The Bootstrap

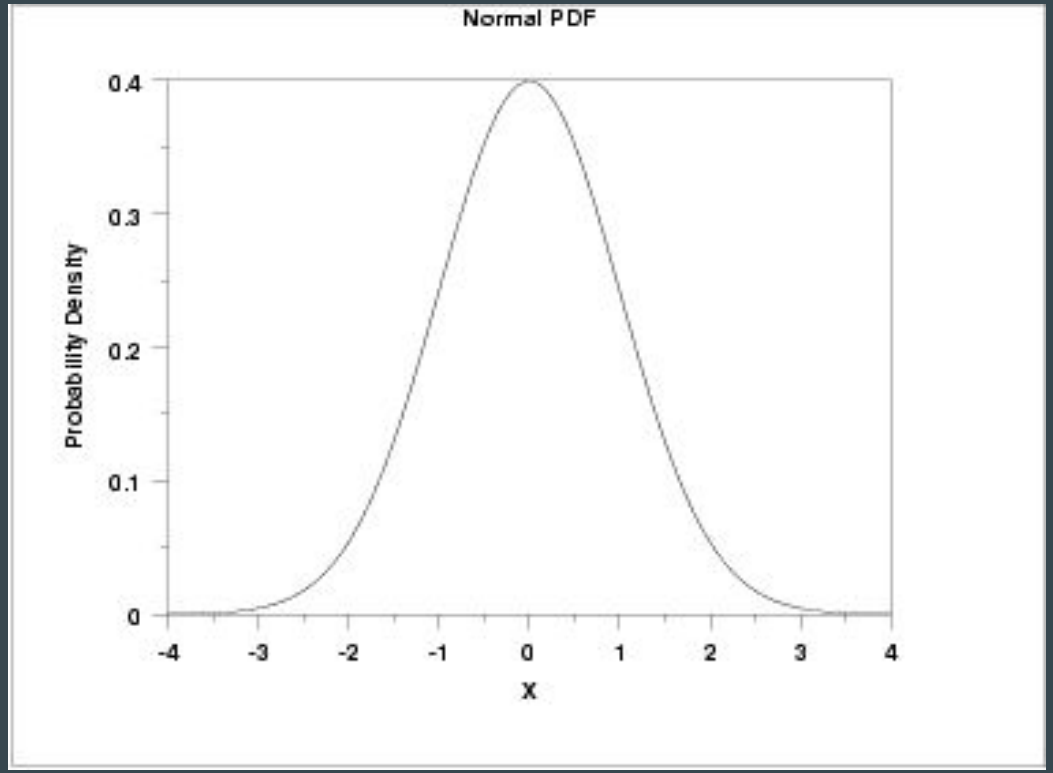
•••

25 February 2019
PLSC 309

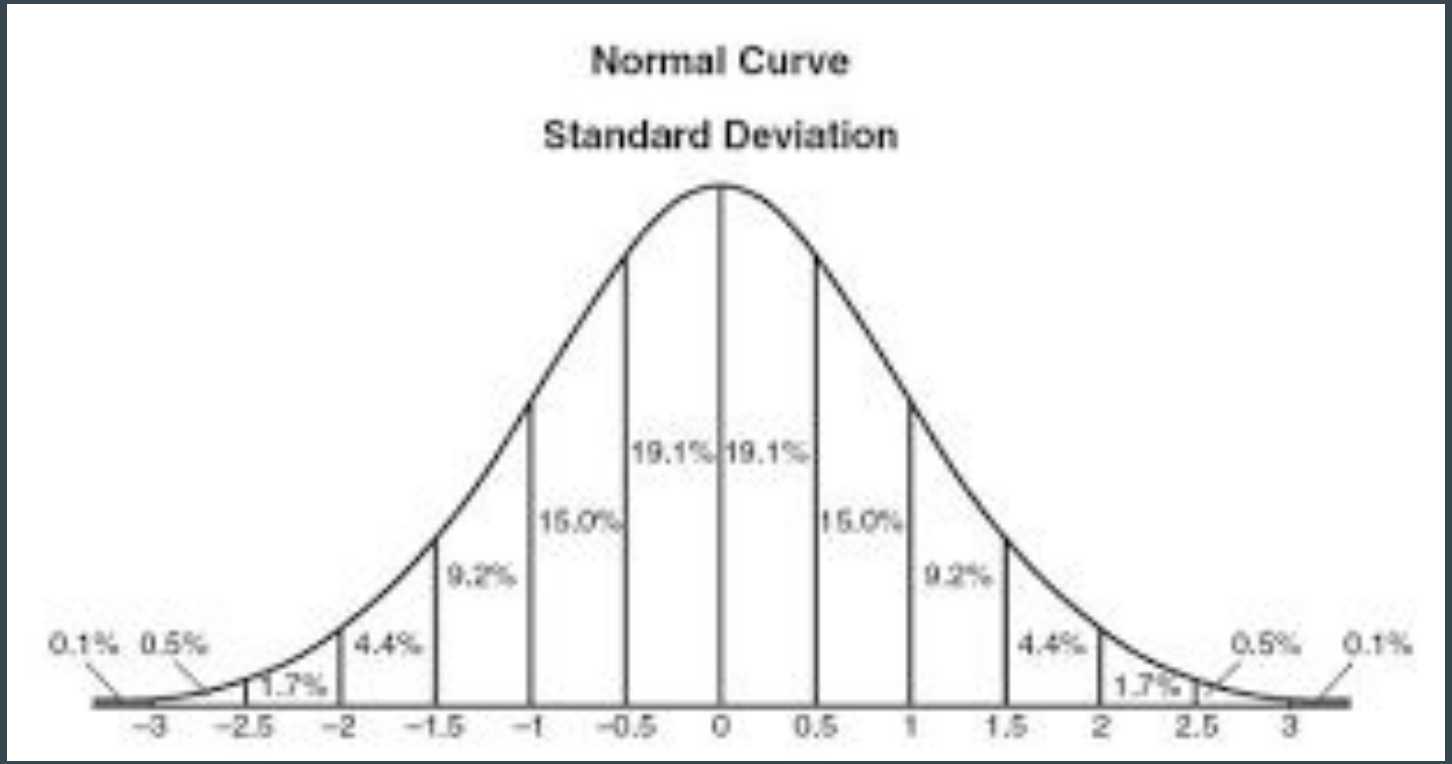
NHST steps

1. Assume that data follows the null distribution
2. Calculate z-score for point estimate you're interested in
    - Use standard error instead of standard deviation
    - $P(x_i) = P(z_i)$
3. Evaluate z-score with respect to the null distribution
    - Normal(μ=0, σ=1)
4. If $p(z_i < \alpha)$ is true
    - Reject H_0 / Accept H_A
5. If $p(z_i > \alpha)$ is true
    - Accept H_0 / Reject H_A

Assume data follows null distribution



Calculate z-score for point estimate



Find associated probability with z-score

$$\text{P-value} = P(H_0 \mid X = z_i)$$

- Given the null hypothesis
- What is the probability that our random variable X takes the value z_i
- z_i is the z-score for your point estimate

Pick a probability that you think is too extreme

- α is the critical value
- Below this probability, you think an observation is so extreme...
- ...that it's unlikely to come from a null distribution
- That means, when you reject the null you will be wrong α% of the time

Pick a probability that you think is too extreme

- α is the critical value
- Below this probability, you think an observation is so extreme...
- ...that it's unlikely to come from a null distribution
- That means, when you reject the null you will be wrong α% of the time
- But that does not mean you will be right 1 - α% of the time!!!
  - You have no idea what H_A looks like, so you can't say anything *about the distribution of* your point estimate
  - You can only say it probably doesn't follow the null

Interpreting the p-value

Assuming that nothing has changed, what is the probability that the value we see for X would be produced by random chance?

The benefits of NHST

Because we assume the null distribution (based on CLT), we only need a few pieces of information to test a hypothesis:

After a presidential debate, the candidate has an approval rating, drawn from a sample of 100 voters, of 53%. The average of past approval polls has been 48% with a 5% standard deviation. Did the debate make a difference in the candidate's approval?

$$Z = 53 - 48 / (5/\text{sqrt}(100)) = 2.5$$

The downsides of NHST

- Lots of assumptions!
- Misleading / difficult interpretation
- Measures what you don’t care about (null distribution) not what you do care about (alternative distribution)

The downsides of NHST

Say you test 100 drugs, administering them to a random sample of patients and calculating the associated p-value, for the effectiveness in treating a condition. In reality, only one of them in successful.

- On average, you will get six positive results
  - 5 false positives due to random chance
  - 1 true result
- In other words, you're wrong 80% of the time!

Basically nobody interprets p-values correctly

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

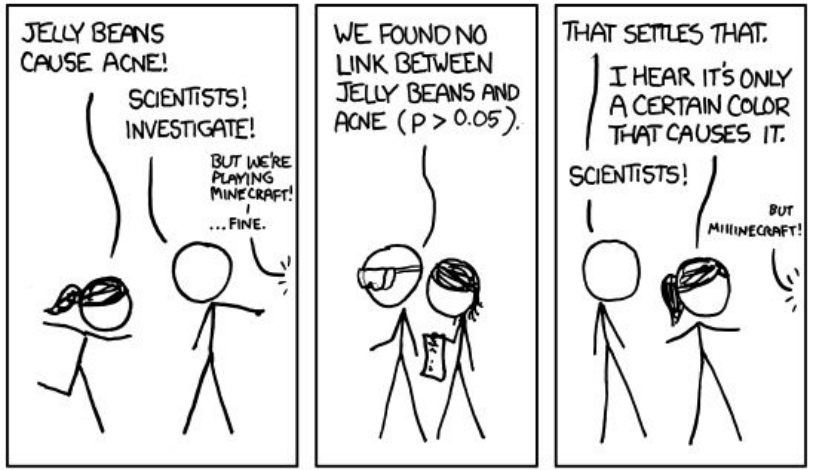
March 7, 2016

MAR. 7, 2016, AT 10:23 AM

Statisticians Found One Thing They Can Agree On: It's Time To Stop Misusing P-Values

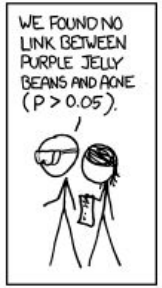
P-hacking

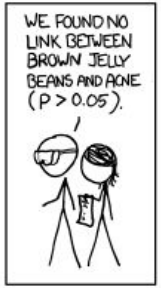
- In addition to p-hacking, p-values open up the problem of *multiple comparisons*
- Since we know that at least 5% of the time the null distribution is true, we will still find a “significant effect” or $p < .05$...
- ...we can just keep trying different samples, or different measurements until we get a significant result

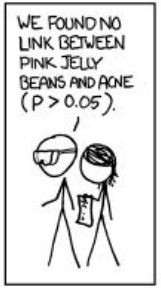

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.

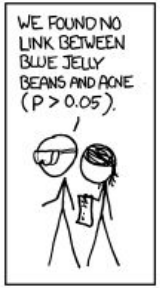
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

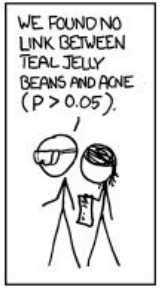
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

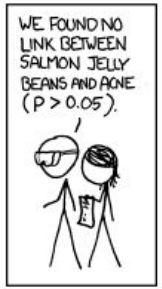

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).

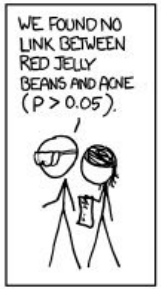

WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).

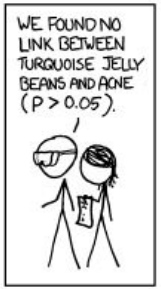

WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).

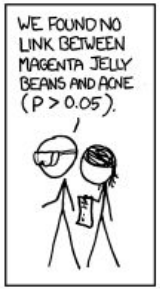

WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).

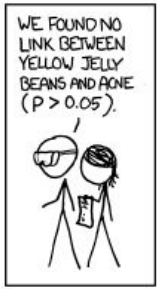

WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

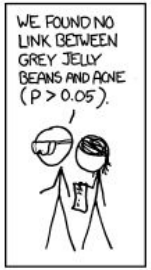

WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).

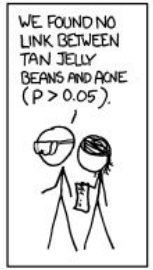

WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).

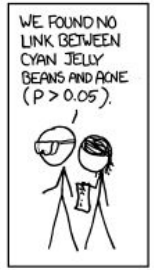

WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).

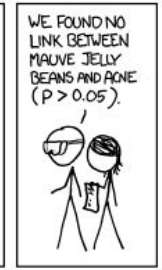

WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

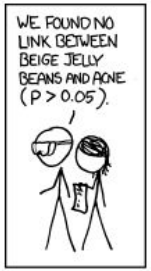

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

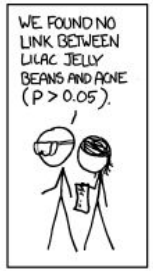

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

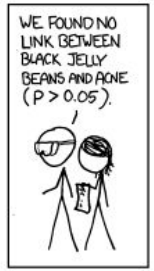

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

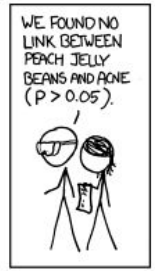

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!

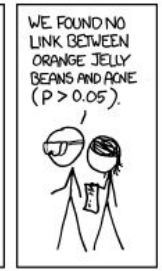

WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

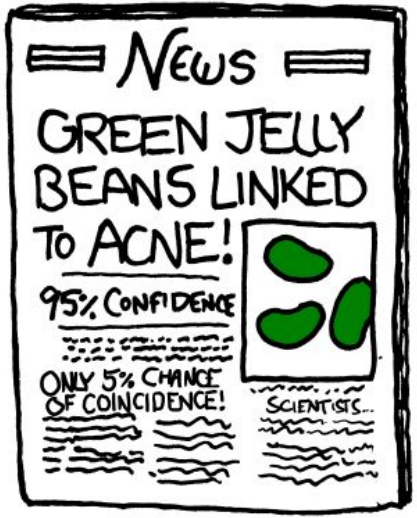

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

P-hacking and publication bias

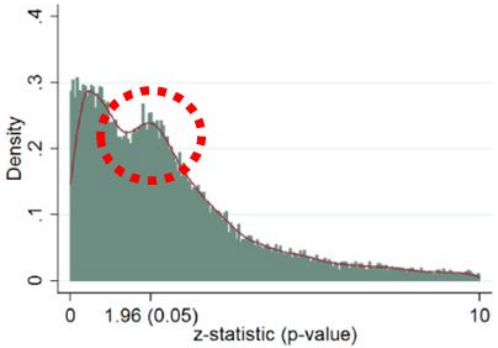
Why Most Published Research Findings Are False

John P. A. Ioannidis

Economics

Brodeur et al (AEJ:A, in press)

"Star Wars: The empirics strike back"

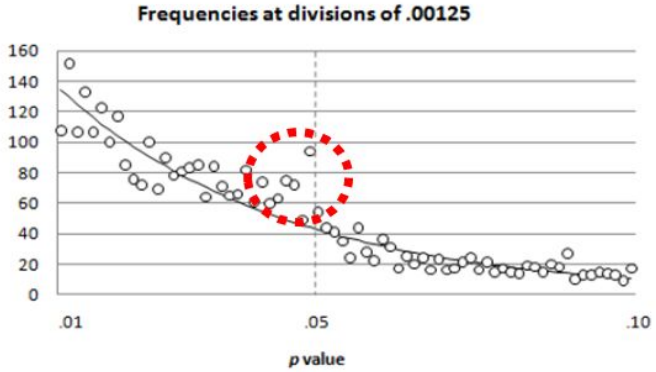


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

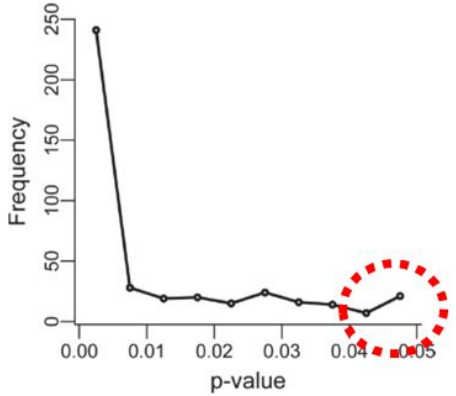
"A peculiar prevalence of p values just below .05"



Biology

Head et al (PLOS Biology 2015)

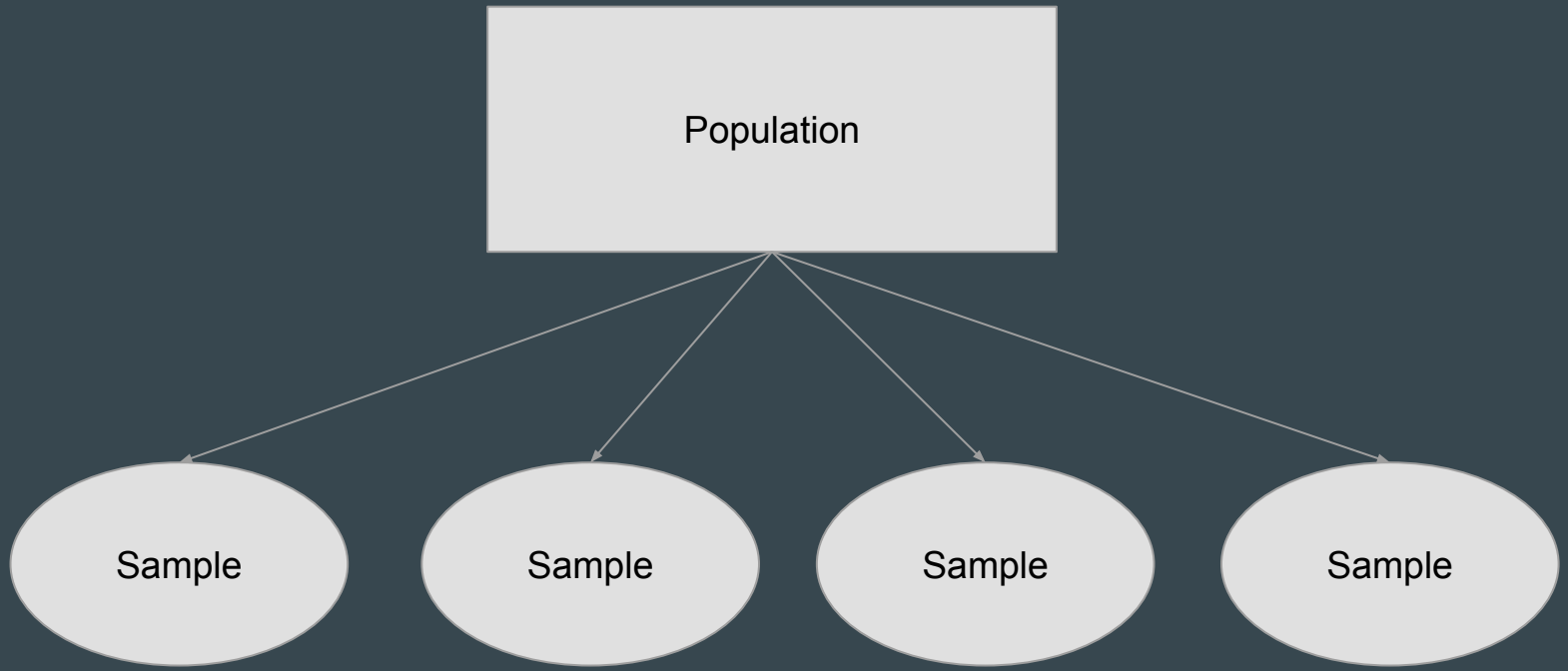
"Extent and Consequences of P-Hacking in Science"



Instead, we can use the bootstrap

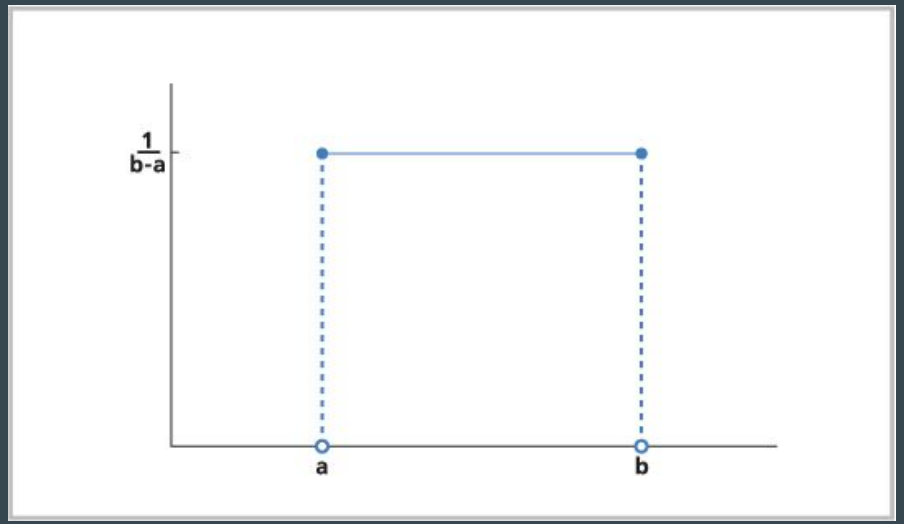
- This will allow us to model our alternative distribution directly
- Needs no assumptions about the distribution
  - No CLT
  - These are called “non-parametric” or “non-functional form” methods
- Can estimate any statistic, not just mean or sums
- “Pull the data up by its own bootstraps”

Review: sampling



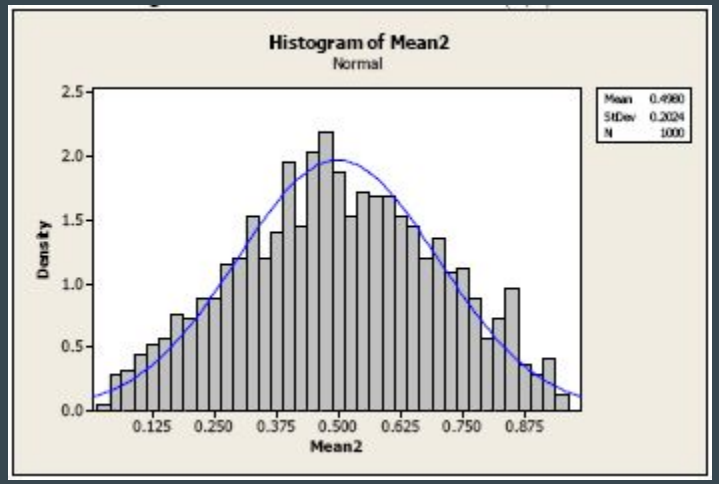
Review: random sampling

- Random sampling applies a uniform distribution to draws from the population
- In other words, each observation in the population has an *equal chance* of being drawn randomly

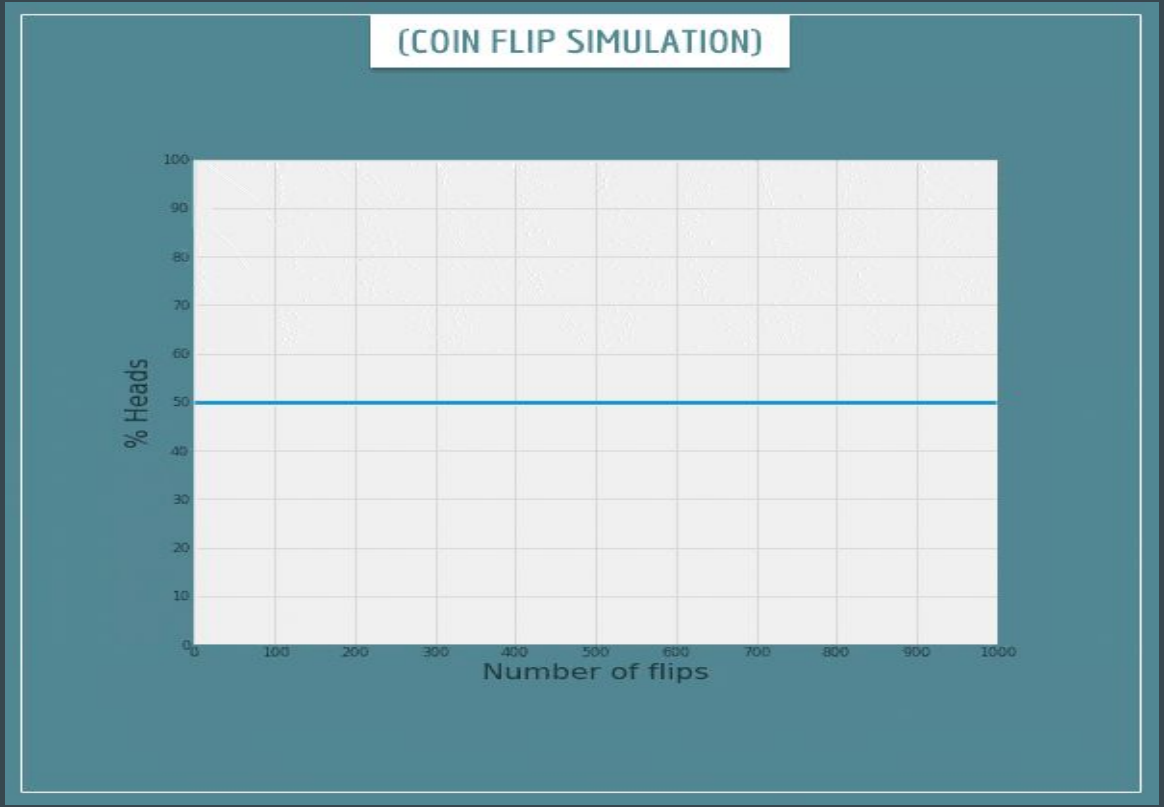


Review: law of large numbers

- Assuming we have a perfectly random sample, as that sample gets larger, *the probability distribution for the sample converges to the probability distribution of the population*



Review: law of large numbers



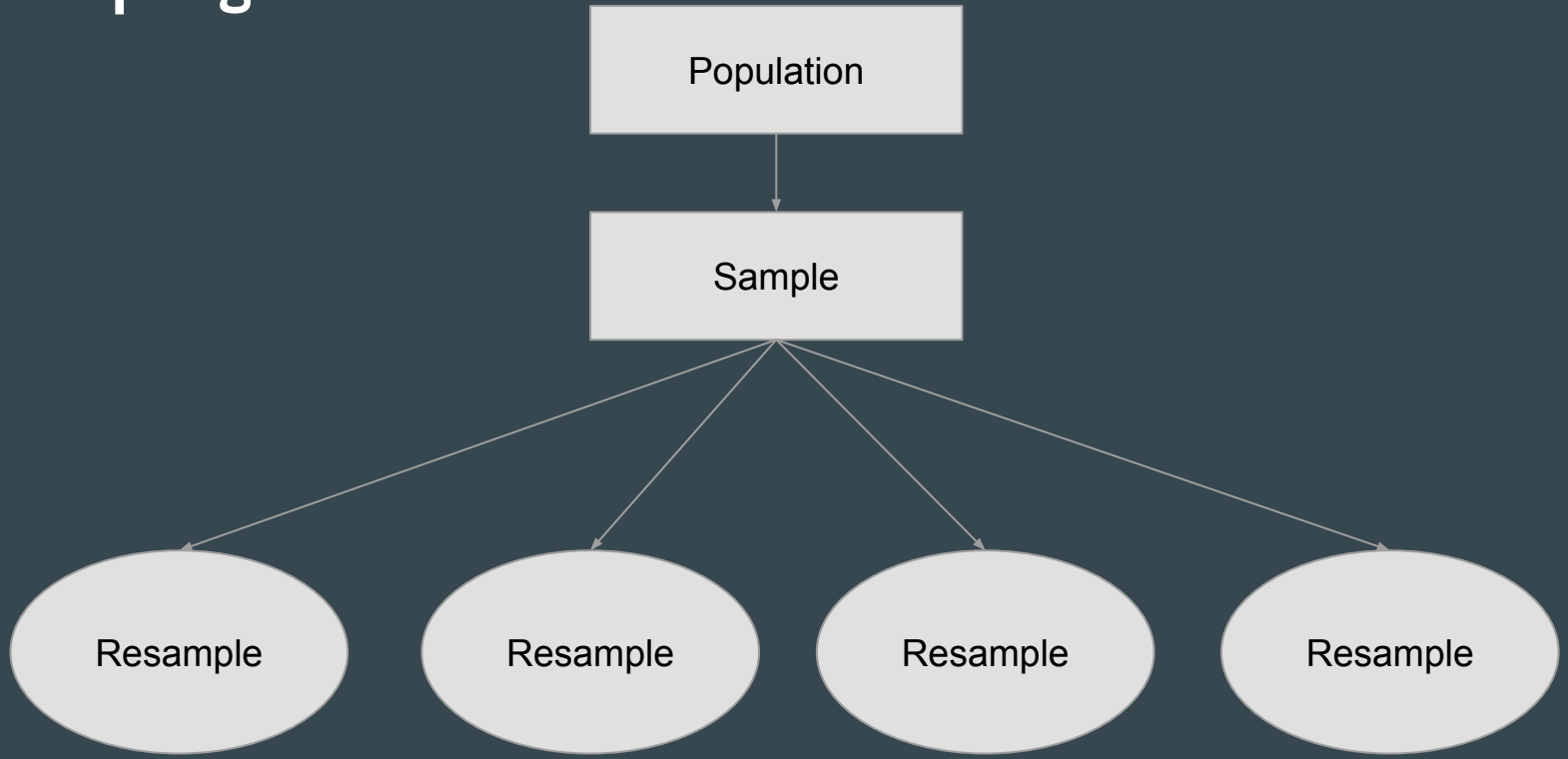
Consequences of law of large numbers

- Law of large numbers say that the sample will eventually converge to the population
- In other words, if we could take as many samples as we wanted, we wouldn't need any null distribution, p-value, etc. *because we would know the true values of the pouplation*

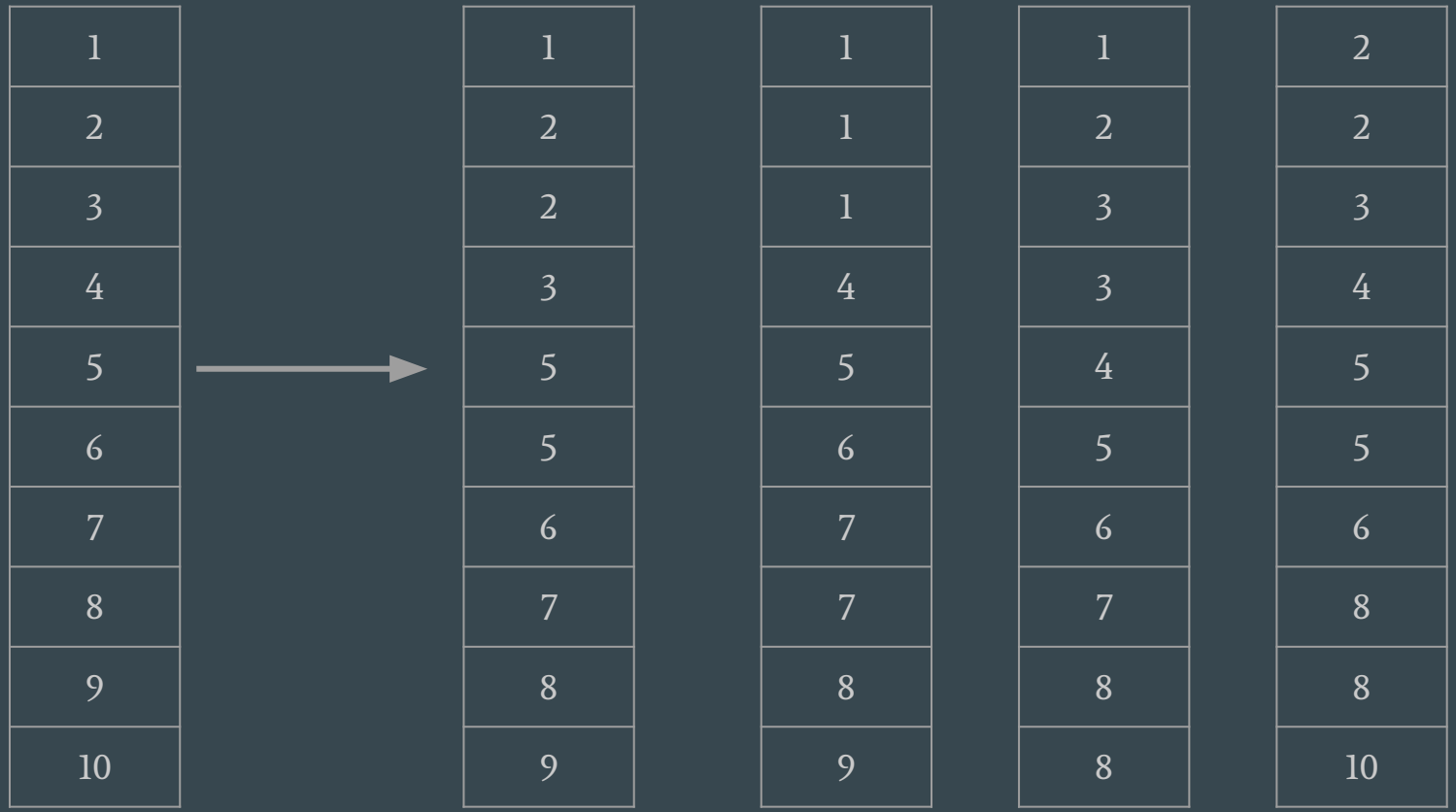
Resampling

- Resampling is the key to the bootstrap
- It mimics the behavior of Law of Large Numbers by creating different combinations of the original sample to recreate the population

Resampling



Resampling


1 2 3 4 5 6 7 8 9 10
1 2 2 3 5 5 6 7 8 9
1 1 1 4 5 6 7 7 8 9
1 2 3 3 4 5 6 7 8 8
2 2 3 4 5 5 6 8 8 10

Resampling

Imagine there is a bowl of 100 marbles.

- A *random sample* would take 30 of those marbles, picked at random, out of the bowl
- A *resample* takes those 30 marbles, and then draws one at a time, recording the value, *and then placing each marble back in the bowl before the next one is drawn*
    - This is called *sampling with replacement*

Logic of sampling with replacement

- A random sample is *representative*, because each observation has an equal chance of being drawn from the population
- When we sample with replacement, we are arguing that the variation in our sample should approximate the variation in our population
- So the “replaced” values (i.e. those sampled twice) stand in for new observations from the population

The bootstrap set-up

1. $x_1, x_2, \ldots, x_n$ is a random sample from a population
2. The population of interest follows a distribution F
    a. Can be any distribution, no assumptions made
3. We are interested in a statistic, μ, from that distribution
    a. I.e. mean, median, etc.
4. We take a resample of our original data, called $x^*_1, x^*_2, \ldots, x^*_n$
5. F* is the distribution of our resamples
6. μ,* is the statistic calculated from our resample

The bootstrap principle

1. F* approximates F
2. The variation in μ, is approximated by μ*

Based on #2, we can calculate a confidence interval for our statistic of interest,

Bootstrap in action

You are running a campaign. Your candidate currently has an approval rating of 45%. After you announce a new policy, you take a random sample of 100 likely voters.

Sample: 100 likely voters

- Observation: voters
- Variable: approve (yes/no; 1/0)

Population: All likely voters

- Sample is representative, because it is random

Bootstrap in action

You are running a campaign. Your candidate currently has an approval rating of 45%. After you announce a new policy, you take a random sample of 100 likely voters.

- Sample with replacement 20 times:
  - Each sample with replacement also contains 100 observations
  - Due to replacement, *these will not be the same 100 observations as the original sample*
- Calculate statistic of interest, μ^* for each resample

Bootstrap in action

You are running a campaign. Your candidate currently has an approval rating of 45%. After you announce a new policy, you take a random sample of 100 likely voters.

- Distribution of mean(X*):

44.8	46.2	46.8	47.1
47.5	48.2	48.7	49.1
49.9	50.4	51.2	51.8
52.6	53.1	53.6	54.2
54.9	55.1	55.7	56.2

Bootstrap in action

You are running a campaign. Your candidate currently has an approval rating of 45%. After you announce a new policy, you take a random sample of 100 likely voters.

- Now we can test our hypothesis by creating a distribution of difference in means (mean(X) - μ), δ*

-0.2	1.2	1.8	2.1
2.5	3.2	3.7	4.1
4.9	5.4	6.2	6.8
7.6	8.1	8.6	9.2
9.9	10.1	10.7	11.2

Bootstrap in action

You are running a campaign. Your candidate currently has an approval rating of 45%. After you announce a new policy, you take a random sample of 100 likely voters.

- 90% CI for δ (1.2 - 10.7)
- Assuming your random sample is representative of the population of interest, your are 90% the population δ is between 1.2 - 10.7%
- In other words, assuming your sample is random and representative, you are 90% confident that likely voters now favor your candidate between 1.2-10.7% more

Don’t misinterpret the bootstrap

- The bootstrap tells you about the *variation* in δ
- NOT an estimate of the *mean* of δ
- This allows you to make confidence intervals
- The middle point of the confidence interval is NOT the expected difference
    - It’s the median of a difference in means... not a particularly interesting or interpretable quantity

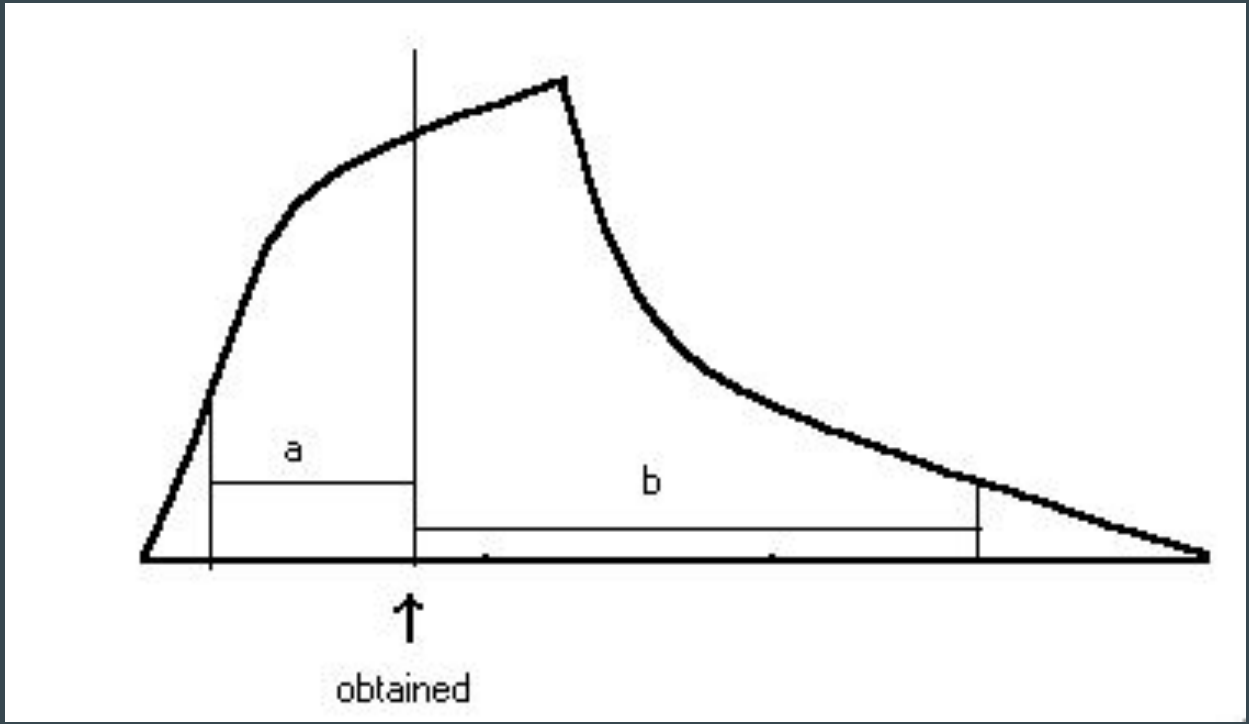
Bootstrap cookbook

1. Take at least 20 samples with replacement from existing sample
2. Calculate the mean of each of those resamples
3. Subtract the sample mean from the population mean μ to get δ
4. Take the α, 1 - α percentiles of your data for a 1 - α% confidence
    a. 80% confidence: 10th and 90th percentile
    b. 90% confidence: 5th and 95th percentile
    c. 95% confidence: 2.5th and 97.5th percentile
    d. 99% confidence: 0.5th and 99.5th percentile

Benefits of the bootstrap

- You are no longer making tortured arguments about the null hypothesis
- You instead model δ, or the variation of mean(H_A) - mean(H_0)
- Assuming a random sample, you can now say something about the likelihood any change is due to random chance
- Still can't say anything about the distribution H_A directly

Benefits of the bootstrap

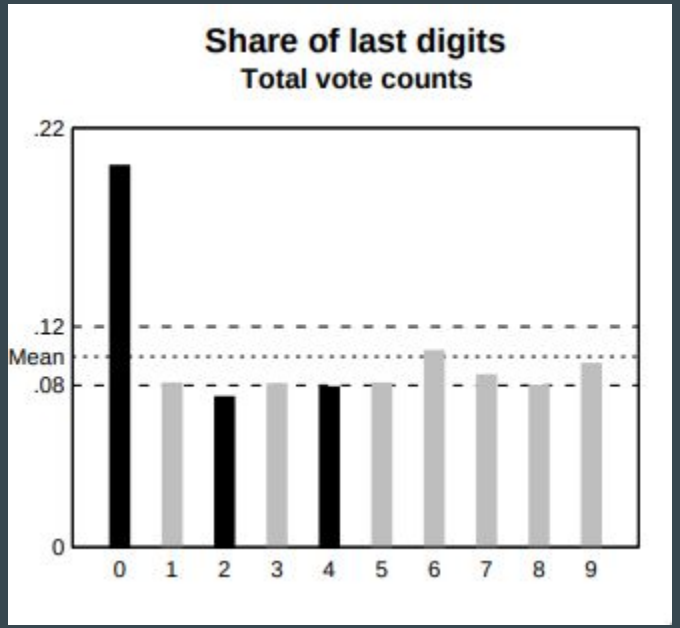


Bootstrap vs. NHST

- NHST is based on the central limit theorem
  - Requires X to be i.i.d.
  - *Also requires X to be a random sample*
  - Only works for sum-based quantities
  - Can only interpret with respect to an assumed null distribution
- Bootstrap approximates the population distribution
  - Does not require X to be independent
  - Requires X to be a random sample
  - Directly interpret variability in difference in means (your quantity of interest)
  - Can create confidence intervals for any statistic (mean, median, etc.)

So why not use the bootstrap?

- Bootstrap wasn't invented until 1979, because humans are unable to truly randomize



Main assumption for the bootstrap

A *random* and *representative* sample

- Random sample: avoids sampling error due to bias (systematic)
  - Selection bias
  - Aggregation bias
  - Desirability bias
- Representative: must be large enough to accurately capture population variation
  - If Var(X) differs from the variance in the population, your *resamples* will be biased
  - No easy answer to how large your sample needs to be
    - Populations with greater variance need larger samples
    - Populations with smaller variance get away with smaller samples

Systematic errors

- Called systematic because they are predictable features of the sampling technique

Randomization errors

- These happen when the sample differs from the population due to pure random chance
- For NHST, when mean(X) differs due to random chance, we get a false positive
  - Doesn't depend on N (size of sample)
- For the bootstrap, variability in X needs to differ to get a false positive
  - You need a similar range of values as the population
  - Otherwise resamples won't stand in for traditional sampling
  - Depends heavily on N (size of sample)

Review: bootstrap

- Law of large numbers states that a large enough sample will converge on the population distribution
- It's also true many random samples will be converge around the population distribution
- Bootstrap simulates this many samples logic by *sampling with replacement*

Review: bootstrap

- Does not require assuming a null distribution
- Does not require a standard deviation estimate for the population
- Does not require our data to follow a particular distribution
- Does require an entire dataset, not just a mean
- Does require a computer to effectively resample

Key intuition of the bootstrap

Hypothesis testing is about distinguishing *random fluctuations* from *systematic changes.* The bootstrap uses random sampling to mimic the probabilistic process of random chance.

Key intuition of the bootstrap

Hypothesis testing is about distinguishing *random fluctuations* from *systematic changes.* The bootstrap uses random sampling to mimic the probabilistic process of random chance.

"Regularization"