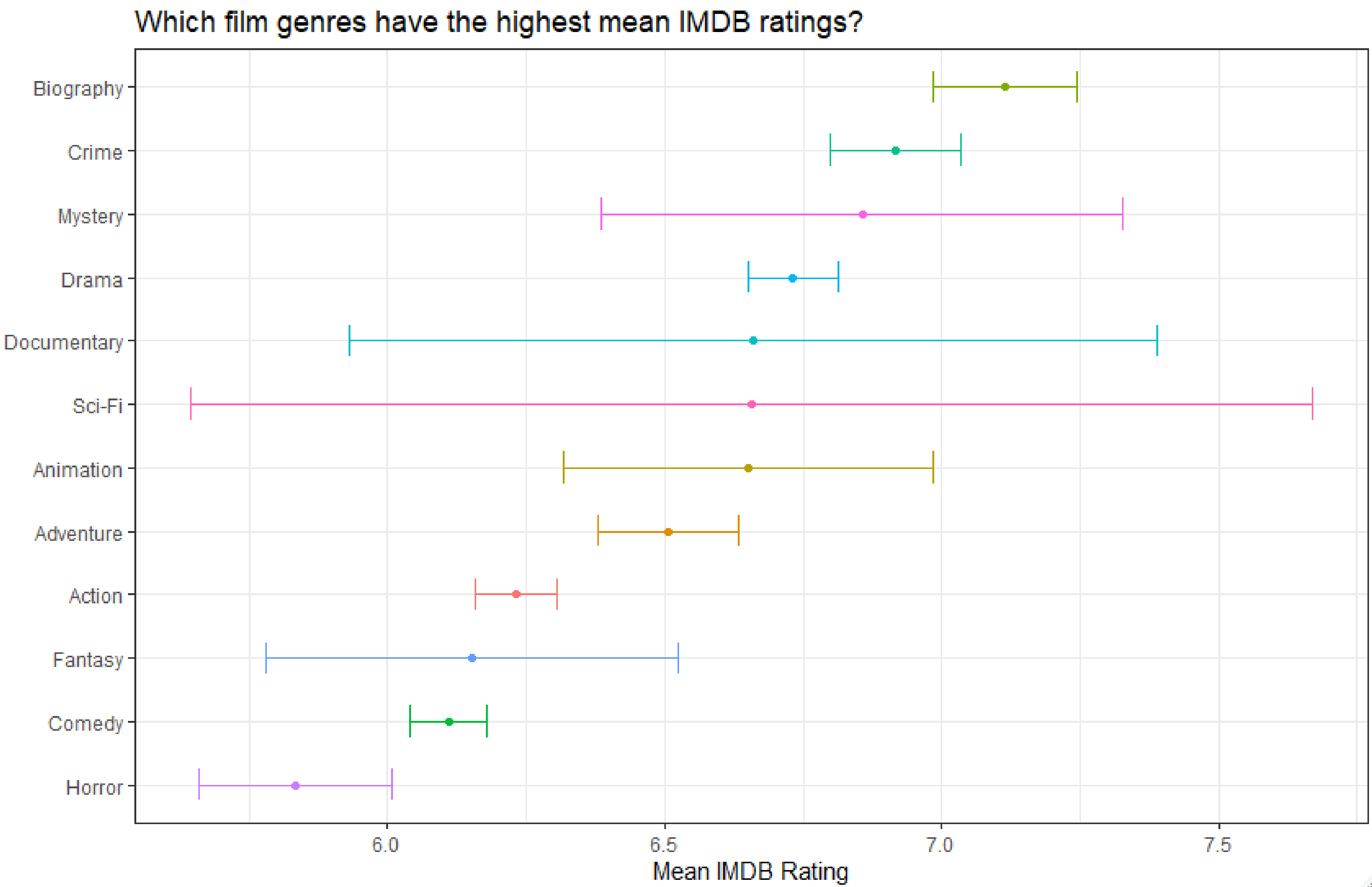
Parameter	Distribution	Conditions	Standard Error
Proportion	Normal	All counts at least 10 $np \geq 10$, $n(1-p) \geq 10$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference in Proportions	Normal	All counts at least 10 $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2 p_2 \geq 10$, $n_2(1-p_2) \geq 10$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Mean	t, df = $n-1$	$n \geq 30$ or data normal	$\sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$
Difference in Means	t, df = smaller of $n_1 - 1$, $n_2 - 1$	$n_1 \geq 30$ or data normal, $n_2 \geq 30$ or data normal	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Central Limit Theorem



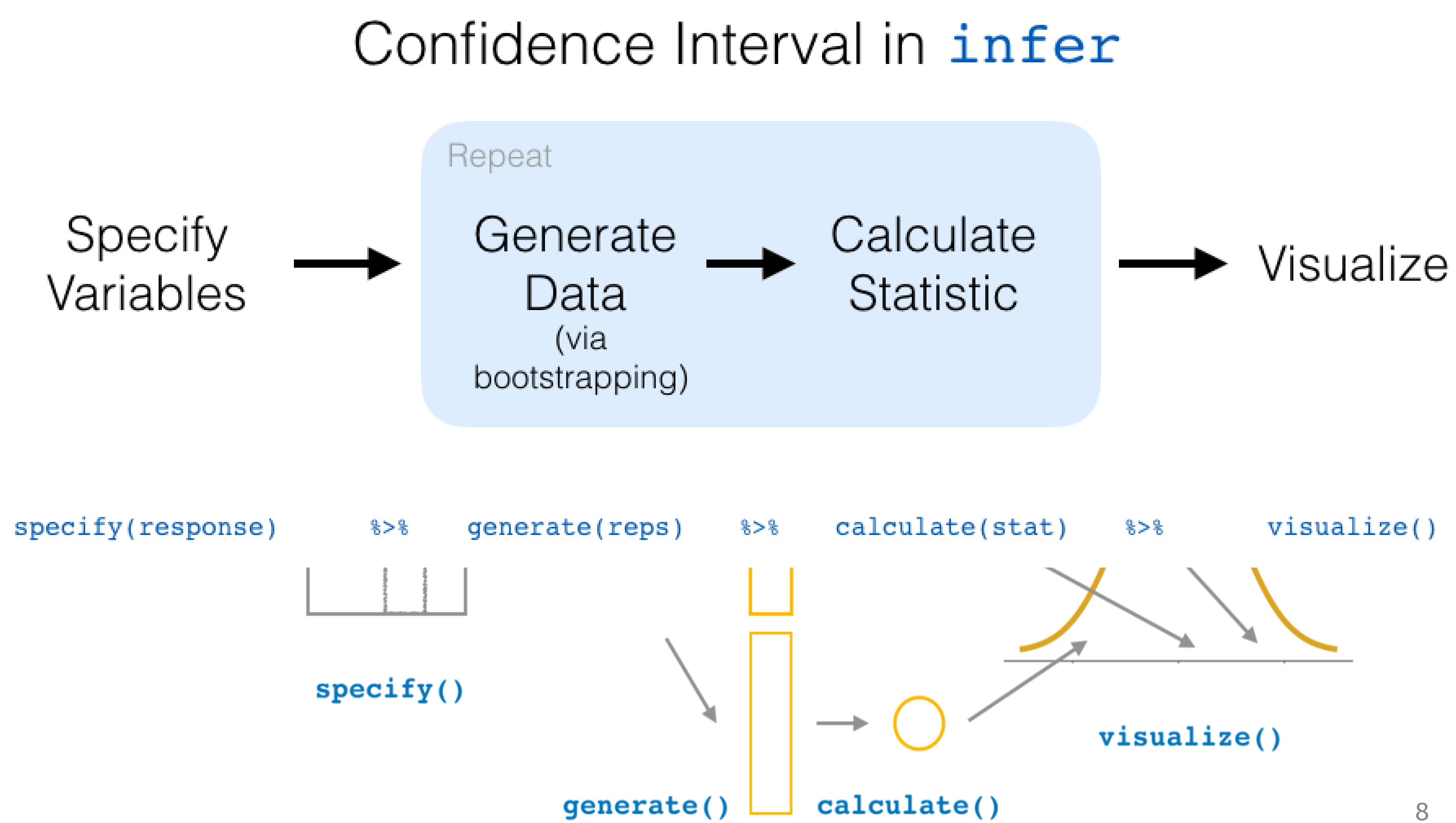
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Visualisation of CIs derived using formula

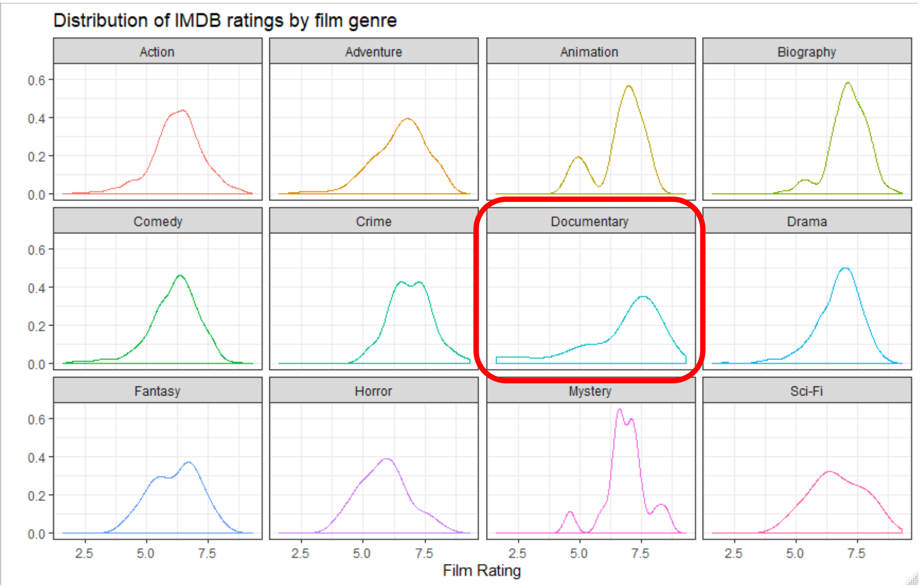


Bootstrapping with infer

```r
library(infer)
```



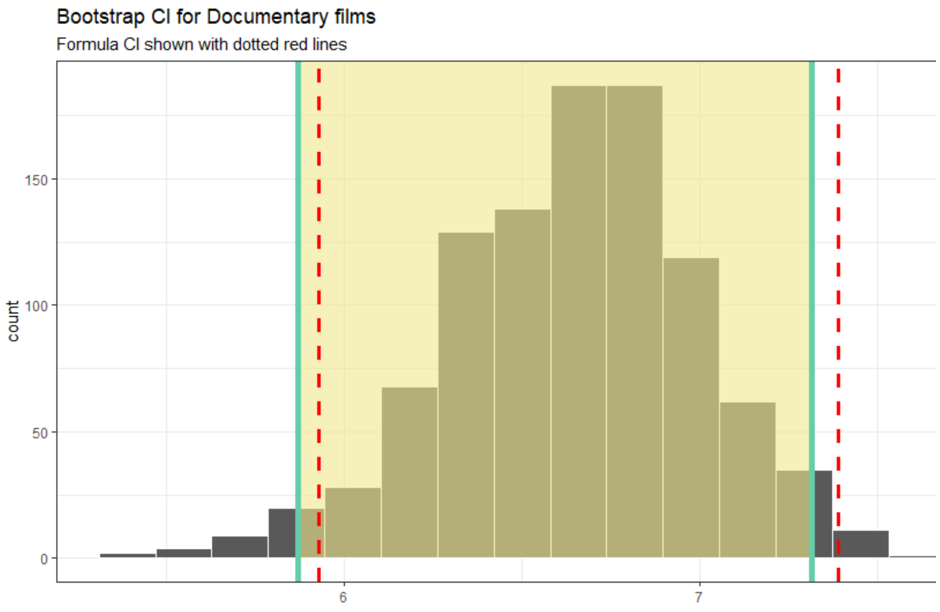
What about *genre==Documentary*

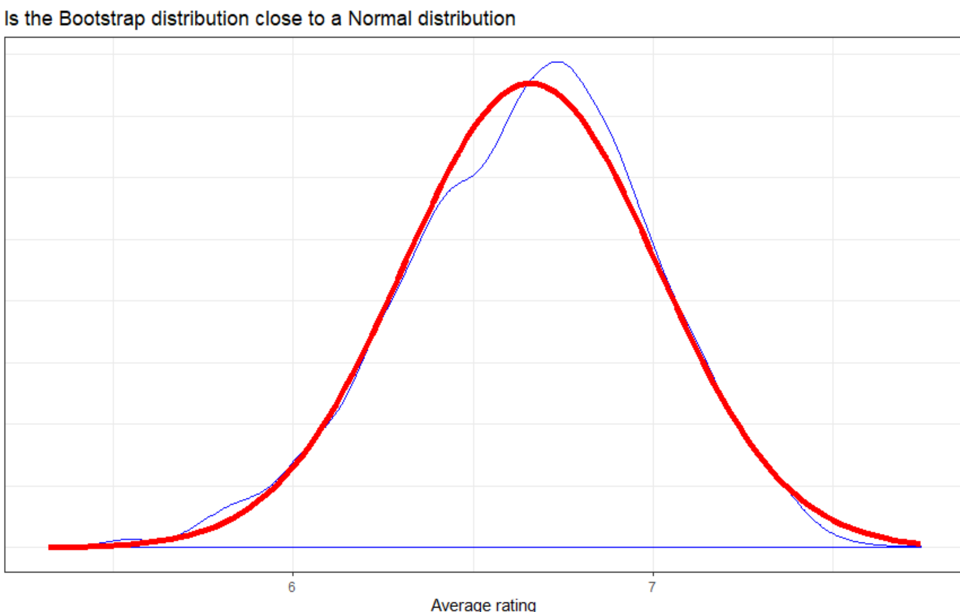


Distribution for ***genre==Documentary*** is heavily left skewed.

Will the bootstrap and formula CIs be similar?

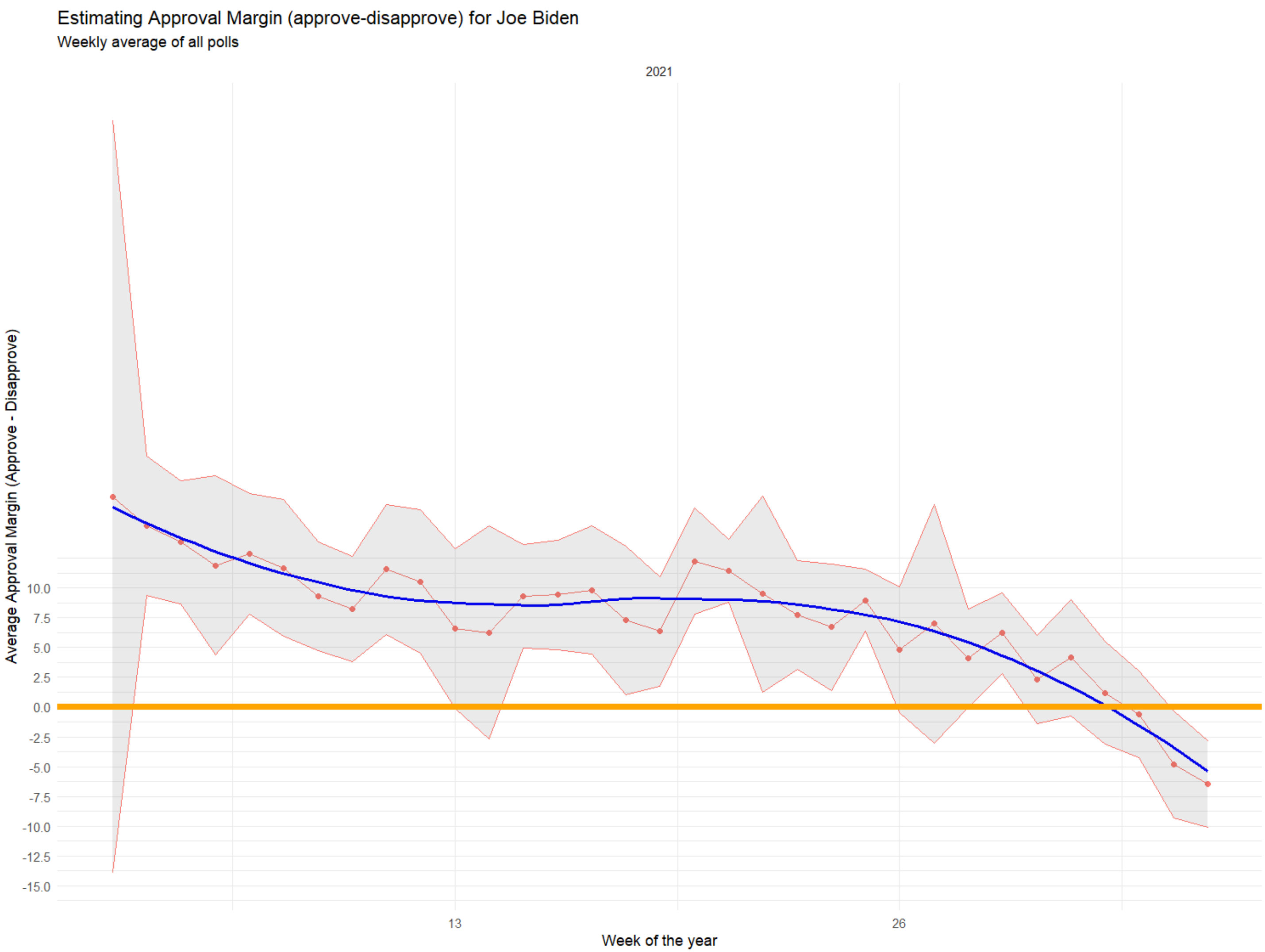
```
> percentile_ci
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   5.87    7.32
> formula_ci %>%
    select(rating_low, rating_high)
# A tibble: 1 x 2
  rating_low rating_high
       <dbl>       <dbl>
1       5.93        7.39
```



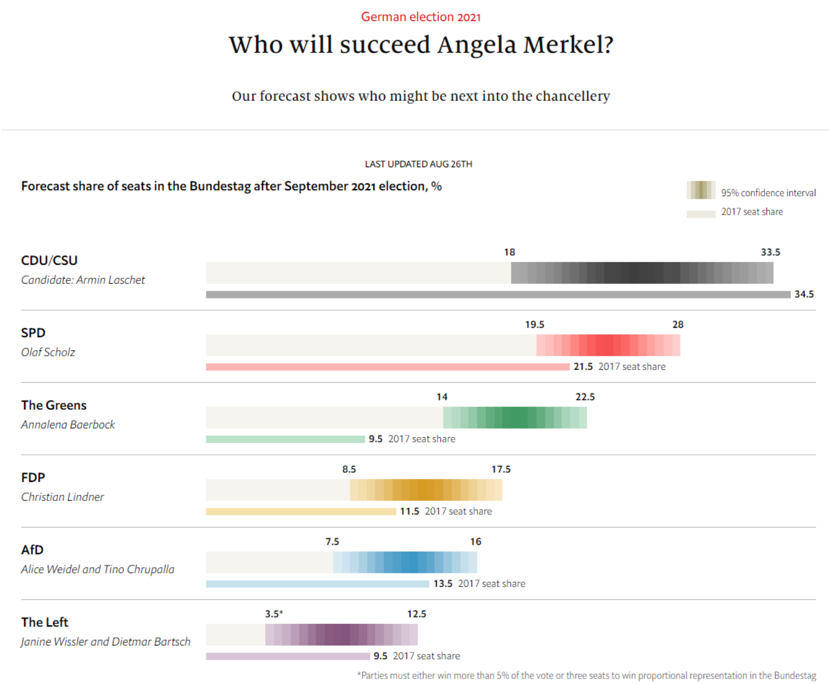


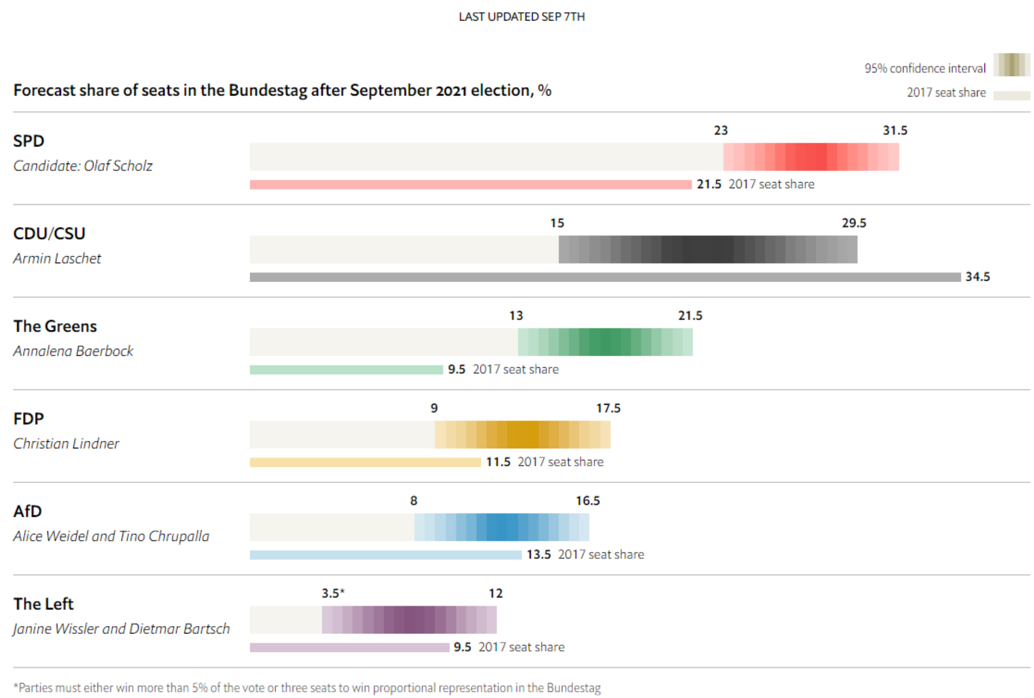
Net approval for Joe Biden

Don’t just show point estimate (net approval mean)
Confidence interval around sample mean visualises uncertainty



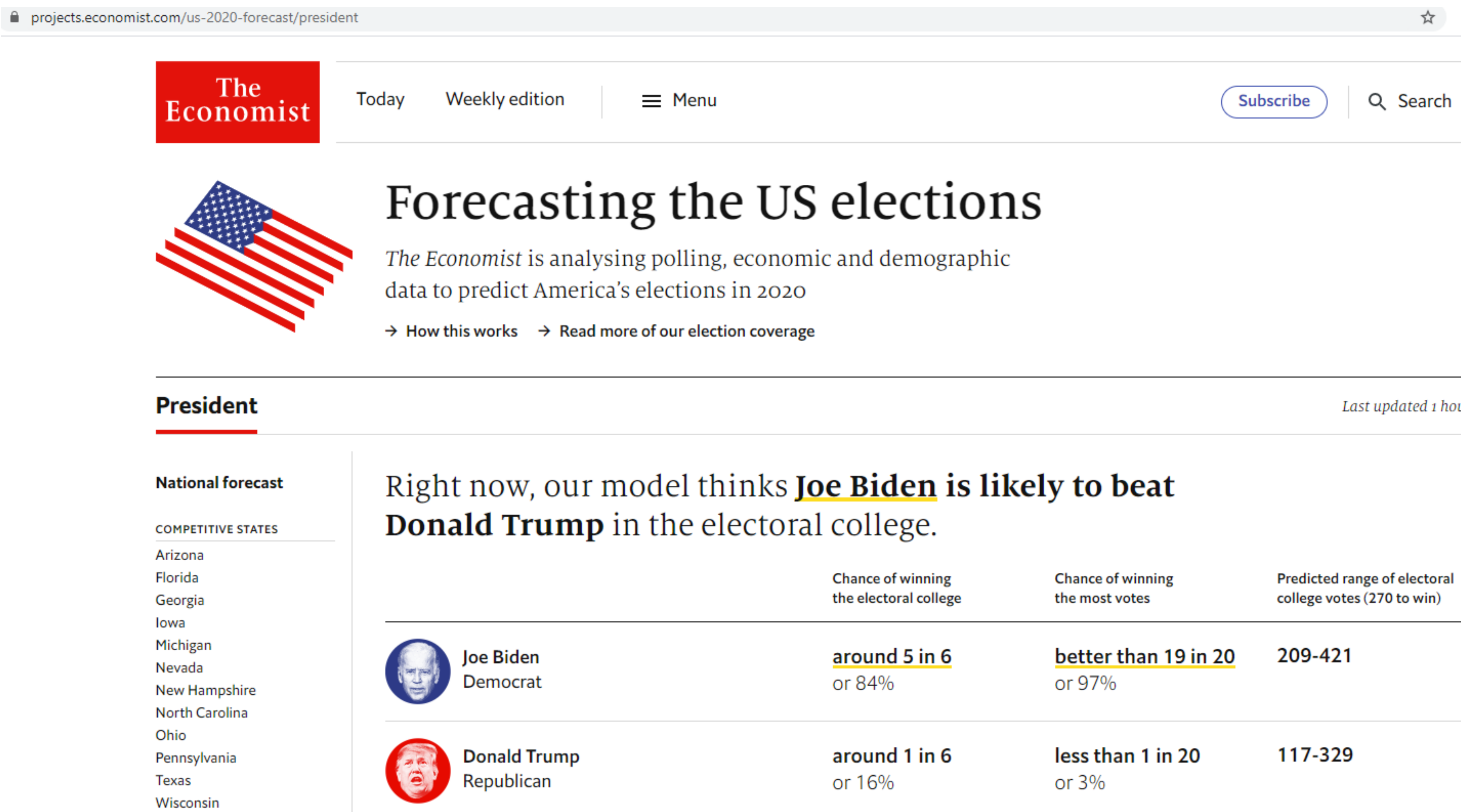
2021 German Federal election





https://www.economist.com/graphic-detail/who-will-succeed-angela-merkel/

State and national presidential election forecasting model


projects.economist.com/us-2020-forecast/president

The Economist

Today | Weekly edition | Menu | Subscribe | Search

Forecasting the US elections

The Economist is analysing polling, economic and demographic data to predict America's elections in 2020

→ How this works → Read more of our election coverage

President

Last updated 1 hou

National forecast

COMPETITIVE STATES

Arizona
Florida
Georgia
Iowa
Michigan
Nevada
New Hampshire
North Carolina
Ohio
Pennsylvania
Texas
Wisconsin

Right now, our model thinks **Joe Biden is likely to beat Donald Trump** in the electoral college.

	Chance of winning the electoral college	Chance of winning the most votes	Predicted range of electoral college votes (270 to win)
Joe Biden Democrat	around 5 in 6 or 84%	better than 19 in 20 or 97%	209-421
Donald Trump Republican	around 1 in 6 or 16%	less than 1 in 20 or 3%	117-329

https://github.com/TheEconomist/us-potus-model

Contents

- Review of Sessions 3-4
- Hypothesis testing: ruling out chance
- Testing for differences in populations
- Hypothesis testing using **infer**

Statistical Decisions, a.k.a Hypothesis Testing

- In statistics we check how surprising is the data we observed/measured from a sample, given an a-priori (or null) hypothesis we believe.
- If, given the null hypothesis, the data we observe is very improbable or surprising, then we reject the null hypothesis
- How improbable is improbable?
  - If the probability for an observed value is below the critical value of 5%, we reject the null hypothesis
  - We call this threshold the critical value or the alpha level .
  - The 5% hurdle rate is chosen arbitrarily and comes from Neyman and Pearson.
  - In medicine, an alpha level of 1% or less is often assumed. In physics you can sometimes find a lot of extreme p-values (p <.00001). In social research, however, an alpha level of 5% has become established.
- We make statements about the data and not about the hypotheses.

https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/testing-of-statistical-hypotheses-in-relation-to-probabilities-a-priori/65C6E3D534996282114D4E16FCA3E73C

Shaken, not stirred

Can James Bond tell the difference between a shaken and a stirred martini?

- We give Mr. Bond a series of 16 taste tests. In each test, we flip a fair coin to determine whether to stir or shake the martini, and ask Mr. Bond to taste it and decide whether it was shaken or stirred.

- If Mr. Bond was correct on 13 of the 16 tastings, does this prove that Mr. Bond can tell whether the martini was shaken or stirred?

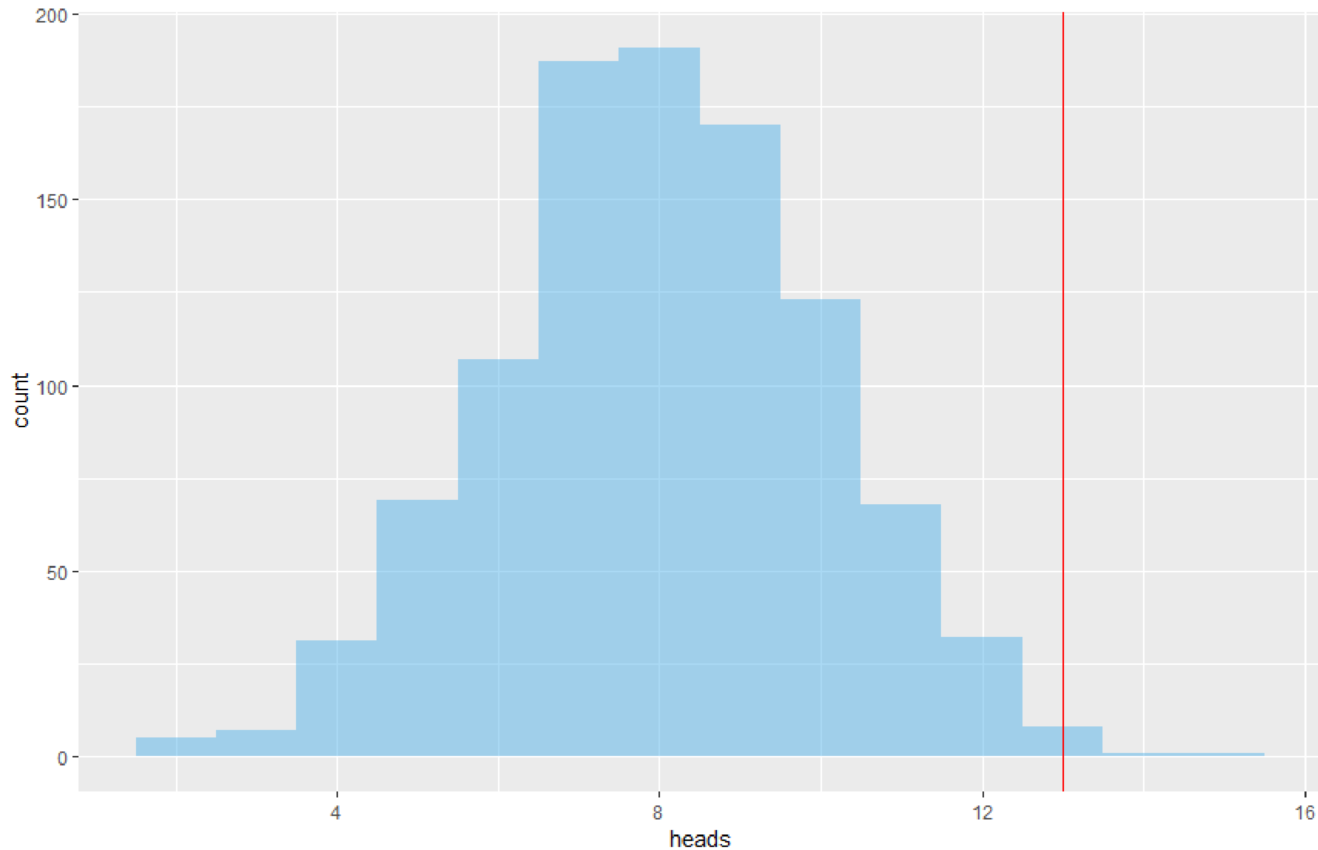
- If someone was clueless and just guessing, then he would have a 50% chance of getting it right in each test. So what is the probability that someone who is clueless would be correct 13/16 times or more?

- ***Can we rule out chance? Is 13/16 a surprising effect or was he just lucky?***

Shaken, not stirred - Simulated

- We can use simulation to see whether we get a similar result
- **rflip(16)** simulates Mr Bond tasting 16 martinis, but we can repeat this process many times, with the **do()** command
- Again, if he is just guessing by flipping a coin, how likely is it to call correctly 13 or more?

```
> Sims <- do (5000) * rflip(16)
> tally ( ~ (heads >= 13), data = Sims, format="prop")
(heads >= 13)
  TRUE  FALSE
0.0096 0.9904
```



Alternative Approach to Hypothesis Testing

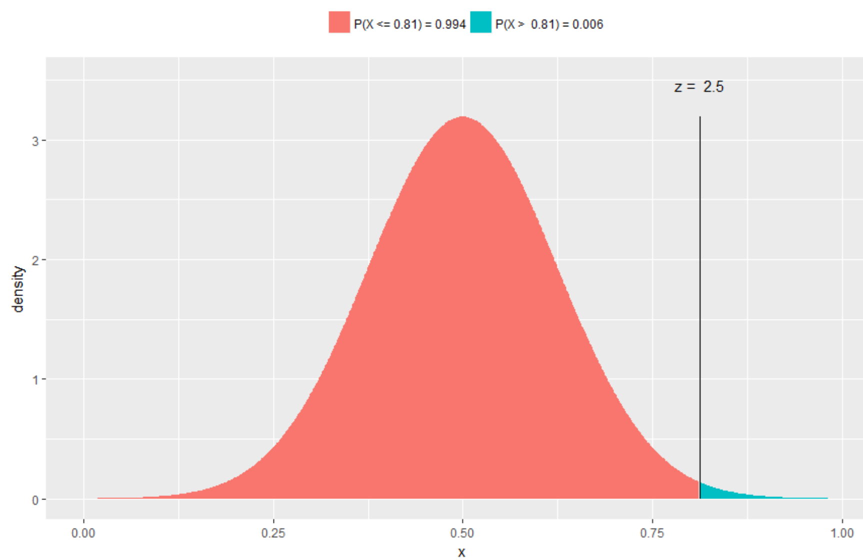
An alternative approach to hypothesis testing is based on the idea of a “p-value”:

We assumed p_0 =0.50, but we actually observed 13/16, or $\hat{p} = \frac{13}{16} = 0.8125$

How far away is what we **observed** (13/16 = 0.8125) from what we **assumed** (0.50)? What is the

$$Test\ statistic = \frac{signal}{noise} = \frac{(0.8125 - 0.50)}{SE} = \frac{0.3125}{\sqrt{\frac{0.5 * 0.5}{16}}} = \frac{0.3125}{0.125} = 2.5$$

```
>  xpnorm(0.8125, mean = 0.5, sd = 0.125)

If X ~ N(0.5, 0.125), then

        P(X <= 0.8125) = P(Z <= 2.5) = 0.9938
        P(X >  0.8125) = P(Z >  2.5) = 0.00621

[1] 0.9937903
```



the “p-value” is the chance of a result this far from the mean, if the null hypothesis were true
(i.e. the area represented by the shaded region [both tails for a 2-tailed test, just one for a 1-tailed test])

if the p-value is very low (typically less than 0.05 or 5%) then the difference is unlikely to be due to sampling error alone, so we reject the null hypothesis

Hypothesis testing example (1/2)

Does driving while sending messages increase the risk of a car accident?

Look at the difference between the two conditions (those who send messages while driving and have a car accident versus those who don't send messages and still have accidents).

You measure a difference of let's say 0.08 (or 8%).

Is this difference

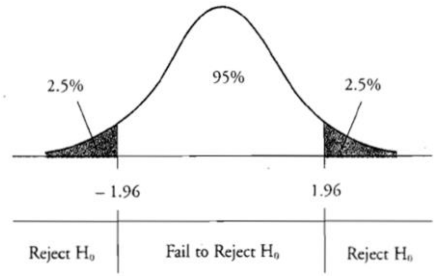
- Probably just some random noise
- Probably a real difference

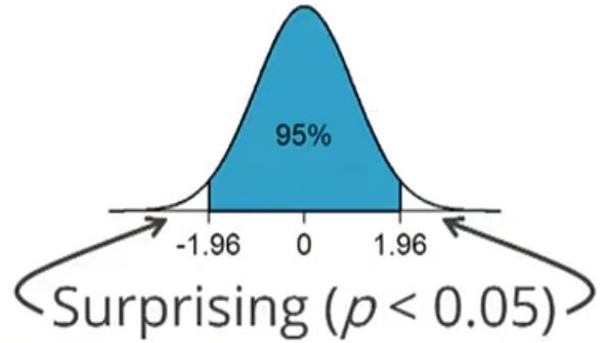
Start with a theory that there is no effect, i.e., difference in means is = 0. The best way to support your theory is try to disprove it, to play devil's advocate

Calculate a test-statistic, from the mean, SD and N and compare it against the Normal distribution, typically centered around zero.

Assuming the null hypothesis that there is no difference is true, data that is 'surprising' if what we measured and is beyond +-1.96 SE's.

Hypothesis testing example (2/2)





p-values tell you how surprising the data is, assuming that the null hypothesis there is no effect is true.

A *p-value* is the probability of getting the observed or more extreme data, assuming the null hypothesis is true.
A *p-value* is the probability of the data, not the probability of a theory. It doesn't mean you have a (1-alpha) probability that your theory is correct.

If the *p-value* <0.05, an effect is not 95% likely to be true.
A *p-value* > 0.05 does not mean there is no true effect; it means that the data we have observed is not surprising. You need large samples to detect small effects.

Just like in confidence intervals, think of *p-values* as a rule to guide behaviour in the long run.

- *p-value* < alpha: Act as if data is not noise. Invitation to explore effect further, it cannot by itself be enough to declare a scientific fact
- *p-value* : remain uncertain or act as if data were noise

Hypothesis Testing - Court Analogy

In the UK, the defendant is presumed ***not guilty***.
Only ***STRONG EVIDENCE*** to the contrary causes the ***not guilty*** claim to be rejected in favour of a ***guilty*** verdict. The phrase "beyond reasonable doubt" is often used to set the cutoff value for when enough evidence has been given to convict.

We should never say "The person is innocent", but instead "There is not sufficient evidence to show that the person is guilty."

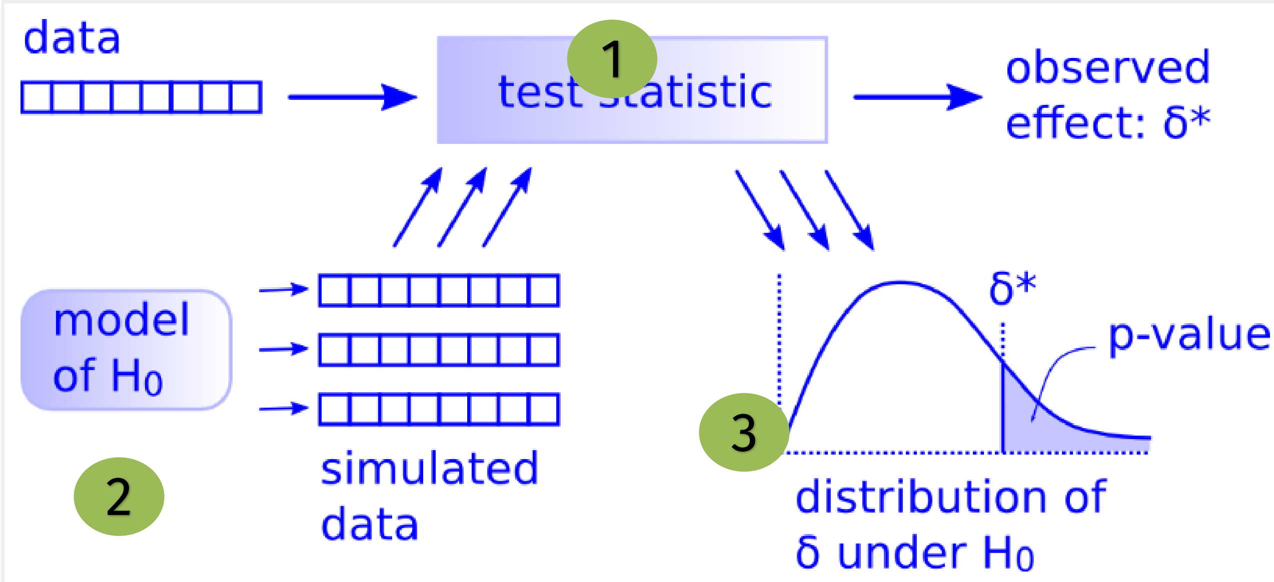
Now let's compare that to how we look at a hypothesis test. The decision about the population parameter(s) must be judged to follow one of two hypotheses.

We initially assume that H0 is true.
The null hypothesis H0 will be rejected (in favour of an alternative hypothesis H1 or Ha) only if the sample evidence strongly suggests that H0 is false. If the sample does not provide such evidence, H0 cannot be rejected.

The analogy to *beyond a reasonable doubt* in hypothesis testing is what is known as the significance level. This will be set before conducting the hypothesis test and is denoted as α .

Common values for α are 0.05 (5%) and 0.01 (1%).
Therefore, we have two possible conclusions with hypothesis testing:
Reject H0 and Fail to reject H0

Inference framework



1. Given a dataset, compute a test statistic that measures the size of the apparent effect. For example, if you are describing a difference between two groups, the test statistic might be the absolute difference in means. Let us call the test statistic from the observed data δ^*.

2. Next, you define a null hypothesis, which is a model of the world under the assumption that the effect is not real; for example, if you think there might be a difference between two groups, the null hypothesis assumes no difference.

3. Compute a p-value, the probability of seeing an effect as big as δ^* *if the null hypothesis were true*. If the p-value is sufficiently small, you can conclude that the apparent effect is unlikely to be due to chance.

Source: Allen Downey https://allendowney.blogspot.com/2016/06/there-is-still-only-one-test.html

Contents

- Review of Sessions 3-4
- Hypothesis testing: ruling out chance
- Testing for differences in populations
- Hypothesis testing using **infer**

Testing for Difference

- Applications for testing differences between samples
  - Average running cost for different makes of vehicle
  - Average salary between different groups of employees
  - Difference in profits between regions, managers, etc.
- Chances are if we take two different samples there will be some difference
  - Is this due to **chance** alone (sampling error) or is there a **significant** difference?

Hypothesis Testing Recipe

1. Set up hypotheses.
   - Claim (null hypothesis): $H0: \delta = 0$
   - Alternative hypothesis: $H_a: \delta \neq 0$

2. Take a sample and calculate :
   - Sample Mean = δ^*, Standard error

3. Choose a significance level: typically 5%
   - Corresponds to 95% confidence (approximately $\pm$ 2 SEs)

4. Various possible methods to decide whether to reject or not:
   - Two separate CIs: If they don't overlap, reject. If they overlap, run t-test
   - CI for delta: If claim (zero) outside of confidence interval
   - t- stat > 2 (approximately)
   - p – value < 5%

Reject!

Comparisons of Two Means: Test Statistic Formula

Large-sample test statistic for the difference between two independent population means:

$$t - stat = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The term $(\mu_1 - \mu_2)_0$ is the difference between μ_1 an μ_2 under the null hypothesis. It is equal to zero in most situations, i.e., there is no difference between the two populations.

The term in the denominator is the standard deviation of the difference between the two sample means (it relies on the assumption that the two samples are independent). This test also assumes unequal variances.

Standard Error of Differences

- For large samples (by CLT), the standard error is

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

- The 95% Confidence Interval for the difference between means is approximately

$$(\bar{x}_1 - \bar{x}_2) \pm 2\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Differences in credit card balances

	income	limit	rating	cards	age	education	own	student	married	region	balance
1	14.891	3606	283	2	34	11	No	No	Yes	South	333
2	106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
3	104.593	7075	514	4	71	11	No	No	No	West	580
4	148.924	9504	681	3	36	11	Yes	No	No	West	964
5	55.882	4897	357	2	68	16	No	No	Yes	South	331
6	80.180	8047	569	4	77	10	No	No	No	South	1151
7	20.996	3388	259	2	37	12	Yes	No	No	East	203
8	71.408	7114	512	2	87	9	No	No	No	West	872
9	15.125	3300	266	5	66	13	Yes	No	No	South	279
10	71.061	6819	491	3	41	19	Yes	Yes	Yes	East	1350
11	63.095	8117	589	4	30	14	No	No	Yes	South	1407
12	15.045	1311	138	3	64	16	No	No	No	South	0
13	80.616	5308	394	1	57	7	Yes	No	Yes	West	204
14	43.682	6922	511	1	49	9	No	No	Yes	South	1081
15	19.144	3291	269	2	75	13	Yes	No	No	East	148
16	20.089	2525	200	3	57	15	Yes	No	Yes	East	0
17	53.598	3714	286	3	73	17	Yes	No	Yes	East	0
18	36.496	4378	339	3	69	15	Yes	No	Yes	West	368
19	49.570	6384	448	1	28	9	Yes	No	Yes	West	891
20	42.079	6626	479	2	44	9	No	No	No	West	1048
21	17.700	2860	235	4	63	16	Yes	No	No	West	89
22	37.348	6378	458	1	72	17	Yes	No	No	South	968

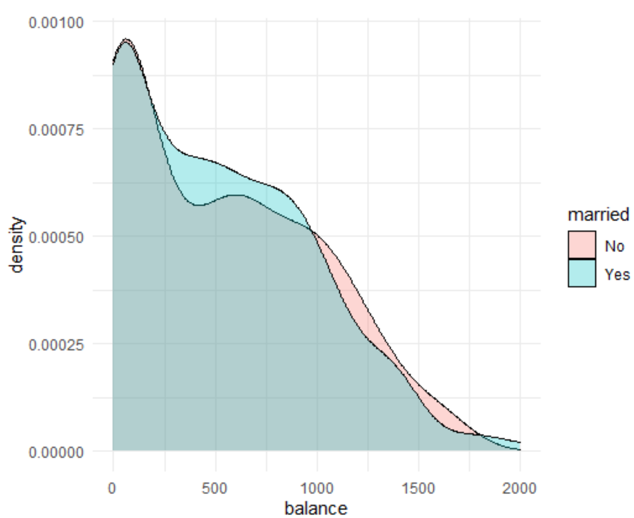
- Data on 400 customers with average ≈ 520$

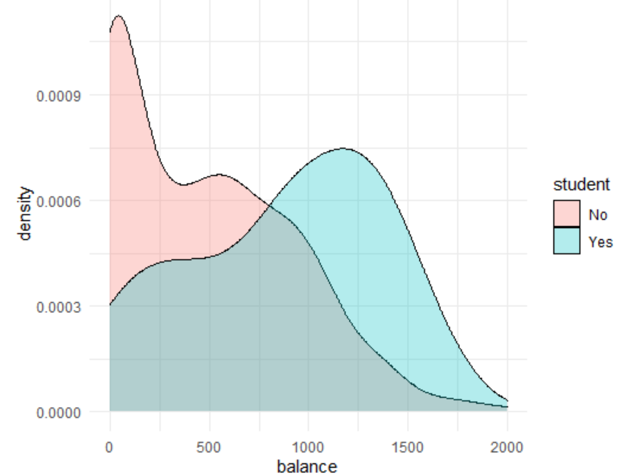
```
> mosaic::favstats(~balance, data=credit)
 min    Q1 median  Q3  max    mean       sd   n missing
   0 68.75  459.5 863 1999 520.015 459.7589 400       0
```

- Do married people have higher balances?
- Do students have higher balances?

```
> mosaic::favstats(balance ~ married, data = credit)
  married min Q1 median  Q3  max     mean       sd   n missing
1      No   0 51    467 910 1687 523.2903 470.8875 155       0
2     Yes   0 81    454 844 1999 517.9429 453.5380 245       0
> mosaic::favstats(balance ~ student, data = credit)
  student min     Q1 median     Q3  max     mean       sd   n missing
1      No   0  13.25    424  807.5 1999 480.3694 439.4145 360       0
2     Yes   0 428.00    953 1256.0 1687 876.8250 490.0020  40       0
```

- We want to look at the difference in the group … SE





Differences in credit card balances

```
> mosaic::favstats(balance ~ student, data = credit)
  student min     Q1 median     Q3  max     mean       sd   n
1      No   0  13.25    424  807.5 1999 480.3694 439.4145 360
2     Yes   0 428.00    953 1256.0 1687 876.8250 490.0020  40
```

$$t-stat = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{480.37 - 876.83}{\sqrt{\frac{439.42^2}{360} + \frac{490^2}{40}}}$$

$$= -4.90$$

- ***t-stat*** is the difference in the group means, measured in units of the SE
- t Stat value of -4.90 is greater than the critical value of 1.96 => reject H_0
- Conclude that the -396$ (480.37-876.38) difference we estimated in our sample means is really different from zero and therefore there is a significant difference

Hypothesis Testing in R

```
> t.test(balance ~ student, data = credit)

        Welch Two Sample t-test

data:  balance by student
t = -4.9028, df = 46.241, p-value = 0.00001205
alternative hypothesis: true difference in means between group No and group Yes is not equal
 to 0
95 percent confidence interval:
 -559.2023 -233.7088
sample estimates:
 mean in group No mean in group Yes
         480.3694          876.8250
```

Differences in haircut spend

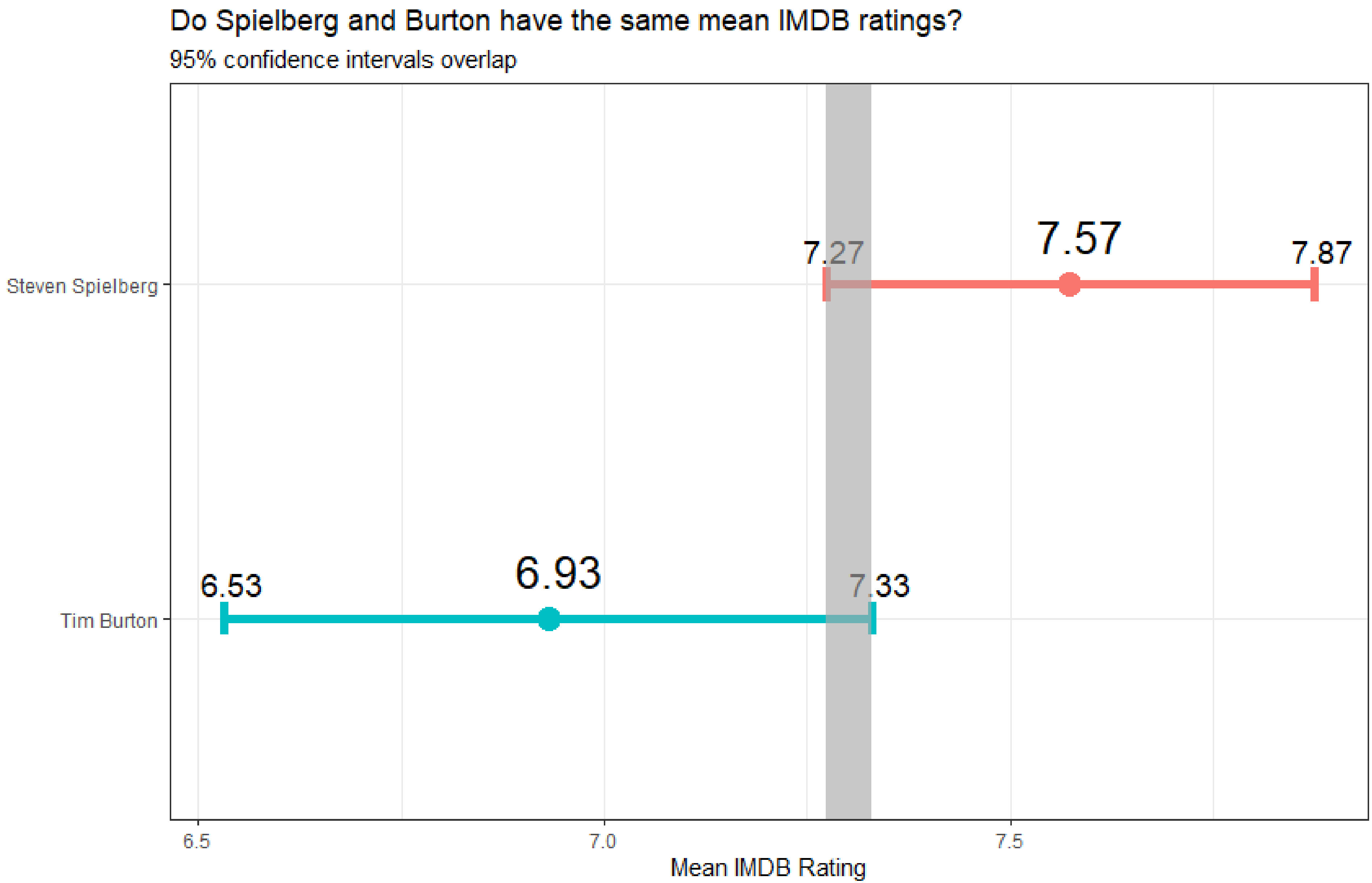
t-Test: Two-Sample Assuming Unequal Variances			
	haircutMen	*haircutWomen*	*Difference*
Mean	30.28	44.99	-14.71
Variance	550.9724873	1479.048264	
Observations	107	89	
Hypothesized Mean Difference	0		
df	140		
t Stat	-3.15		
P(T<=t) one-tail	0.10%		
t Critical one-tail	1.66		
P(T<=t) two-tail	0.20%		
t Critical two-tail	1.98		

$$t-stat = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

$$= \frac{30.28 - 44.99}{\sqrt{\frac{550.97}{107} + \frac{1479.05}{89}}}$$

$$= \frac{-14.71}{4.67} = -3.15$$

- "t Stat" is the difference in the group means, measured in units of the SE
- H_0: average male haircut - average female haircut ≤ 0 (one-tailed)
- t Stat value of -3.15 is greater than the 5% critical value of 1.66 => reject H_0,
- conclude that the 14.71$ difference we estimated in our sample is really different from zero and therefore there is a significant difference in haircut spend



Do Spielberg and Burton have the same mean IMDB ratings?
95% confidence intervals overlap
Steven Spielberg
Tim Burton
7.27
7.57
7.87
6.53
6.93
7.33
6.5
7.0
7.5
Mean IMDB Rating

Difference between two proportions

If we assume that both proportions are equal, then we would expect the difference ***(p1-p2)*** to be zero.

What about the standard error of the difference in the two proportions? This is a pooled variance calculation

$$SE_{diff} = \sqrt{SE_1{}^2 + SE_2{}^2} = \sqrt{\frac{\mathrm{p}_1 * (1 - \mathrm{p}_1)}{\mathrm{n}_1} + \frac{\mathrm{p}_2 * (1 - \mathrm{p}_2)}{\mathrm{n}_2}}$$

We can create a Confidence Interval (CI) for the difference.
The difference ***p1-p2*** follows a Normal distribution with a mean=0, as shown below

$$(p_1 - p_2) \sim \mathrm{N}\left(0, \sqrt{\frac{\mathrm{p}_1 * (1 - \mathrm{p}_1)}{\mathrm{n}_1} + \frac{\mathrm{p}_2 * (1 - \mathrm{p}_2)}{\mathrm{n}_2}}\right)$$

Contents

- Review of Sessions 3-4
- Hypothesis testing: ruling out chance
- Testing for differences in populations
  - Experiments
- Hypothesis testing using **infer**

https://www.theguardian.com/lifeandstyle/2017/nov/23/three-coffees-a-day-linked-to-a-range-of-health-benefits

Fashion **Food** Recipes Love & sex Health & fitness Home & garden Women Family Travel Money

Coffee

Three coffees a day linked to a range of health benefits

Research based on 200 previous studies worldwide says frequent drinkers less likely to get diabetes, heart disease, dementia and some cancers

Staff and agencies

Thu 23 Nov 2017 00.54 GMT

▲ The findings supported other studies showing the health benefits of drinking coffee. Photograph: Wu Hong/EPA

People who drink three to four cups of coffee a day are more likely to see health benefits than problems, experiencing lower risks of premature death and heart disease than those who abstain, scientists have said.

The research, which collated evidence from more than 200 previous studies, also found coffee consumption was linked to lower risks of diabetes, liver disease, dementia and some cancers

Source: https://www.theguardian.com/lifeandstyle/2017/nov/23/three-coffees-a-day-linked-to-a-range-of-health-benefits

Is there **any relation** between coffee consumption and heart disease?

association

- any relation
- link

*Drinking coffee was consistently **linked** with a lower risk of death from all causes and from heart disease. The largest reduction in relative risk of premature death is seen in people consuming three cups a day, compared with non-coffee drinkers.*

Does coffee consumption **lead to** a reduction in heart disease?

causality
This question is often harder to answer.

What does it take to prove a cause-and-effect relationship between coffee and reduced risk of heart disease?

Scope of Inference

Random sampling and/or random assignment

	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

- One of the early studies linking smoking and lung cancer compared patients already hospitalized with lung cancer to similar patients without lung cancer (hospitalized for other reasons), and recorded whether each patient smoked. Then, proportions of smokers for patients with and without lung cancer were compared.

- Does this study employ random sampling and/or random assignment?

Types of studies

1. Observational study: Observing what naturally goes on in the world without directly interfering with it. We can only infer correlation.

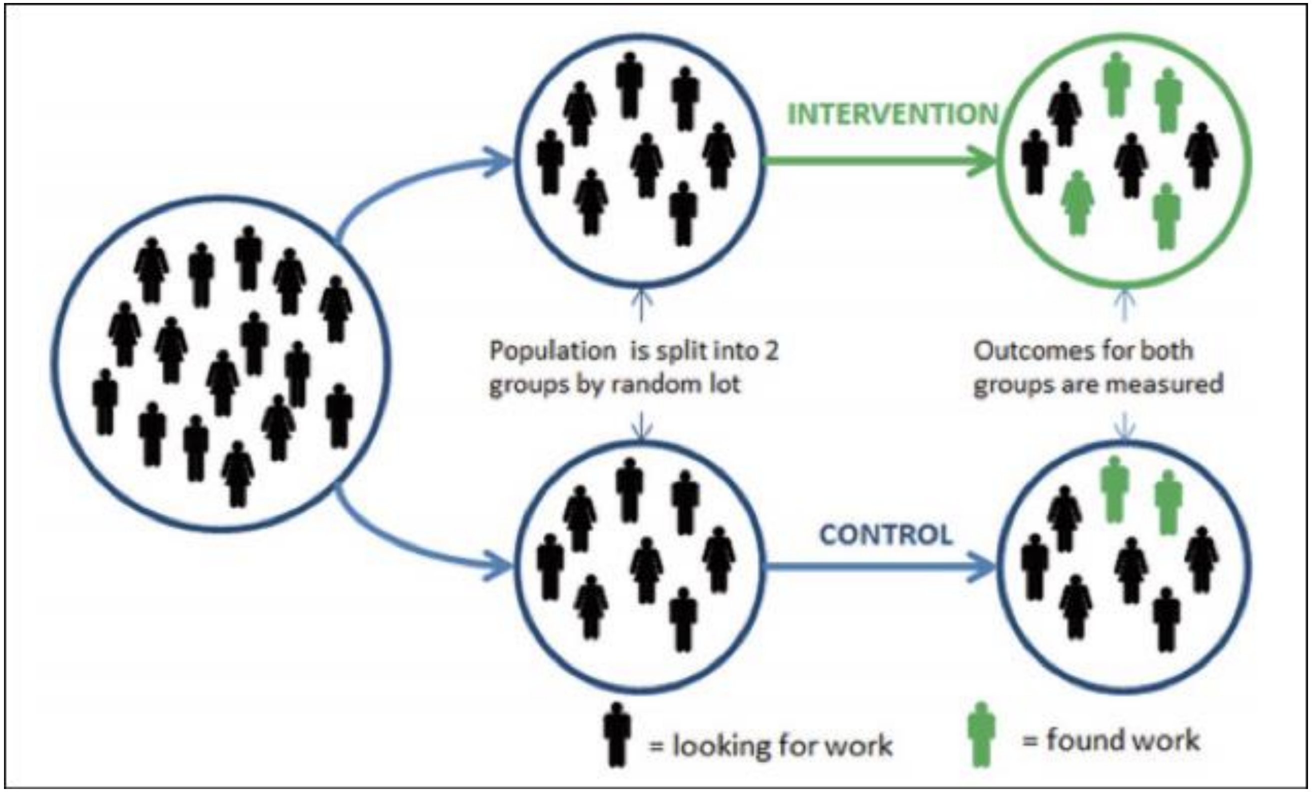
2. Experiment- Randomized Controlled Trial (RCT): One (or more) variable is systematically manipulated to see their effect (alone or in combination) on an outcome variable. Statements can be made about cause and effect.

   – In a controlled experiment, subjects are randomly assigned a treatment, and the effect of the treatment is examined. For example in a drug trial half of the subjects are randomly chosen and given the drug *(treatment group)*, while the other half are given a placebo *(control group)*.
   – Random assignment of the treatment ensures that the treatment is uncorrelated with any observables or unobservables → Causation can be inferred

The real purpose of the scientific method is to make sure nature hasn't misled you into thinking you know something you actually don't know. – Robert Pirsig, *Zen and the Art of Motorcycle Maintenance*

How RCTs work

Where feasible, randomised control trials (RCTs) are generally the most reliable tool we have for finding out which of two interventions works best. We simply take a group of people; we split them into two groups at random; we give one intervention to one group, and the other intervention to the other group; then we measure how each group is doing, to see if one intervention achieved its supposed outcome any better.



Contents

- Review of Sessions 3-4
- Hypothesis testing: ruling out chance
- Testing for differences in populations
- Hypothesis testing using **infer**

Hypothesis testing with the **infer** package

Step 1: Calculate a sample statistic, or ***δ****. This is the main measure you care about: the difference in means, the average, the median, the proportion, the difference in proportions, the chi-squared value, etc.

Step 2: Use simulation to invent a world where ***δ*** is null. Simulate what the world would look like if there was no difference between two groups, or if there was no difference in proportions, or where the average value is a specific number. Look at ***δ*** in the null world. Put the sample statistic in the null world and see if it fits well.

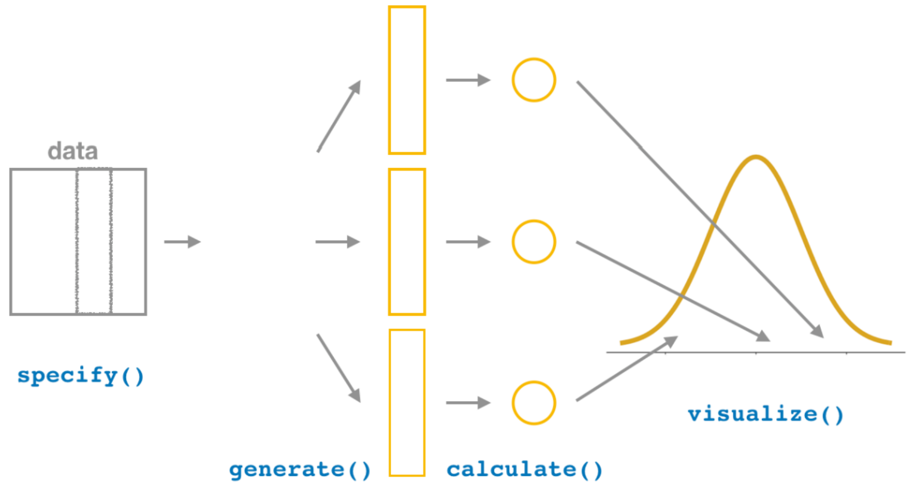
Step 3: Calculate the probability that ***δ*** could exist in null world. This is the ***p-value***, or the probability that you'd see a ***δ*** at least that high in a world where there's no difference.

Step 4: Decide if ***δ*** is statistically significant. Choose some evidentiary standard or threshold for deciding if there's sufficient proof for rejecting the null world. Standard thresholds (from least to most rigorous) are 0.1, 0.05, and 0.01.

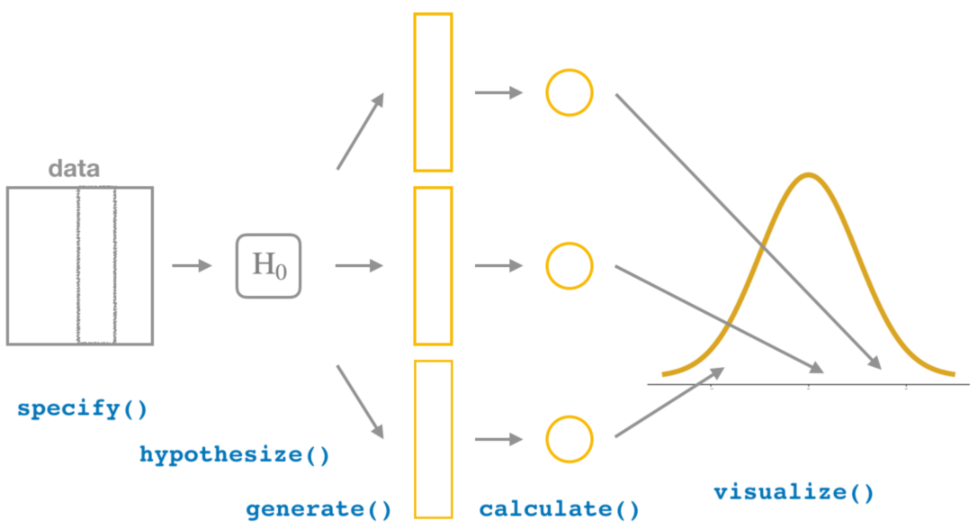
Bootstrap CIs, Hypothesis testing

1. `specify()` the variables of interest in your data frame
2. `generate()` replicates of bootstrap resamples with replacement
3. `calculate()` the summary statistic of interest
4. `visualize()` the resulting bootstrap distribution and the confidence interval.

Bootstrap CI

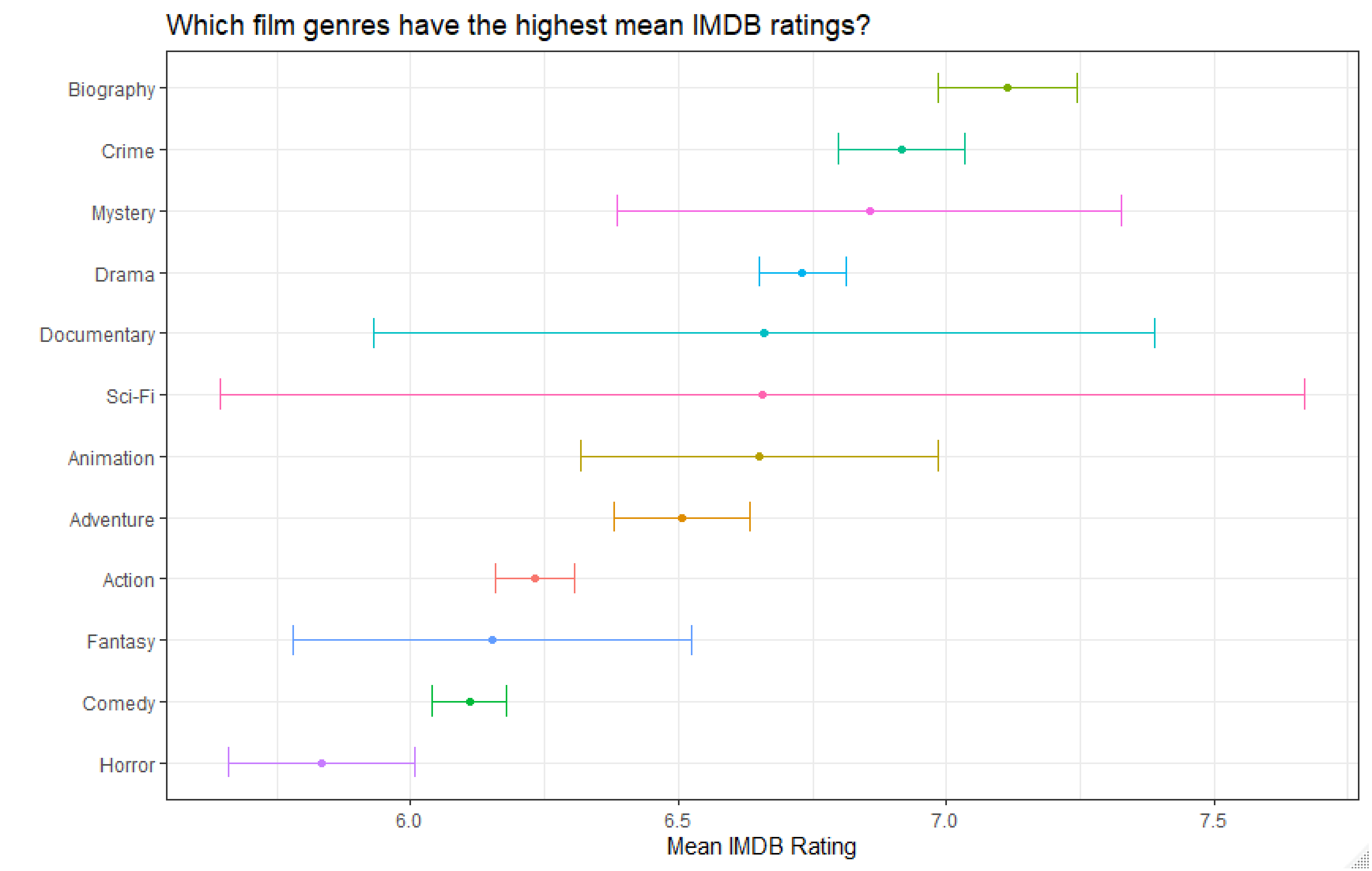


Hypothesis testing



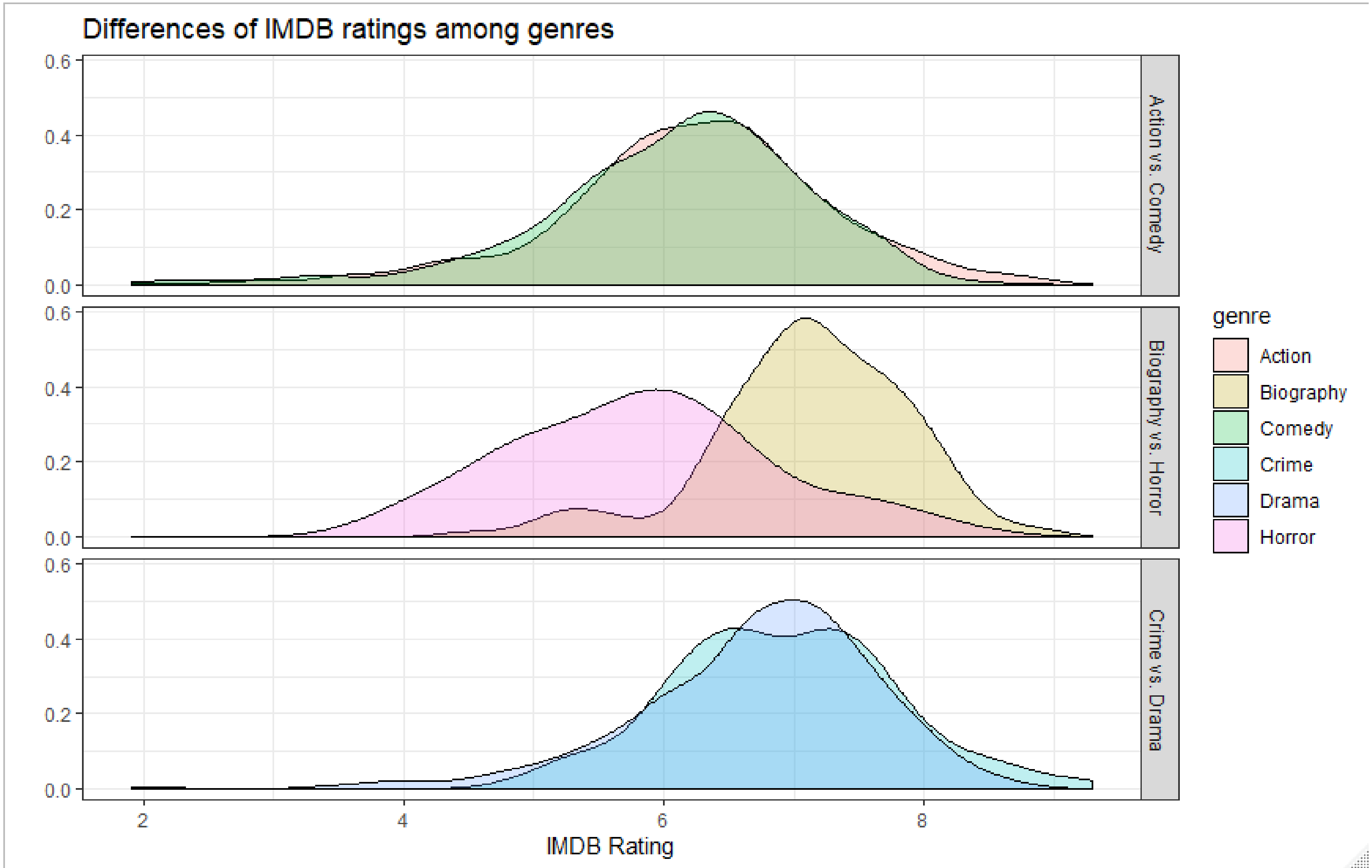
Testing for differences among mean ratings

- Consider ***Biography*** and ***Horror*** movies (first-last on the plot)
  - Difference is so pronounced, that we don't even bother to run a test
- What about ***Crime*** vs. ***Drama*** though? Or ***Action*** vs. ***Comedy?***



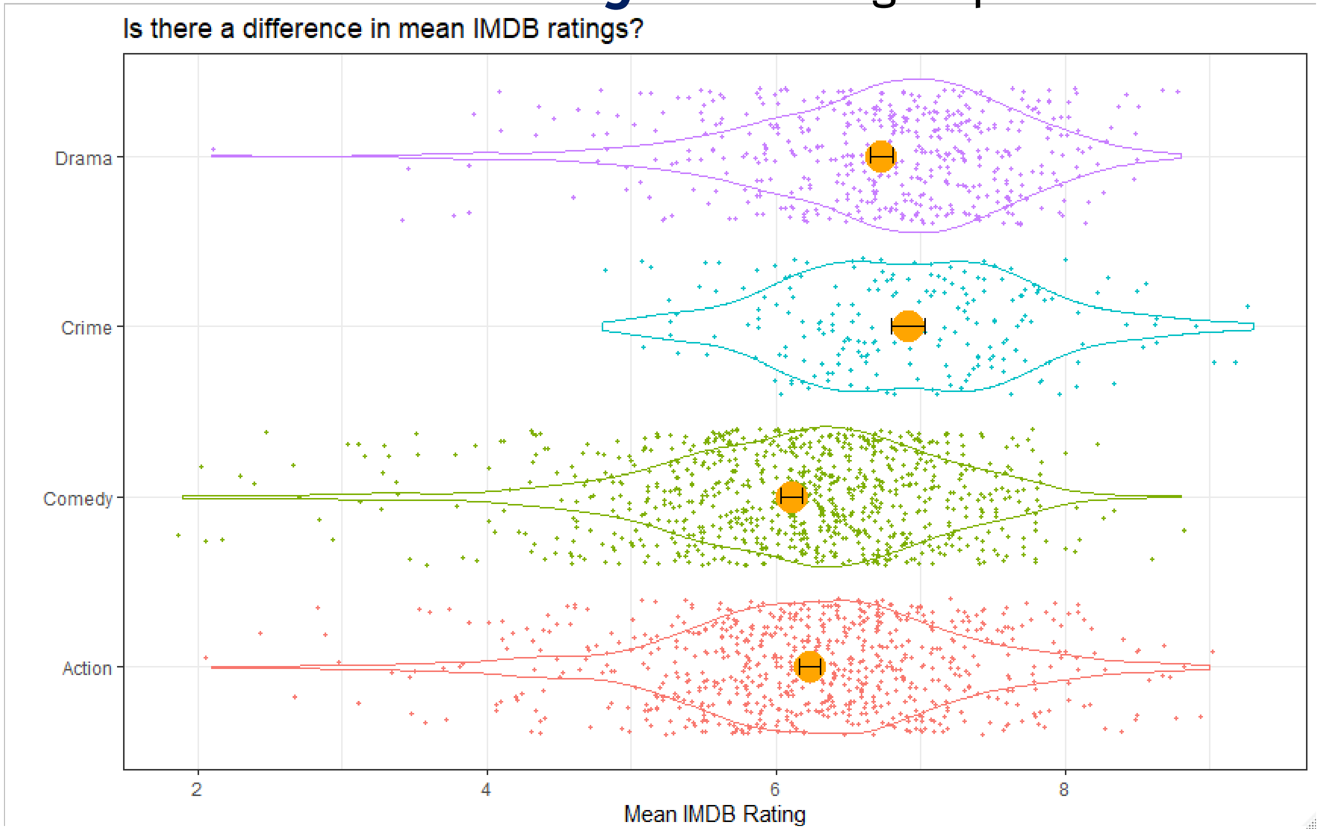
Testing for differences among mean ratings

- When comparing means we do not care about differences in individual values
- Even in ***Biography*** vs. ***Horror***, there can be an individual Horror movie that has a higher rating from a Biography movie
- We care whether ***mean rating*** between groups are the same or not



Testing for differences among mean ratings

- When comparing means we do not care about differences in individual values
- Even in ***Biography*** vs. ***Horror***, there can be an individual Horror movie that has a higher rating from a Biography movie
- We care whether ***mean ratings*** between groups are the same or not



Crime vs Drama

Action vs Comedy

```
> genre_formula_ci
# A tibble: 4 x 9
  genre  mean_rating sd_rating count t_critical se_rating margin_of_error rating_low rating_high
  <chr>        <dbl>     <dbl> <int>      <dbl>     <dbl>           <dbl>      <dbl>       <dbl>
1 Crime         6.92     0.849   202       1.97    0.0598          0.118        6.80        7.03
2 Drama         6.73     0.917   498       1.96    0.0411          0.0807       6.65        6.81
3 Action        6.23     1.03    738       1.96    0.0379          0.0745       6.16        6.31
4 Comedy        6.11     1.02    848       1.96    0.0351          0.0690       6.04        6.18
```

Overlap

Standard Error of Differences

- For large samples (by CLT), the standard error is

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

- The 95% Confidence Interval for the difference between means is approximately

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Action vs Comedy: Who has higher mean rating?

```
> genre_formula_ci
# A tibble: 4 x 9
  genre  mean_rating sd_rating count t_critical se_rating margin_of_error rating_low rating_high
  <chr>        <dbl>     <dbl> <int>      <dbl>     <dbl>           <dbl>      <dbl>       <dbl>
1 Crime         6.92     0.849   202       1.97    0.0598          0.118        6.80        7.03
2 Drama         6.73     0.917   498       1.96    0.0411          0.0807       6.65        6.81
3 Action        6.23     1.03    738       1.96    0.0379          0.0745       6.16        6.31
4 Comedy        6.11     1.02    848       1.96    0.0351          0.0690       6.04        6.18
```

1. Build two CIs for mean rating using formula and check whether they overlap

$$\text{Action CI} = 6.23 \pm 1.96 * \left(\frac{1.03}{\sqrt{738}}\right) \approx [6.16, 6.31]$$

$$\text{Comedy CI} = 6.11 \pm 1.96 * \left(\frac{1.02}{\sqrt{848}}\right) \approx [6.04, 6.18]$$

2. Build CI for mean difference in ratings using formula and check whether CI contains zero

$$(\bar{x}_1 - \bar{x}_2) + 1.96 * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

$$= (6.23 - 6.11) + 1.96 * \sqrt{\frac{1.03^2}{738} + \frac{1.02^2}{848}} =$$

$$= [0.02118, 0.224] = [2.12\%, 22.4\%]$$

Hypothesis testing steps

Test Statistic and Observed Effect $\boldsymbol{\delta}^*$

When describing a difference between two independent groups, the test statistic might be the absolute difference in the sample means $\boldsymbol{\delta}^* = |\overline{\boldsymbol{x}}_{\boldsymbol{A}} - \overline{\boldsymbol{x}}_{\boldsymbol{B}}|$

Does the observed statistic (what you measured/estimated) seem like a surprising effect?

Model of H_0

$H_0\colon \mu_\mathrm{A} - \mu_\mathrm{B} = 0$

$H_1\colon \mu_\mathrm{A} - \mu_\mathrm{B} \neq 0$

We assume as a null hypothesis that the population means are equal. We would like to assume this is true and see whether we have enough *evidence* to reject this hypothesis

i

Calculation of ***t-stat*** and ***p-value*** using formula

$$t - stat = \frac{observe - assume}{SE} = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

$$= \frac{6.23 - 6.11}{\sqrt{\frac{1.03^2}{738} + \frac{1.02^2}{848}}} = 2.37$$

ii

Bootstrapping: Distribution of δ under H_0

Simulation and ‘Null’ Worlds

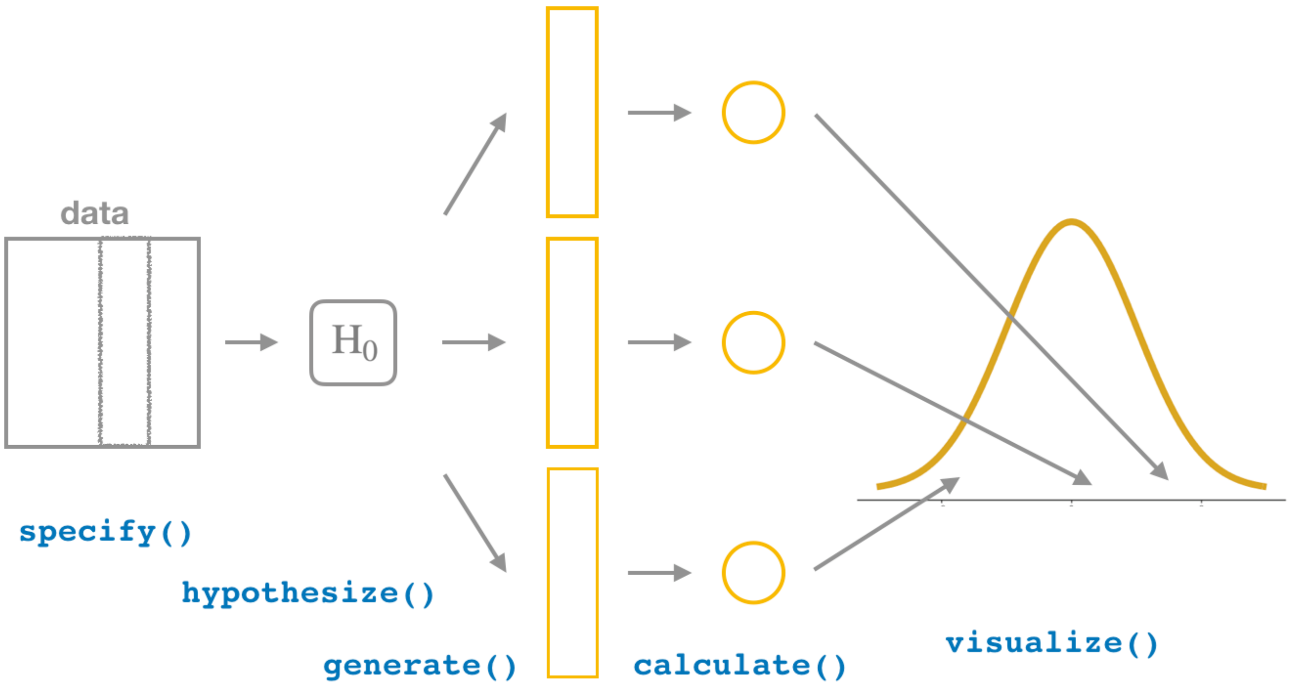
But how do you determine what δ would look like in a null world?

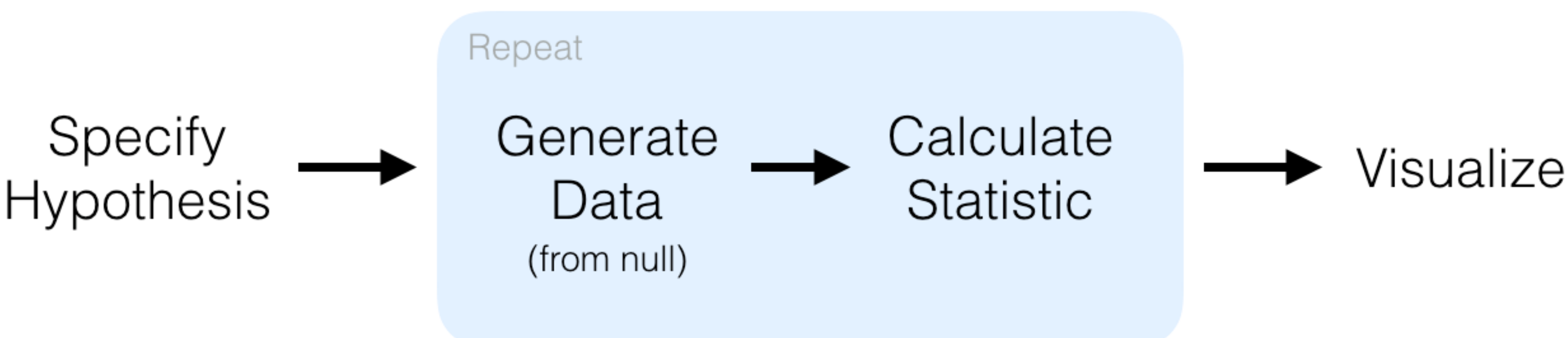
Option 1: Use math and formulas to determine probabilities

Option 2: Use brute force with simulation

Using simulation, you can test any hypothesis without formulas

Hypothesis test






```
specify(response) %>%
    hypothesize(null)    %>%    generate(reps)    %>%    calculate(stat)    %>%    visualize()
```

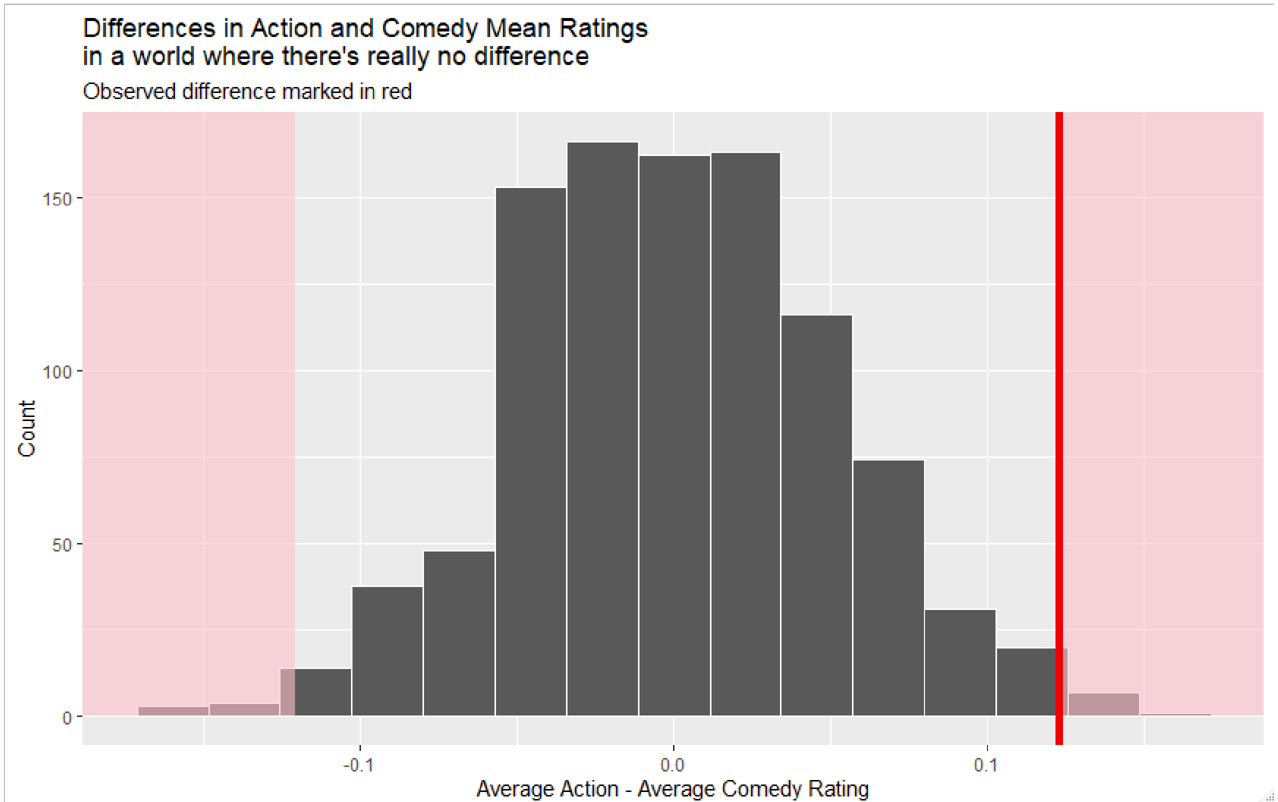
Hypothesis testing with the **infer** package

```
set.seed(1234)
ratings_in_null_world <- action_comedy %>%

  # Specify the variable of interest
  specify(rating ~ genre) %>%

  # Hypothesize a null of no (or zero) difference
  hypothesize(null = "independence") %>%

  # Generate a bunch of simulated samples
  generate(reps = 1000, type = “permute") %>%

  # Find the mean difference of each sample
  calculate(stat = "diff in means“,
       order = c(“Action", “Comedy"))


ratings_in_null_world %>% visualize()
```

p-value

Want to see where our observed sample mean difference of 0.123 (the red vertical line) falls on this null/randomization distribution.

Since we are interested in a difference, "more extreme" corresponds to values in both tails on the distribution. Let's shade our null distribution to show a visual representation of our p-value:




```
> # 4. Calculate a p-value, or the probability that we would see a red line at
> # least that extreme in null world
> diff_means_null_world %>%
    get_pvalue(obs_stat = observed_difference, direction = "both")
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.022
```

Hypothesis test approaches

1. By hand

$$t-stat = \frac{observe - assume}{SE} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

$$= \frac{6.23 - 6.11}{\sqrt{\frac{1.03^2}{738} + \frac{1.02^2}{848}}} = \boxed{2.37}$$

2. Using *t.test()*

```
> t.test(rating ~ genre, data = action_comedy)

        Welch Two Sample t-test

data:  rating by genre
t = 2.37, df = 1551, p-value = 0.018
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.021171 0.223988
sample estimates:
mean in group Action mean in group Comedy
              6.2322               6.1097
```

3. Using infer() package

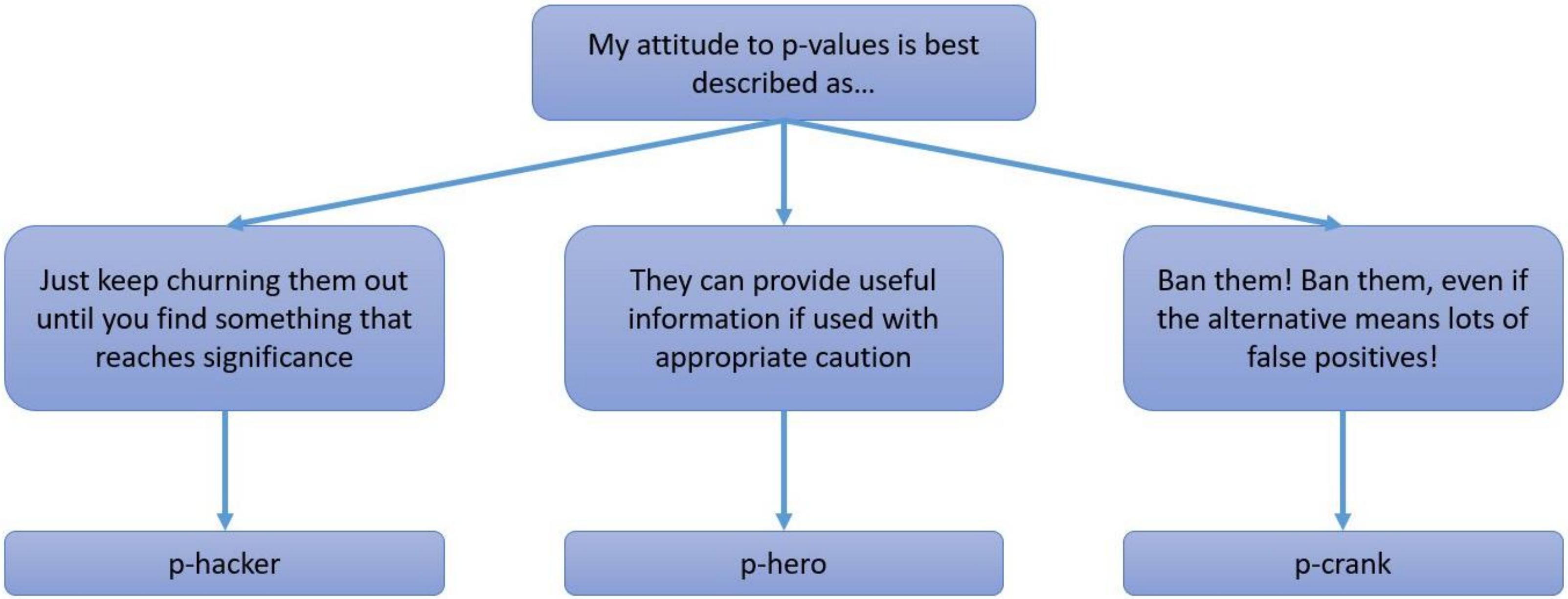
```
set.seed(1234)
ratings_in_null_world <- action_comedy %>%

  # Specify the variable of interest
  specify(rating ~ genre) %>%

  # Hypothesize a null of no (or zero) difference
  hypothesize(null = "independence") %>%

  # Generate a bunch of simulated samples
  generate(reps = 1000, type = "permute") %>%

  # Find the mean difference of each sample
  calculate(stat = "diff in means",
       order = c("Action", "Comedy"))


pub_private_null %>% visualize()
```

```
> diff_means_null_world %>%
    get_pvalue(obs_stat = observed_difference, direction = "both")
# A tibble: 1 x 1
  p_value
    <dbl>
    0.022
```

p-values

Think of p-values as a rule to guide behaviour in the long run.

- p-value < *alpha* (typically 0.05): Act as if data is not noise. You have a surprising effect and your results are an invitation to explore effect further, it cannot by itself be enough to declare a scientific fact
- p-value > *alpha*: remain uncertain or act as if data were noise



If the p-value <0.05, an effect is not 95% likely to be true.
A p-value > 0.05 does not mean there is no true effect; it means that the data we have observed is not surprising. You need large samples to detect small effects.

Fischer (1971) on why one study is never enough

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*.

Type I and Type II Errors

Type I error, false positive

- α = Pr(reject null hypothesis when H_0 is true)

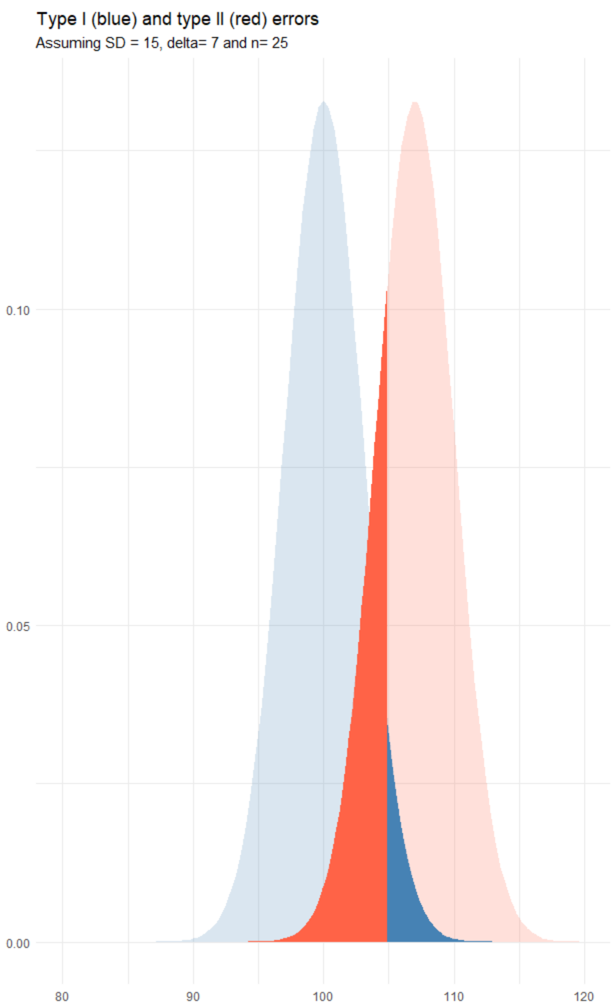
Type II error, false negative

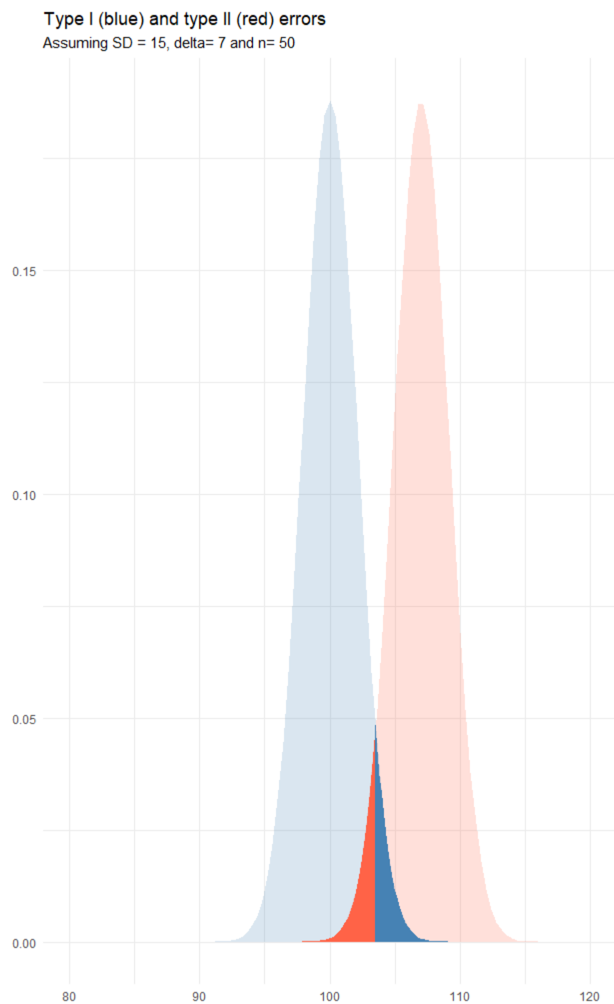
- β = Pr(fails to reject null hypothesis when H_a is true)

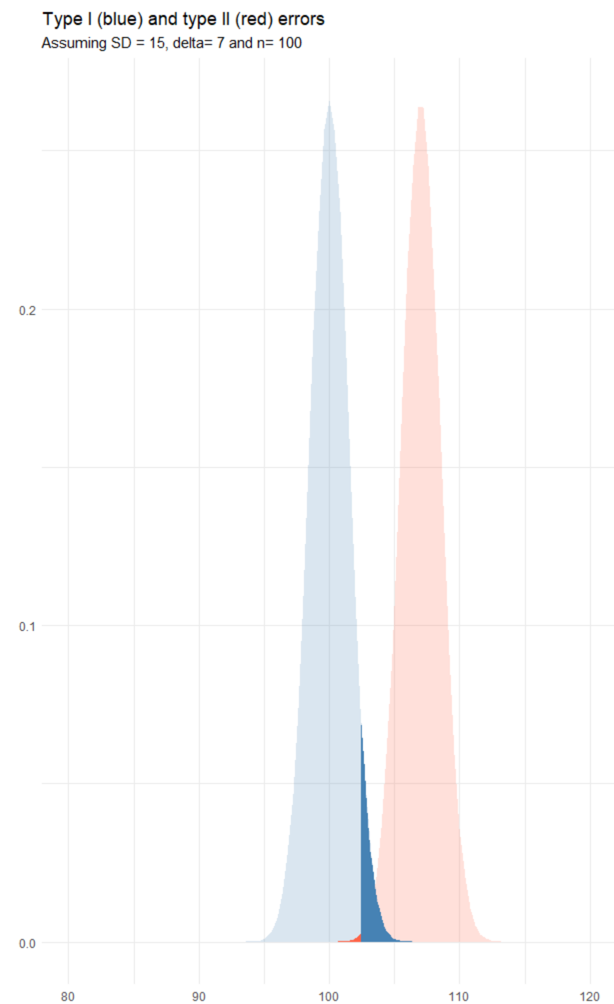
	H_0 true	H_a true
Reject H_0	Type I error	√
Fail to reject H_0	√	Type II error

Reducing Type I and Type II errors

We can reduce the probability of a Type I and a Type II error simultaneously by increasing the sample size *n*







Distributions of p-values

If statistical decisions are made based on the critical value of *alpha = 0.05*, doesn't that mean that we often come to wrong conclusions?
p-values are only meaningful if we understand them in the context of the experiment.

We can best interpret p-values if we look at them not as individual events but as distributions over long periods of time.

Samples vary, but many studies often show a clear picture.

Statistical power is the probability that you will observe a significant effect, if there is a true effect. It ranges from 0 to 1 and in an ideal world we would like our test to have as high a power as possible

Variability of a Sample

- **IQ** follows a Normal distribution with **μ = 100, σ = 15.**
- You take a sample of 25 LBS students and the sample mean IQ = 107.
- We can calculate the t-statistic $$t - stat = \frac{107 - 100}{15/\sqrt{25}} = \frac{7}{3} = 2.33$$

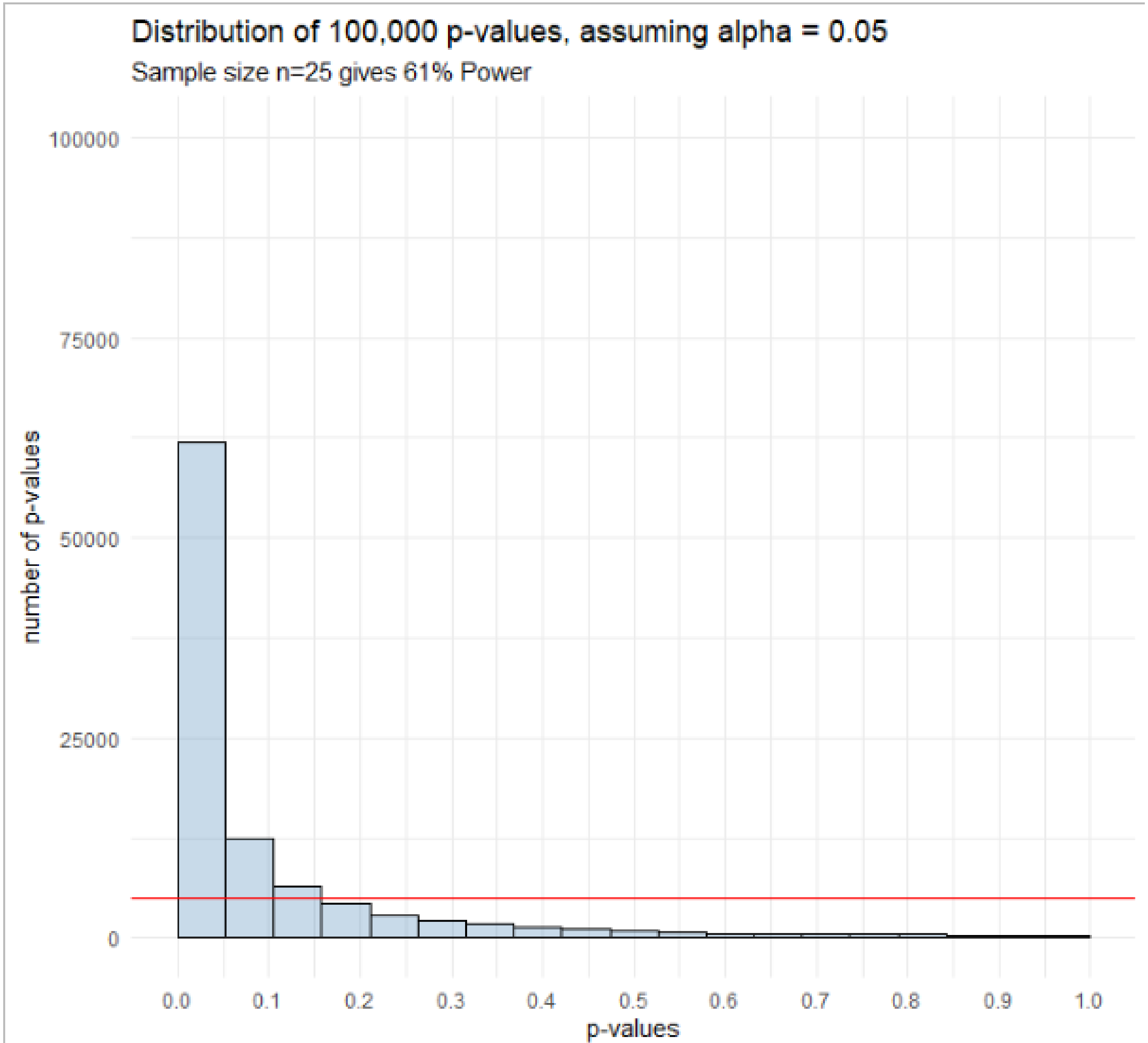
- p-value = 2.8% < 5%

```
> 2*(1-pt(2.33, 24))
[1] 0.02854225
```

- We reject the null hypothesis that LBS students have mean IQ = 100

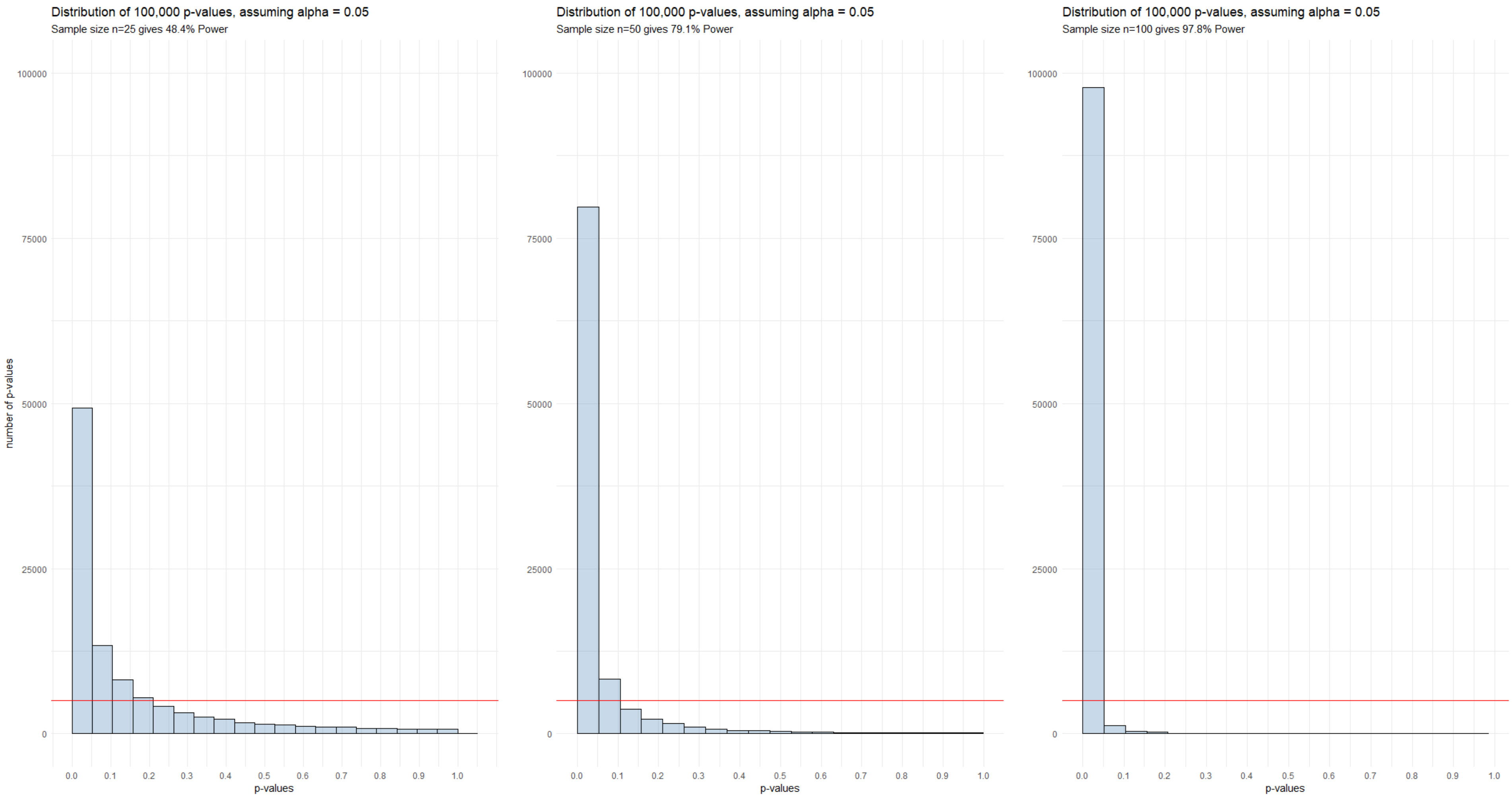
- What p-values would you get if you did the same experiment 100,000 times and compared a sample that always had a mean of 107 with the assumed population μ=100 .
- How often would you be able to reject the null hypothesis?
- Most people think that we should be able to reject correctly 95% of the times.
- We can graph the answer to this question by plotting the p-values on the x-axis and the frequency of the p-values on the y-axis

Statistical Power



Delta IQ = 7, different sample sizes

Statistical power is the probability that you will observe a significant effect, if there is a true effect. It ranges from 0 to 1 and in an ideal world we would like our test to have as high a power as possible



Statistical Power

The **power** of a test depends on delta and sample size

Compare the power of a statistical test to the power of a telescope:

- There are lots of stars in space, smaller- larger, near- far
- The more powerful the telescope is, the more chance you have of seeing a small star.
- If you cannot see anything
    - either there are no stars where you are looking,
    - or your telescope is not powerful enough.

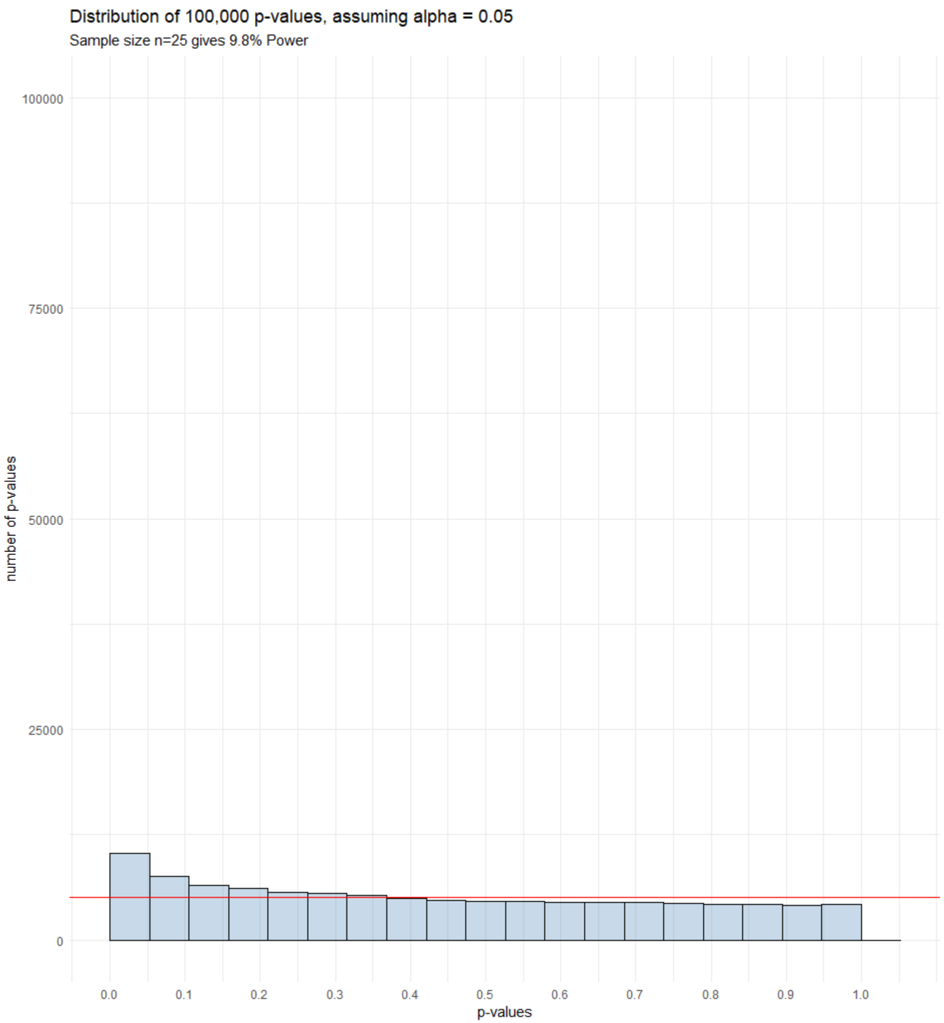
You never can draw a 100% certain conclusion about this.

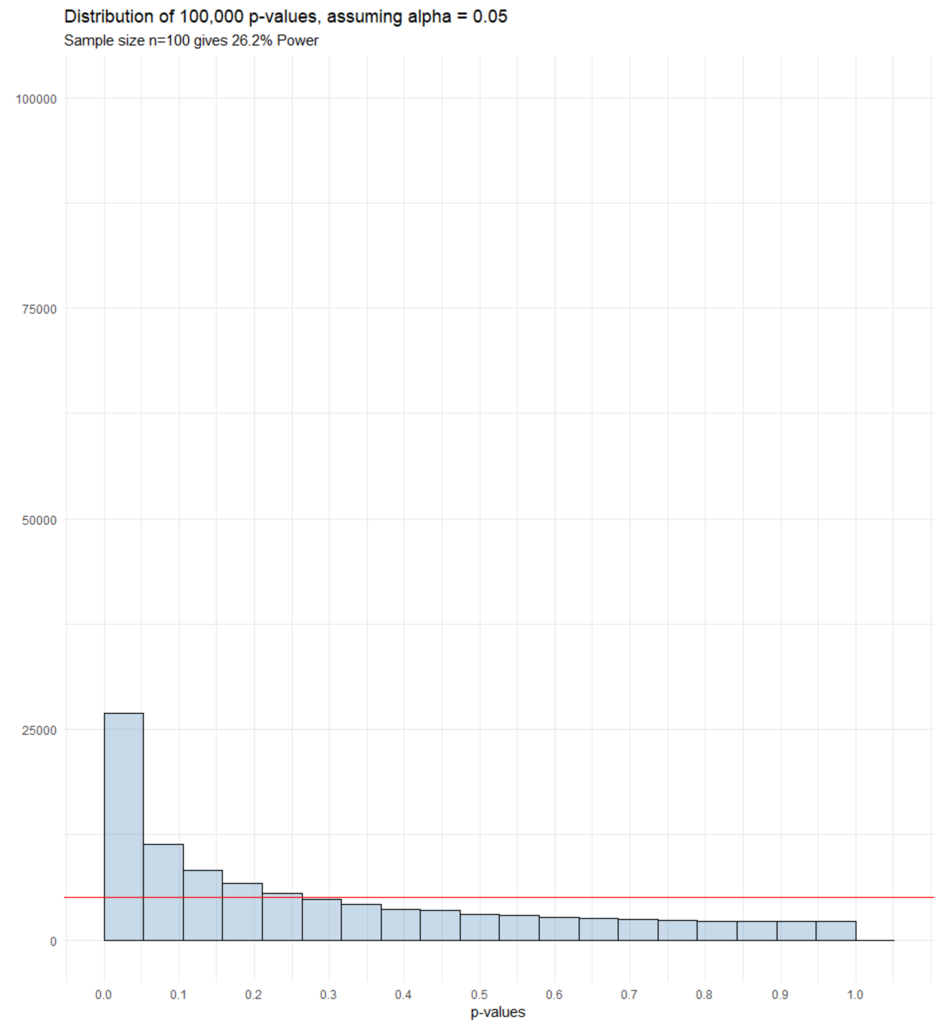
Delta IQ = 2, different sample sizes

In the earlier examples, the difference in IQ was 7 points (107 vs 100).
What if the sample mean of IQ was 102, meaning a difference of just 2 (107 vs 100)

$$t - stat = \frac{102-100}{15/\sqrt{25}} = \frac{2}{3} = 0.67$$

$$t - stat = \frac{102-100}{15/\sqrt{100}} = \frac{2}{1.5} = 1.33$$





Statistical Power

The power of a test depends on delta and sample size

Compare the power of a statistical test to the power of a telescope:

- There are lots of stars in space, smaller- larger, near- far
- The more powerful the telescope is, the more chance you have of seeing a small star.
- If you cannot see anything
    - either there are no stars where you are looking,
    - or your telescope is not powerful enough.

You never can draw a 100% certain conclusion about this.

```
> power.t.test(n=NULL, delta = 7, sd = 15, power = 0.90, sig.level = 0.01)

     Two-sample t test power calculation 

              n = 138.3163
          delta = 7
             sd = 15
      sig.level = 0.01
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group

> power.t.test(n=NULL, delta = 2, sd = 15, power = 0.90, sig.level = 0.01)

     Two-sample t test power calculation 

              n = 1675.591
          delta = 2
             sd = 15
      sig.level = 0.01
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Volume 25, Issue 5 October 2016 , pp. 428-435

How to prove that your therapy is effective, even when it is not: a guideline

Part of: Special Articles

P. Cuijpers [a1] [a2] and I. A. Cristea [a3] [a4]

DOI: https://doi.org/10.1017/S2045796015000864 Published online by Cambridge University Press: 28 September 2015

Related commentaries (2)

Abstract

Aims.

Suppose you are the developer of a new therapy for a mental health problem or you have several years of experience working with such a therapy, and you would like to prove that it is effective. Randomised trials have become the gold standard to prove that interventions are effective, and they are used by treatment guidelines and policy makers to decide whether or not to adopt, implement or fund a therapy.

Methods.

You would want to do such a randomised trial to get your therapy disseminated, but in reality your clinical experience already showed you that the therapy works. How could you do a trial in order to optimise the chance of finding a positive effect?

Results.

Methods that can help include a strong allegiance towards the therapy, anything that increases expectations and hope in participants, making use of the weak spots of randomised trials (risk of bias), small sample sizes and waiting list control groups (but not comparisons with existing interventions). And if all that fails one can always not publish the outcomes and wait for positive trials.

Conclusions.

Several methods are available to help you show that your therapy is effective, even when it is not.

https://www.cambridge.org/core/journals/epidemiology-and-psychiatric-sciences/article/how-to-prove-that-your-therapy-is-effective-even-when-it-is-not-a-guideline/F666A4AB2E7A70FE12F7A3C3B97380CA#

Session Summary

We covered

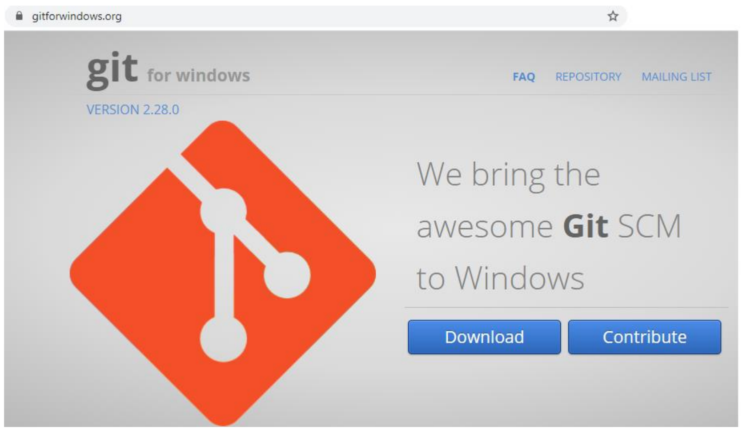
- Distribution of the sample mean
- Experiments
- Hypothesis tests
    - By hand, using formulas and normality assumptions
    - With simulation, using **infer** package

Version control with Git(Hub)

Blogdown + Hugo + Netlify

Local Git installation

Windows



https://gitforwindows.org/

macOS

Option 1 (*highly recommended*): Install the Xcode command line tools (**not all of Xcode**), which includes Git.

Go to the shell and enter one of these commands to elicit an offer to install developer command line tools:

```
git --version
git config
```

Accept the offer! Click on “Install”.

Here’s another way to request this installation, more directly:

```
xcode-select --install
```

Install the Xcode command line tools

Source: https://happygitwithr.com/install-git.html#install-git

Git + Rstudio configuration

Chapter 7 Introduce yourself to Git

In the shell (Appendix A):

```
git config --global user.name 'Jane Doe'
git config --global user.email 'jane@example.com'
git config --global --list
```

substituting your name and **the email associated with your GitHub account**.

The usethis package offers an alternative approach. You can set your Git user name and email from within R:

```
## install if needed (do this exactly once):
## install.packages("usethis")

library(usethis)
use_git_config(user.name = "Jane Doe", user.email = "jane@example.org")
```

Source: https://happygitwithr.com/hello-git.html

Cache credentials for HTTPS

Chapter 10 Cache credentials for HTTPS

If you plan to push/pull using HTTPS, you want to cache your credentials (e.g. password), so you don't need to enter them over and over again. Alternatively, you could set up SSH keys (chapter 11). I suggest you set up one of these methods of authentication on each computer you want to connect to GitHub from.

I find HTTPS easier to get working quickly and **strongly recommend** it when you first start working with Git/GitHub. HTTPS is what GitHub recommends, presumably for exactly the same reasons. I started with HTTPS, preferred SSH for a while, and have returned to HTTPS. Either is fine, you can change your mind later, and you can use HTTPS on one machine and SSH on another.

Remember: the transport protocol is controlled by the URL you use for remote repo access.

HTTPS remotes look like `https://github.com/<OWNER>/<REPO>.git`.
SSH remotes look like `git@github.com:<OWNER>/<REPO>.git`.

10.1 You should get a personal access token (PAT)

10.1.1 Why a PAT?

Password-based authentication for Git is deprecated, i.e. you really should *not* be sending your username and password every time you push or pull. Here, I'm referring to the username and password you would use to login to GitHub in the browser.

What should you do instead?

Get a **personal access token** (PAT) and use that as your credential for HTTPS operations. (The PAT will actually be sent as the password and the username is somewhat artificial, consulted only for credential lookup.)

If you turn on two-factor authentication (a.k.a. "2FA") for GitHub and you use HTTPS, you absolutely **must** send a personal access token. And, really, it's a good idea for everyone to turn on 2FA and for everyone who uses HTTPS to use a PAT.

The final selling point is that once you configure a PAT, several R packages, including usethis and gh, will be able to work with the GitHub API on your behalf, automagically. Therefore, a properly configured PAT means all of this will work seamlessly:

- Remote HTTPS operations via command line Git and, therefore, via RStudio
- Remote HTTPS operations via the gert R package and, therefore, usethis
- GitHub API operations via the gh R package and, therefore, usethis

10.1.2 How to get a PAT?

Source: https://happygitwithr.com/hello-git.html

Git + RStudio

1. Create a new public repo, called ***my_website***, on GitHub and initialise with a README.
2. Copy the HTTPS/SSH link (the green **Code** button).
3. Open up RStudio.
4. Navigate to *File -> New Project -> Version Control -> Git*
5. Paste your copied link into the "Repository URL:" box.
6. Choose the project path ("Create project as subdirectory of:") and click Create Project.
7. Look at the top-right panel in your RStudio IDE. Do you see the ***Git*** tab?
8. There should already be some files in there, which we'll ignore for the moment
9. Go to ***Files*** tab in the bottom-right panel and click on the README file.
10. Add some text like "Hello World!" and save (Ctrl/Cmd + S) the README.
11. Do you see any changes in the "Git" panel?

Main Git operations

The four main Git operations, in this order, are:

1. Stage (or "add")

- Tell Git that you want to add changes to the repo history (file edits, additions, deletions, etc.)

2. Commit

- Tell Git that, yes, you are sure these changes should be part of the repo history. This is when you take a snapshot of your work as it is now. Git won't allow you to commit if you don't add a helpful, explanatory message

3.Pull

- Get any new changes made on the GitHub repo (i.e. the upstream remote), either by you on another machine or your collaborators.

4.Push

- Push any (commited) local changes to the GitHub repo

Always ***pull first*** from the upstream repo ***before you push*** any changes, even when working alone. It's a good habit that will save you troubles later on

Git from the shell/terminal

Besides using the Rstudio Git panel and its GUI, you can also give Git commands from Rstudio shell/terminal.

1. Add ("stage") a file or group of files: **git add NAME-OF-FILE-OR-FOLDER**
You can use wildcard characters to stage a group of files (e.g. sharing a common prefix).
There are a bunch of useful flag options too:

Stage all files	**git add -A**
Stage updated files only (modified or deleted, but not new)	**git add -u**
Stage new files only (not updated)	**git add .**

2. Commit your changes **git commit -m "Helpful message"**

3. Pull from the upstream repository (i.e. GitHub) **git pull**

4. Push local changes that you commited to the upstream repo **git push**

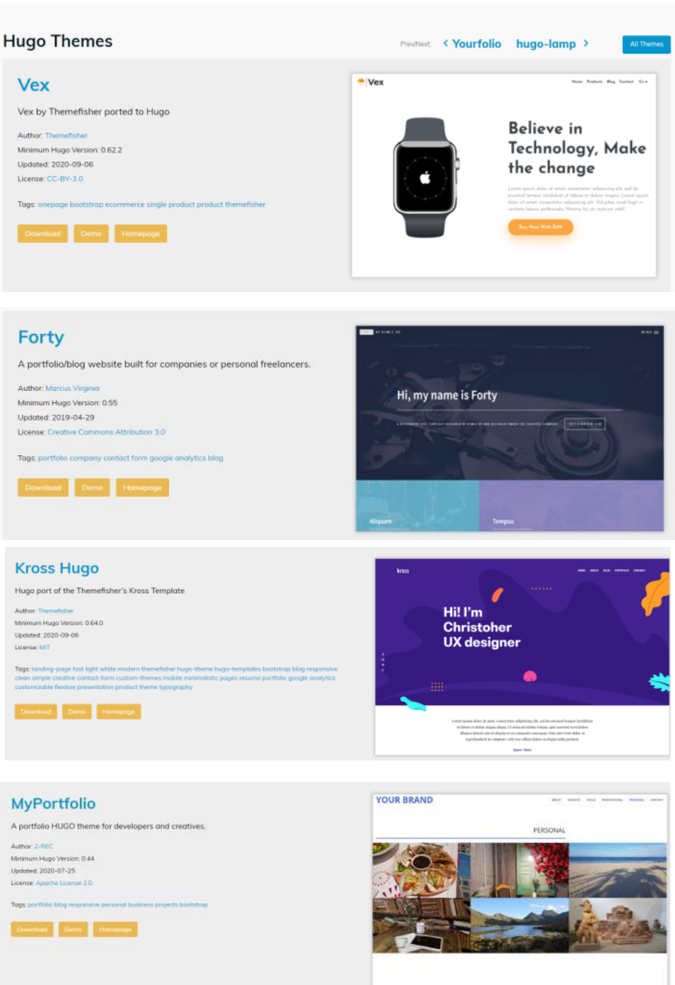
Clone the repo.	**git clone REPOSITORY-URL**
See the commit history (hit spacebar to scroll down or q to exit)	**git log**
See changes	**git status**

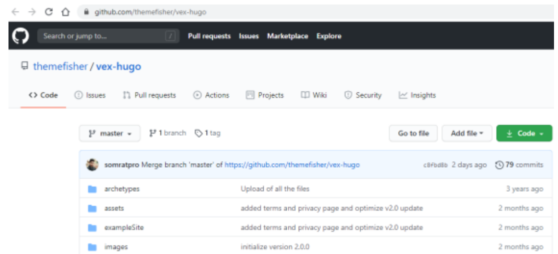
Blogdown, hugo, your website

Steps to follow to get our website

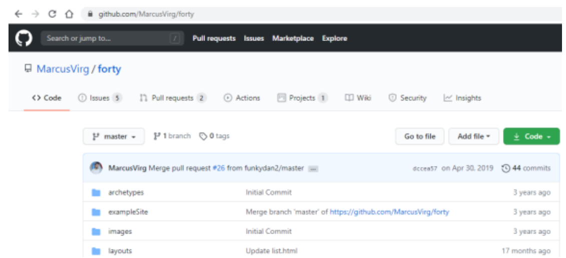
1. Create a new public repo, called test, on GitHub and initialise with a README.
2. Copy the HTTPS/SSH link (the green Code button).
3. Open up RStudio.
4. Navigate to File -> New Project -> Version Control -> Git
5. Paste your copied link into the "Repository URL:" box.
6. Choose the project path ("Create project as subdirectory of:") and click Create Project.
7. Look at the top-right panel in your RStudio IDE. Do you see the Git tab?
8. There should already be some files in there, which we'll ignore for the moment
9. Go to Rstudio console (bottom-left) and type
   - **library(blogdown)**
   - **install_hugo()**
   - **hugo_version()**
10. Go to https://themes.gohugo.io/ to choose your favourite theme. We use theme Forty, but you can choose anything you like
11. All hugo themes are on github, so to use them you need to note the USER/NAME address.

Github USER/REPO address

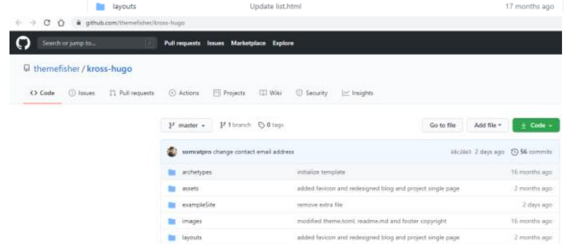




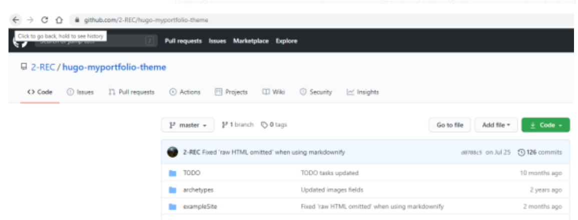
https://github.com/**themefisher/vex-hugo**



https://github.com/**MarcusVirg/forty**

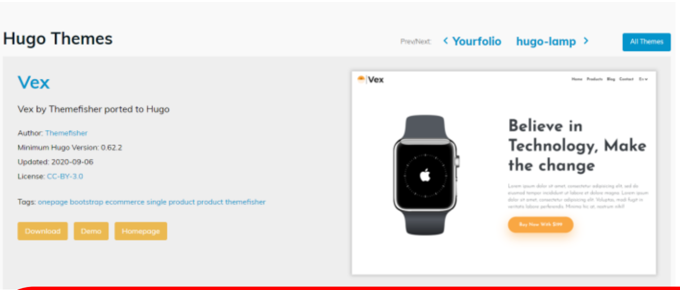


https://github.com/**themefisher/kross-hugo**

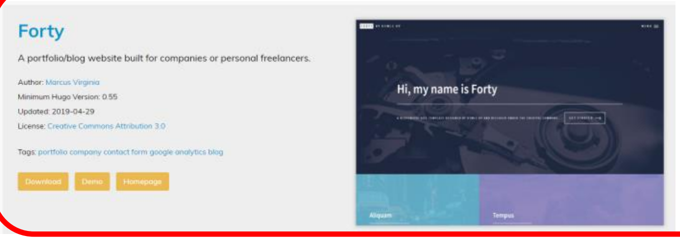


https://github.com/**2-REC/hugo-myportfolio-theme**

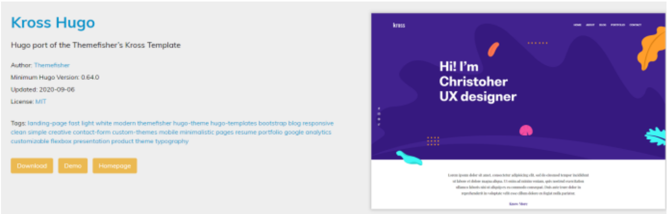
Build a template site with your chosen theme



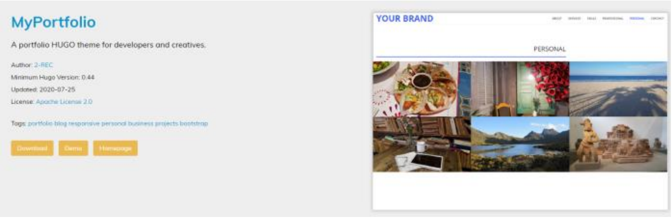

```r
blogdown::new_site(theme = "themefisher/vex-hugo",
        sample = TRUE,
        theme_example = TRUE,
        empty_dirs = TRUE,
        to_yaml = TRUE)
```




```r
blogdown::new_site(theme = "MarcusVirg/forty",
        sample = TRUE,
        theme_example = TRUE,
        empty_dirs = TRUE,
        to_yaml = TRUE)
```

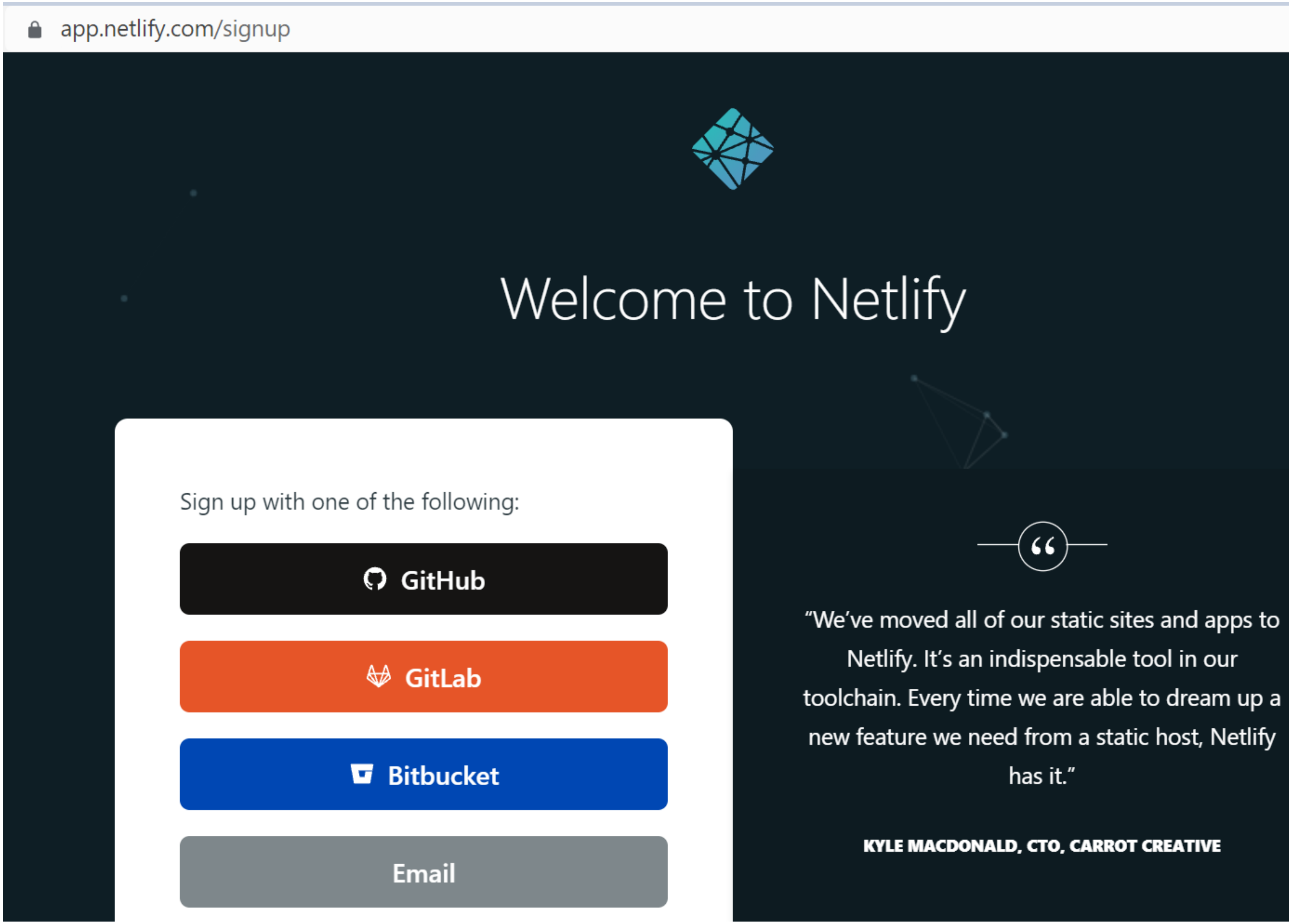



```r
blogdown::new_site(theme = "themefisher/kross-hugo",
        sample = TRUE,
        theme_example = TRUE,
        empty_dirs = TRUE,
        to_yaml = TRUE)
```

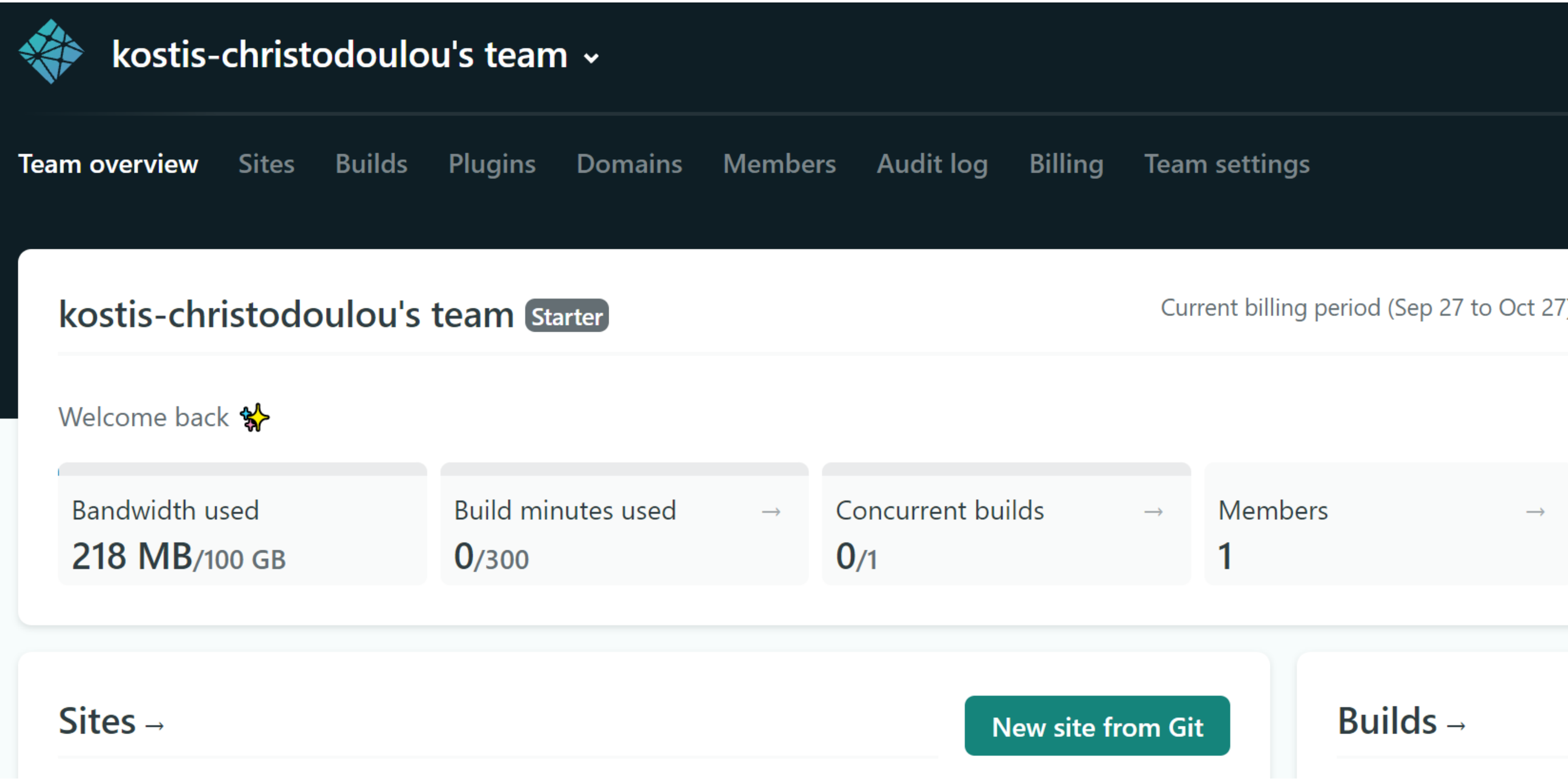


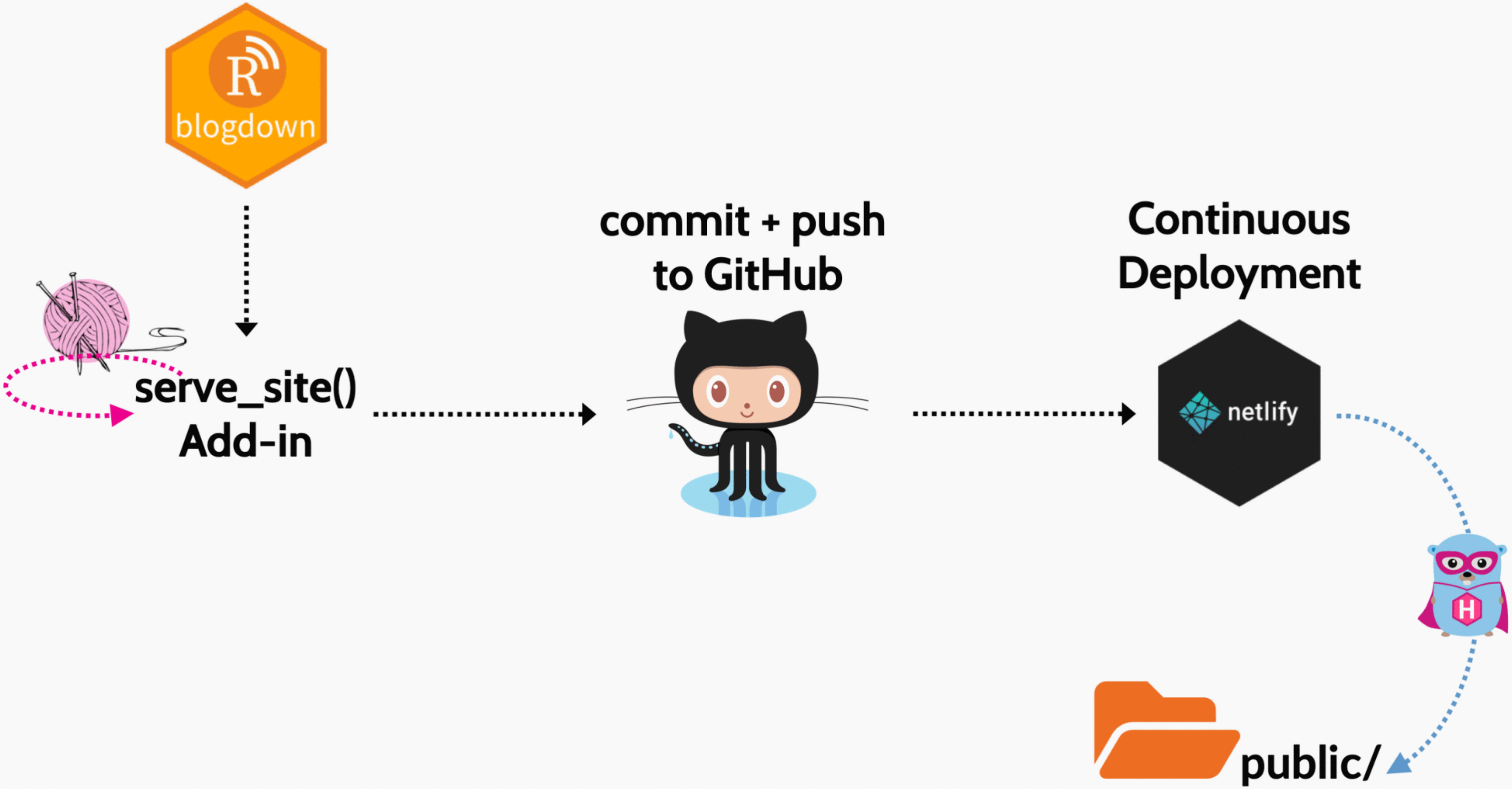

```r
blogdown::new_site(theme = "2-REC/hugo-myportfolio-theme",
        sample = TRUE,
        theme_example = TRUE,
        empty_dirs = TRUE,
        to_yaml = TRUE)
```

Sign up at www.netlify.com using your github



Click on **New site from Git**





blogdown
serve_site()
Add-in
commit + push
to GitHub
Continuous
Deployment
netlify
H
public/

For continuous deployment, change **netlify.toml**

```
[build]
  publish = "public"
  command = "hugo"

[build.environment]
  HUGO_VERSION = "0.83.1"
  HUGO_ENABLEGITINFO = "true"

[context.production.environment]
  HUGO_ENV = "production"

[context.branch-deploy.environment]
  HUGO_VERSION = "0.83.1"

[context.deploy-preview.environment]
  HUGO_VERSION = "0.83.1"
```

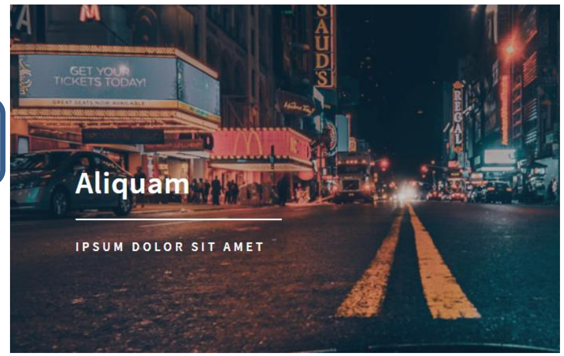
Same as hugo_version()

Portfolio website – change tile image, title, subtitle

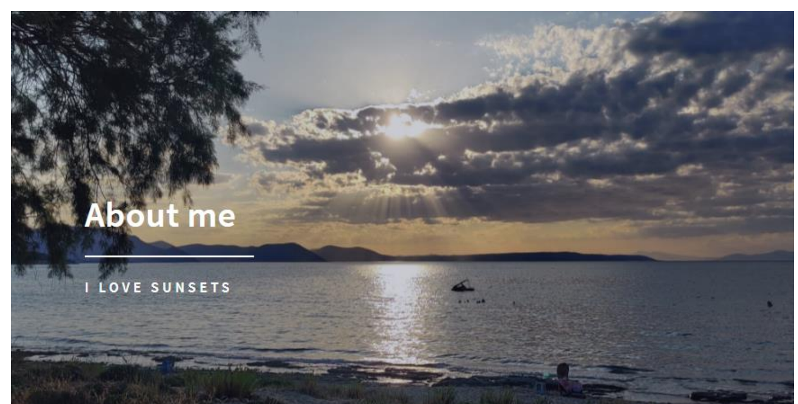
config.yaml

\themes\forty\static\img

```
tiles:
  enable: yes
  showcase:
    - image: pic01.jpg
      subtitle: Ipsum Dolor Sit Amet
      title: Aliquam
      url: blogs/aliquam
    - image: pic02.jpg
      subtitle: Feugiat Amet Tempus
      title: Tempus
      url: blogs/tempus
```




```
  showcase:
    - image: backround_sunset.jpg
      subtitle: I love sunsets
      title: About me
      url: blogs/aliquam
    - image: pic02.jpg
      subtitle: Feugiat Amet Tempus
      title: Tempus
      url: blogs/tempus
```

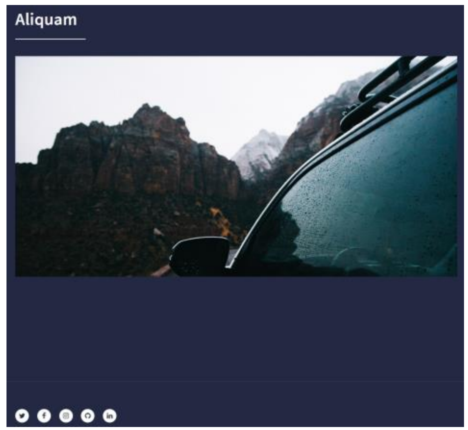


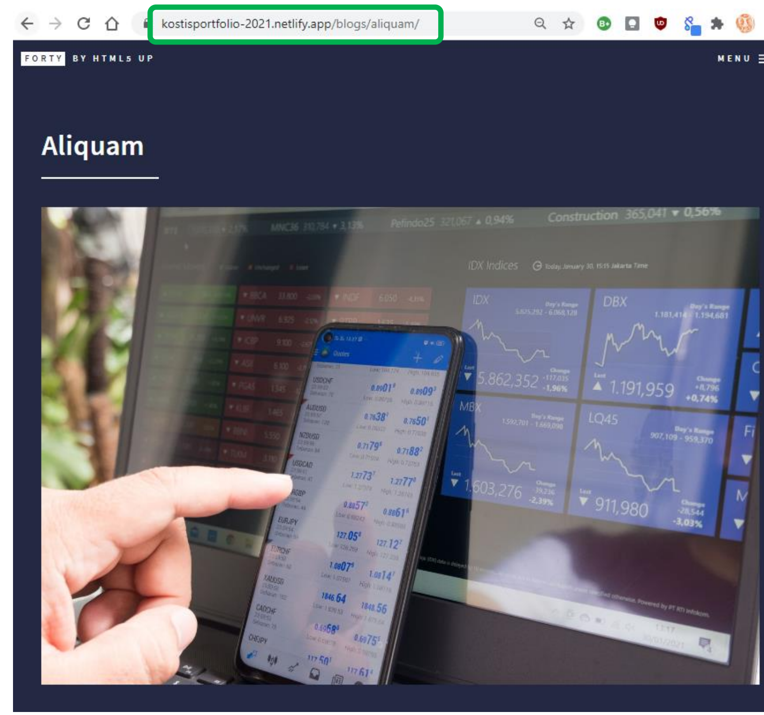
Change image inside blog
slug = blog address

config.yaml

```yaml
tiles:
  enable: yes
  showcase:
    - image: pic01.jpg
      subtitle: Ipsum Dolor Sit Amet
      title: Aliquam
      url: blogs/aliquam
    - image: pic02.jpg
      subtitle: Feugiat Amet Tempus
      title: Tempus
      url: blogs/tempus
```

\static\img\blogs





\content\blogs\blog4.md

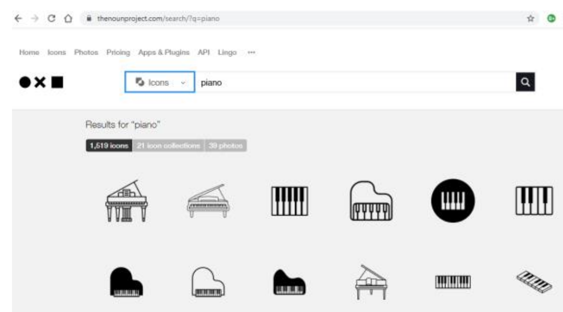
```yaml
---
categories:
- ""
- ""
date: "2017-10-31T22:42:51-05:00"
description: Nullam et orci eu lorem consequat tincidunt vivamus et
  sed nunc rhoncus condimentum sem. In efficitur ligula tate urna.
  sed magna lacinia magna pellentesque lorem ipsum dolor. Nullam et
  consequat tincidunt. Vivamus et sagittis tempus.
draft: false
image: forex_prices.jpg
keywords: ""
slug: aliquam
title: Aliquam
---
```

Where to find royalty-free photos-icons?

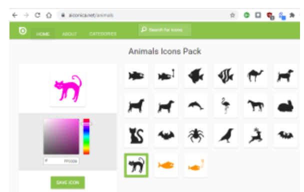
- Use the Creative Commons filters on *Google Images* or *Flickr*
- Unsplash https://unsplash.com/
- freephotos.cc https://freephotos.cc/en
- Pexels https://www.pexels.com/
- Pixabay https://pixabay.com/
- StockSnap.io https://stocksnap.io/
- Burst https://burst.shopify.com/

Icons and Vectors

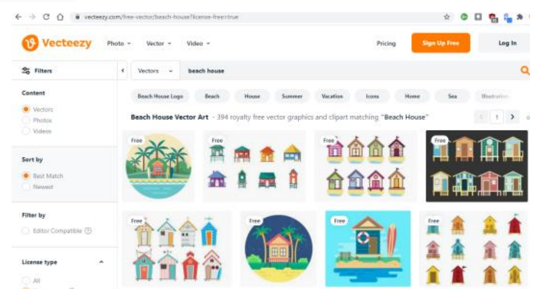
- Noun Project https://thenounproject.com/



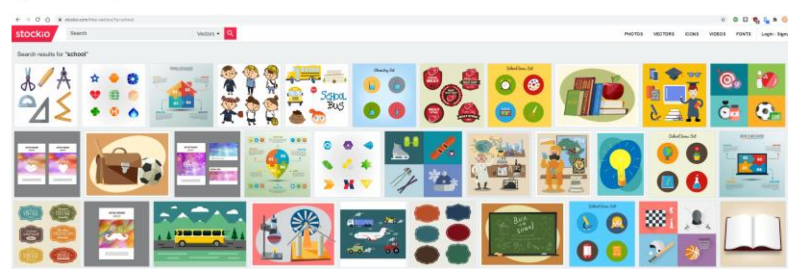
- Aiconica https://aiconica.net/



- Vecteezy https://www.vecteezy.com/



- Stockio https://www.stockio.com/



Take your pre-programme Rmd save it in \content\blogs\

\content\blogs\blog4.md

```
---
categories:
- ""
- ""
date: "2017-10-31T22:42:51-05:00"
description: Nullam et orci eu lorem consequat tincidunt vivamus et
  sed nunc rhoncus condimentum sem. In efficitur ligula tate urna.
  sed magna lacinia magna pellentesque lorem ipsum dolor. Nullam et
  consequat tincidunt. Vivamus et sagittis tempus.
draft: false
image: forex_prices.jpg
keywords: ""
slug: aliquam
title: Aliquam
---

```

```
---
title: 'Pre-programme Assignment'
author: "Kostis"
date: "What date?"
output:
  html_document:
    theme: flatly
    highlight: zenburn
    toc: yes
    toc_float: yes
---

```{r load-libraries, warning=FALSE, message=FALSE, echo=FALSE}
library(tidyverse)  # Load ggplot2, dplyr, and all the other tidyverse packages
library(gapminder)  # gapminder dataset
library(here)
library(janitor)
```

```

Change the YAML in your Rmd to be like blog4.md

```
---
categories:
- ""
- ""
date: "2017-10-31T22:42:51-05:00"
description: Nullam et orci eu lorem consequat tincidunt vivamus et sagittis magna
  sed nunc rhoncus condimentum sem. In efficitur ligula tate urna. Maecenas massa
  sed magna lacinia magna pellentesque lorem ipsum dolor. Nullam et orci eu lorem
  consequat tincidunt. Vivamus et sagittis tempus.
draft: false
image: forex_prices.jpg
keywords: ""
slug: aliquam
title: Aliquam

```{r load-libraries, warning=FALSE, message=FALSE, echo=FALSE}
library(tidyverse)  # Load ggplot2, dplyr, and all the other tidyverse packages
library(gapminder)  # gapminder dataset
library(here)
library(janitor)
```

The goal is to test your software installation, to demonstrate competency in Markdown,
```

blogdown::serve_site() will knit the Rmd

```
Quitting from lines 144-148 (kostis_pre_programme.Rmd)
Error: 'C:/Users/kchristodoulou/Desktop/my_gorgeous_website/data/brexit_results.csv' does not exist.
Execution halted
Error: Failed to render content/blogs/kostis_pre_programme.Rmd
```

1. Create a folder \data\ and save brexit_results.csv
2. Don't knit. You may have to restart R (Cmd + Shift + F10)
3. Delete `blog4.md`, because it has the same slug (shortcut address)
4. Run `blogdown::serve_site()`

5. Once it has rendered, you need to
   - `git add -A`
   - `git commit -m "a useful message"`
   - `git pull`
   - `git push`