Section 2: Estimation and A/B Testing

Vadim Sokolov

Suggested Reading

OpenIntro Statistics, Chapters 4,5&6

Last Section

How to deal with uncertainty?

- Random Variables and Probability Distributions
- Joint and Conditional (Happy/Rich), Independence (Sally Clark),
- Expectation and Variance (Bookies vs Betters, Tortoise and Hare)
- Binomial Distribution (Patriot Coin Toss), Normal distribution (Crash of 1987)
- Decision Making under uncertainty (Marriage Problem and Probability and Decision Trees)
- Bayes Rule (Practice Hard $\neq$ Play in NBA)

This Section

- Estimating Parameters and Fitting Distributions
- Confidence and Prediction Intervals
- Means, Proportions, Differences
- A/B Testing

Why R? R is free!

All the code for the course and assignments are in `Rexamples.R`!

- Most Data Scientists are using R.
- R's syntax is very simple.
- There are a large number of add-on statistical packages that can be installed and run in R. You don't need to code anything.

You can use your favorite software: `Excel, Stata, Matlab`.

Statistics with R

You can download `R` and `Rstudio` for Windows, Linux, Mac at

`http://www.r-project.org/`

`http://www.rstudio.com/`

`Rstudio` is a very useful front-end interface to `R`.

- Links, code for class, video tutorials are up on the course-page.

A list of books on doing statistics in R is at

`http://www.r-project.org/doc/bib/R-books-html`

`http://www.r-bloggers.com`

Start by watching `R` videos for Libraries and Packages/Files and Data.

Standard R Commands

Given vectors *x* and *y* we can apply `mean( ... )` computes the sample mean

`median( ... )` computes the median

`var( ... )` computes the sample variance

`sd( ... )` computes the sample standard deviation

`cov( ... )` computes the sample covariance

`cor( ... )` computes the sample correlation

`pnorm( ... )` calculates normal probabilities

`hist( ... )` makes histograms

`lm( ... )` for linear model (a.k.a regression)

`summary( ... )` provides a summary analysis of the output

Stats in Excel

`average( ... )` computes the sample mean

`median( ... )` computes the median

`var( ... )` computes the sample variance

`stdev( ... )` computes the sample standard deviation

`covar( ... , ... )` computes the sample covariance

`normsdist( ... )` calculates normal probabilities

`histogram( ... )` makes histograms

Regression commands include `linest(yrange,xrange)` for linear model (a.k.a multiple regression)

`slope(yrange,xrange)` provides the regression β

Google 2019

Let X be daily returns. We assume that returns are independent and identically distributed as

$$X \sim N\left(\mu, \sigma^2\right)$$

Question: What are the values of μ and σ? Let's assume that each observation in the random sample $\{x_1, x_2, \ldots, x_n\}$ is independent and distributed according to the model above, $X_i \sim N(\mu, \sigma^2)$. Then we use the sample mean, $\bar{x}$, and sample variance, s^2,

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

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

Google 2019

Prices and Returns

Data: $z_1, z_2, \ldots, z_n$

estimated

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

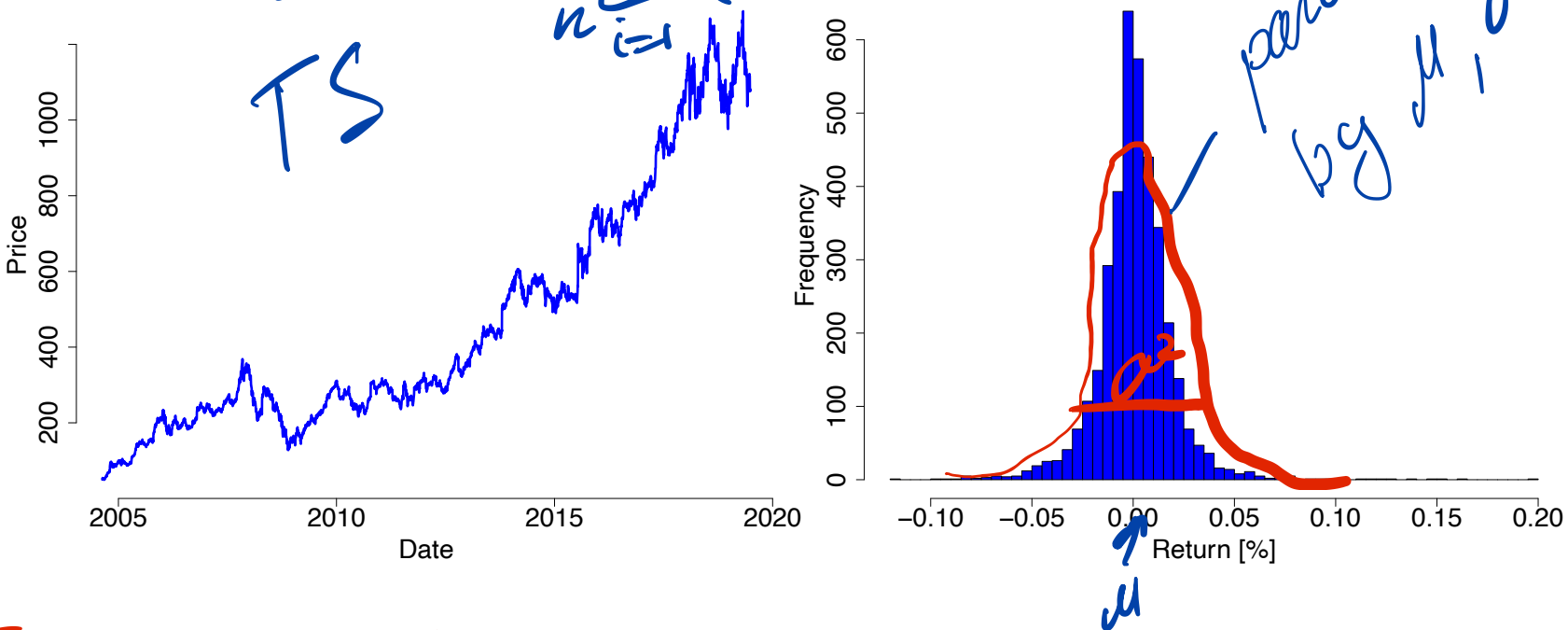
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (\hat{\mu} - z_i)^2$$

efficient

Daily Returns

TS

parametrized by μ, σ^2



- Fit normal model into data (histogram)?
- How to estimate μ, σ^2?

Google 2019

Assume that I invest all my money in Google! Your job is to tell me the following:

1. What is my expected one year return?
2. How about the risk? Volatility?
3. What's the probability that I lose at least 3% in a day?

Google 2019

We can use our model and R to answer our questions

```
>mean(ret) [1] 0.0007078
>sd(ret)   [1] 0.01864
```

Annual expected return $0.0007078 * 252 = 17.8\%$ a year Daily risk or volatility of 1.86% How about losing money?

$$Z = \frac{X - \mu}{\sigma} = \frac{-0.03 - 0.00070}{0.0186} = -1.65$$

```
>pnorm(-1.65,0,1)
[1] 0.049
```

There's a 4.9% chance I can lose at least 3% in a day

The Genius at the Royal Mint (Sir Isaac Newton, 1643-1727)

Ideal Weight: 100g Actual Coins: 100 ± 10g

The goal is to guarantee: 100 ± 1g

Bad money drives out the good:

1. English coins were worth less as currency in England than as precious metal in Europe.
2. Large variability in the weights of coins make it hard to stop coin clipping.

Trial of the Pyx was held by Newton to check whether the coins are good.

2500 coins into box weight

$2500 \cdot 100 \pm 2500$. Correct?

The variation of total is smaller than or equal to the individual coin

Ineffectiveness of the Trial

Batches of $n = 2,500$ coins, each supposed to weigh 100 grams with an allowable margin of error of ± 1 gram. Trial set the bounds for the average weight to be ± 1 gram! A bad mistake.

$$\text{Variability of an Average} = \frac{\text{Variability of a Single Measurement}}{\sqrt{\text{Sample Size}}} = \frac{\sigma}{\sqrt{n}}$$

Sd of individual coin

$$\frac{\pm 1\,g}{\sqrt{2500}} = 0.02$$

According to modern statistics, the bound should be:

$$100 \pm \frac{1}{\sqrt{2500}} = 100 \pm 0.02$$

$$\pm 2500 \cdot \sqrt{n}$$

$$2500 \cdot 100 \pm 2500 \cdot 0.02$$

Sampling Distribution of Sample Mean

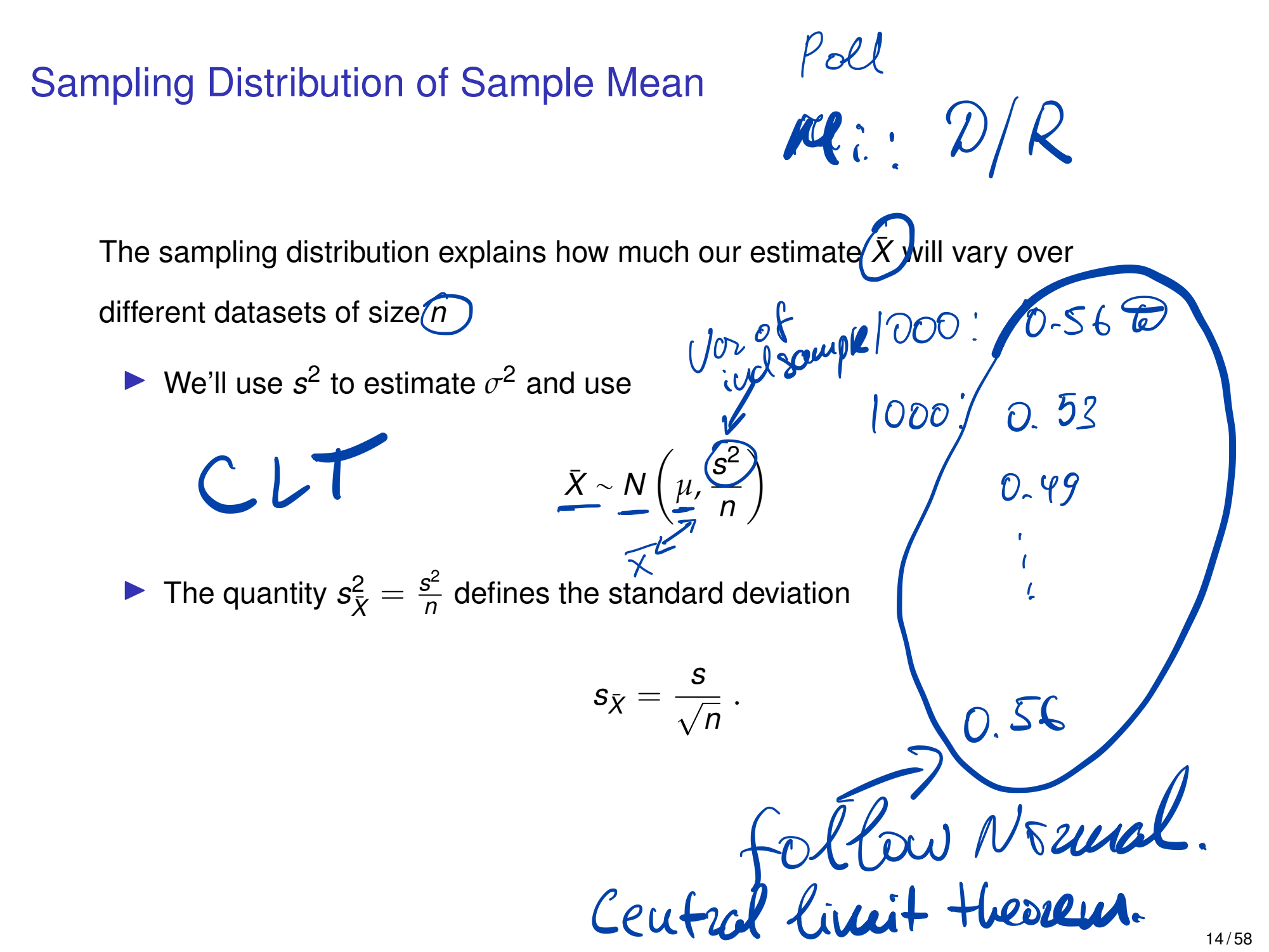
The sampling distribution explains how much our estimate $\bar{X}$ will vary over different datasets of size n

- We'll use s^2 to estimate σ^2 and use

$$\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$$

- The quantity $s^2_{\bar{X}} = \frac{s^2}{n}$ defines the standard deviation

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} .$$



One sample

$(0, 1, 0, 0, 1 \ldots \cdot 1)$

Calculate Variance of individual sample

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.53$$

↑ estimated from sample

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{p})^2 = \sigma^2 = \hat{p}(1-\hat{p})$$

We are uncertain about $\hat{p}$

$$\hat{p} \sim N\left(0.53, \frac{\sigma^2}{n}\right)$$

↑ avg from sample

s.d of $\hat{p}$ = s.d of sample / $\sqrt{n}$ ✓

Var of $\hat{p}$ = Var of ind. sample / n

Confidence Intervals

In summary, our best guess at μ is $\bar{x}$

- How large a mistake can we make? The distribution is $\bar{X} \sim N\left(\mu, s_{\bar{X}}^2\right)$ where $s_{\bar{X}} = s/\sqrt{n}$.
- $[\bar{x} \pm 1.96 s_{\bar{X}}]$ give us a 95% range of plausible values for μ

  We call it a 95% Confidence Interval (C.I.)

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$$

We are 95% conf. that true value of $\bar{x}$ is inside this interval

What if I want 90% or 99% confidence?

Prediction intervals

1. A confidence interval estimates the mean, μ.
2. A *prediction interval* answers the question:

"What range of values are plausible for a single future observation"?

$100(1-\alpha\%)$ of the data lie in the interval

$$\overline{x} \pm z_{\alpha/2}\, s\sqrt{1+1/n}.$$

For example, if $\alpha = 0.05$, we have $z_{0.05/2} = 1.96$.

Standard Error and Confidence Interval for a Proportion

if Sample = (0, 1.0, .., 1) avg

Suppose that we have a proportion $\hat{p} = X/n$ rather than a mean $\frac{1}{n}\sum_{i=1}^{n} x_i$

We can compute the standard error for the proportion:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

called proportion

Then we can calculate a 95% Confidence Interval

Var of ind Sample

$$\hat{p} \pm 1.96 \times s_{\hat{p}}$$

Var of $\hat{p}$ is $\frac{\hat{p}(1-\hat{p})}{n}$

$\hat{p}(1-\hat{p})$

Example: Google Search Algorithm

Google is testing a new search algorithm. 2,500 searches and check how often people clicked through. Here's our data:

Table: Google Search Algorithm

	Algo1	Algo2
success	1755	1818
failure	745	682
	2500	2500

The probability of success is estimated to be $\hat{p}_1 = 0.702$ for the current algorithm and $\hat{p}_2 = 0.727$ for the new algorithm.

Is the new algorithm better? For sure??

Example: Google Search

Are we 95% confident that Algo 2 is better?

We could calculate 95% confidence intervals separately

For Algo 1: $\hat{p}_1$

$$0.702 \pm 1.96\sqrt{\frac{0.702 \times 0.298}{2500}} = (0.693, 0.711)$$

s.d of $\hat{p}_1$

For Algo 2: $\hat{p}_2$

$$0.727 \pm 1.96\sqrt{\frac{0.727 \times 0.273}{2500}} = (0.718, 0.735)$$

What do you think?

Maybe its more accurate to do the difference!!

$0.711 < 0.718$

Intervals do not overlap.

The diff between $\hat{p}_1$ & $\hat{p}_2$ is statistically significant @ 95%

Standard Error for the Difference in Means

$\hat{p}_1 - \hat{p}_2 \sim N(\hat{p}_1 - \hat{p}_2, \ldots)$

Suppose that we have two samples

We can compute the standard error for the difference in means:

$s.d\ \hat{p}_1 - \hat{p}_2 = \sqrt{Var(\hat{p}_1) + Var(\hat{p}_2)}$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s^2_{\bar{x}_1}}{n_1} + \frac{s^2_{\bar{x}_2}}{n_2}}$$

We can compute the standard error for the difference in proportions:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$

$Var(X - Y) = Var(X) + Var(Y)$

Confidence Interval for the Difference in Means & Proportions

We can now compute the confidence interval for the difference in means:

$$(\hat{X}_1 - \hat{X}_2) \pm 1.96 \times s_{\bar{X}_1 - \bar{X}_2}$$

or the confidence interval for the difference in proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \times s_{\hat{p}_1 - \hat{p}_2}$$

Example: Google Search

More accurate to calculate CI for diff. compared to ind. intervals.

Now we get

$$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{0.702 \times 0.298}{2500} + \frac{0.727 \times 0.273}{2500}} = 0.0128$$

The Confidence Interval becomes

$$(0.702 - 0.727) \pm 1.96 \times 0.0128 = (-0.05, 0.00)$$

$\hat{p}_1 - \hat{p}_2 < 0$

-10^{-6}

$\hat{p}_1 < \hat{p}_2$

What's our conclusion now?

C 95% CI for $\hat{p}_1 - \hat{p}_2$

$(-0.05, -0.01)$

Let's revisit the Patriots example ...

$$\hat{p} = \frac{19}{25}$$

The data tells us that the Patriots have won 19 out 25 tosses.

Assume they were using the same coin the entire time and that the Patriots always choose heads...

What is the data telling us about the probability of heads in this coin?

Patriots and Coin Tosses

Is true value of $\hat{p} = 0.5$?

Our best guess at p is $\hat{p} = \frac{19}{25} = 0.76$

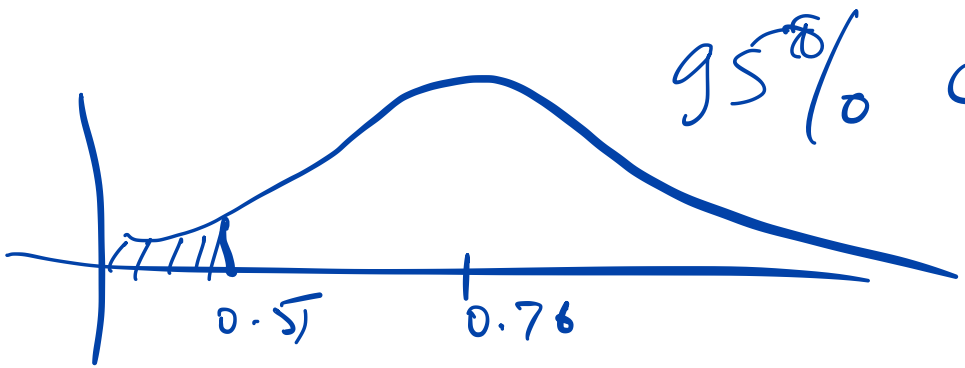
We have a 95% Confidence Interval given by

$$0.76 \pm 1.96\sqrt{\frac{0.76 \times 0.24}{25}} = 0.76 \pm 1.96 \times 0.0854$$

The interval is $(0.592, 0.927)$.

Did it happen by chance?

Why so wide? What do we conclude??



95% conf not by chance

Conclusions

There's a number of things:

1. Estimates are based on random samples and therefore random (uncertain) themselves

   We need to account for this uncertainty!

2. Standard Error measures the uncertainty of an estimate

3. We construct 95% Confidence Intervals

   This provides us with a plausible range for the quantity we are trying to estimate.

A/B Testing :

Marketing: Is new marketing comp better than the previous.

Internet: What should be the color of bg. of my page

The key elements of hypothesis testing are:

1. Setting Up a Hypothesis Test
2. Significance Levels
3. p-values
4. Testing a Mean and Difference in Means
5. Testing a Proportion
6. Tests for Small Samples

Drug trials: Is new drug better than old

Politics: A/B exp for fund raising

Null Hypothesis: Old is better than New

Alternative Hyp: New is better.

Coke versus Pepsi

Pepsi: $> \frac{1}{2}$ prefer Pepsi

Start by assuming H_0 is true

The most famous hypothesis test in history in whether people can decide the difference between Coke and Pepsi

Double Blind: neither the experimenter or subject know the allocation

- Pepsi claimed that more than $\frac{1}{2}$ of Diet Coke drinkers said they preferred to drink Diet Pepsi
- Suppose we take a random sample of 100 drinkers and find that 56 favor Pepsi?

Data: $\hat{p} = \frac{56}{100} = 0.56$

Pepsi claim becomes Alt Hypothesis.

$H_0: p < \frac{1}{2}$ $H_1: p > \frac{1}{2}$

$H_0: p < \frac{1}{2}$ $H_1: p > \frac{1}{2}$

if P-value $\leq 0.05 \Rightarrow$ accept H_1

$P > 0.05 \Rightarrow$ Cannot reject H_0

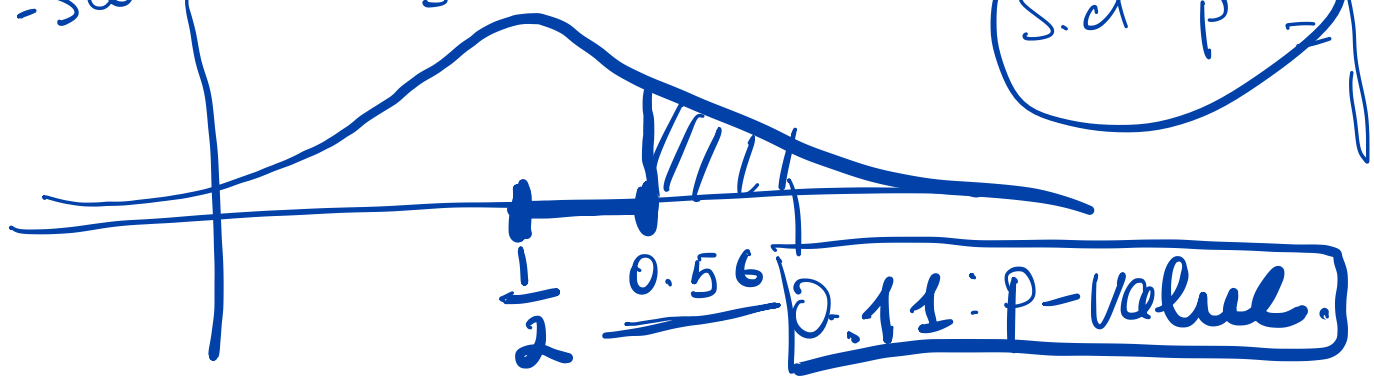
Assume H_0 is true. $\hat{p} = 0.56$

Sign. level: 95%

$$\text{T-score} \\ \text{z-score} = \frac{\frac{1}{2} - 0.56}{\text{s.d. of } \hat{p}}$$

$$\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{100}$$

$$\text{s.d } \hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$$



0.11: P-value.

$P(p > 0.56)$? $1 - p\text{norm}(0.56, \frac{1}{2}, \text{s.d } \hat{p})$ by Excel

There is 11% chance to observe 0.56 or greater given that H_0 is true.

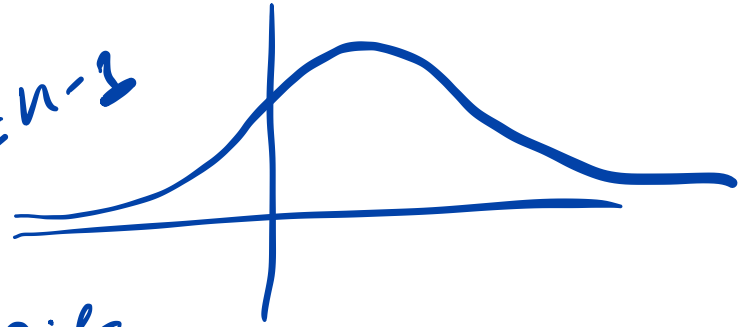
A/B Testing

Scientific Testing is the underpinning of many things ...

Controlled experiment (designed) or observational study

1. A Statistic. Sample mean, proportion, difference in means, ...
2. A Null hypothesis: $H_0 : \mu = \mu_0, H_0 : p = p_0, \ldots$ A level of significance α
3. Sampling distribution: The probability distribution of the statistics values (Normal, t)

Z and T-Score



How many standard deviations are you away from the mean? We'll address this with a Z or T-score

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

If σ is unknown, we'll use s instead

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

For small samples, Student's t distribution with $(n-1)$ degrees of freedom.

Guinness: Testing the quality of beer in 1908.

What is Hypothesis Testing?

A hypothesis is a statement about a population developed for the purpose of testing with data

- Step 1: Null Hypothesis (H_0): assume to be true unless there is sufficient evidence to the contrary.

  Alternative Hypothesis (H_1): test against the null. If there is evidence that H_0 is false, we accept H_1.

- Step 2: Select the significance level α. $\alpha = 0.05$ (the 5% level) is the most commonly used. $\alpha = 0.01$ (the 1% level) is prevalent in medical and quality assurance examples.

Hypothesis Testing

Making a decision

- Step 3: Compute the Test Statistic (Z or T) or p-value
- Step 4: Formulate the Decision Rule

  For example, reject the Null hypothesis if $|Z| > 1.96$ or p-value < 0.05
- Step 5: Make a Decision, Compute the p-value.

  p-value: The smallest significance level at which a null hypothesis can be rejected.

1) Reject Accept H_1

2) Cannot Reject H_0

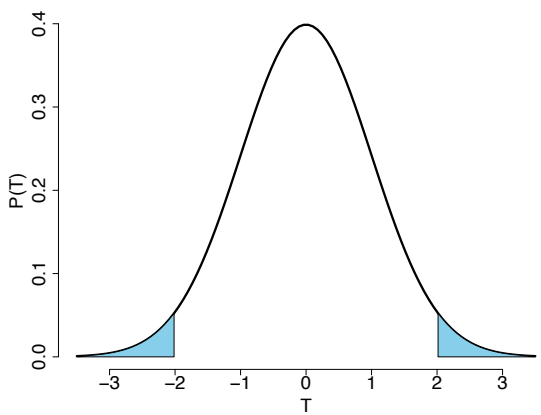
Two Sided Test vs One Sided Test

We calculate sample statistic $\bar{x}$ from data.

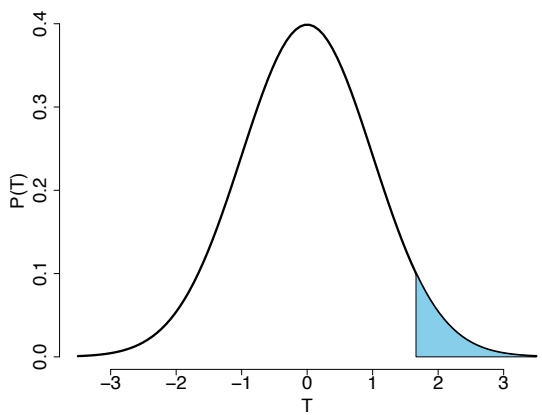
Are we sure that the true value x is different from some value μ?

It can be different in three ways: not equal, greater or less.

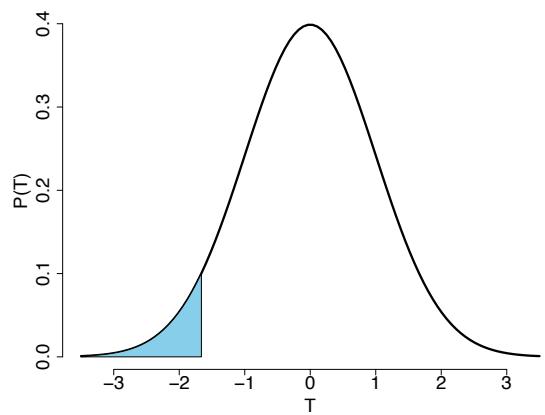
We calculate T-score.



H_0: $x = \mu$; H_1: $x \neq \mu$



H_0: $x \leq \mu$; H_1: $x > \mu$



H_0: $x \geq \mu$; H_1: $x < \mu$

If T score is inside the white area, cannot reject H_0 at $p = 0.05$.

Recap:

Data: $\{x_1, x_2, \ldots, x_n\}$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

1) S.d. of $\bar{x}$ is lower compared to ind. sample x_i. $\bar{x}$ is uncertain

$$\text{S.d. of } \bar{x} = \frac{\text{s.d of ind. sample}}{\sqrt{n}}$$

if you want to double your certainty about $\bar{x}$, half s.d. $2\sqrt{n}$ $\sqrt{4n}$

2. CLT: $\bar{X} \sim N\left(\bar{X}, \frac{S^2}{n}\right)$ does not depend on distr. of X_i

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

3. Property of Normal 95% CI for $\bar{X}$ is $\bar{X} \pm 1.96\boxed{\frac{S}{\sqrt{n}}}$ Def: → std. error

Hypothesis test: Value calculated from data

Null Hypothesis.

$\underline{d}$, e.g $d = \bar{X}$, $d = \bar{X}_1 - \bar{X}_2$

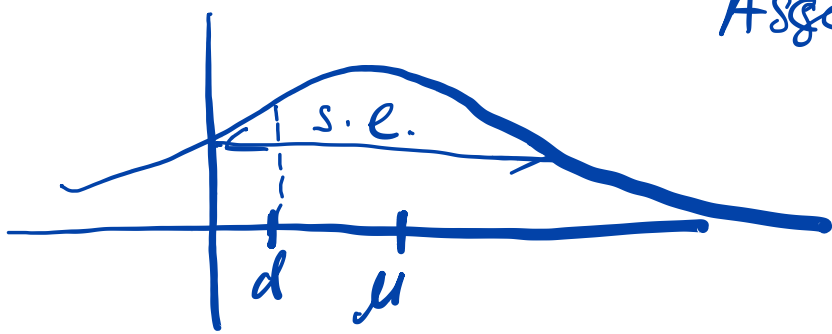
H_0: $\underline{d} \leq \underline{\mu}$ (given) H_1: $d > \mu$

Google search: $d = \hat{p}_2 - \hat{p}_1$, $\mu = 0$

H_0: $\hat{p}_2 - \hat{p}_1 \leq 0$ H_1: $\hat{p}_2 - \hat{p}_1 > 0$

Accepted Truth; Devil's advocate

Assume H_0 is correct

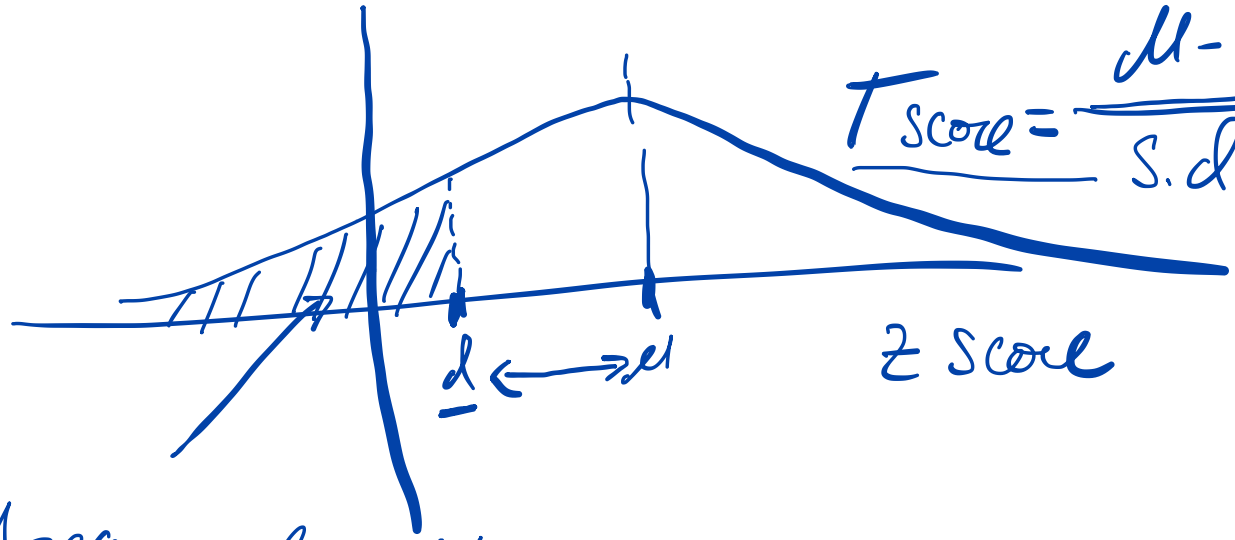


s.e = s.d of d

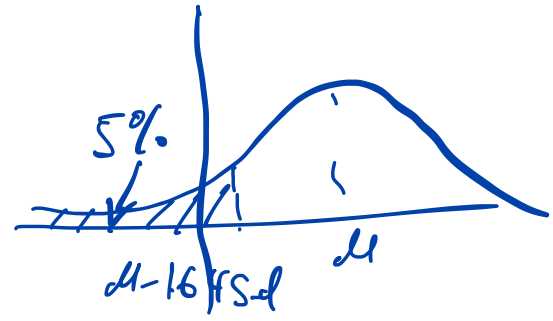
want to know how far d from μ.

Close: cannot reject H_0

Far: Reject H_0 & accept H_1



$$T\ score = \frac{\mu - d}{S.d\ of\ d}$$



Area under the curve

p-value: Probability that d is observed by chance, assuming that H_0 is correct

Reject H_0: 95% level:

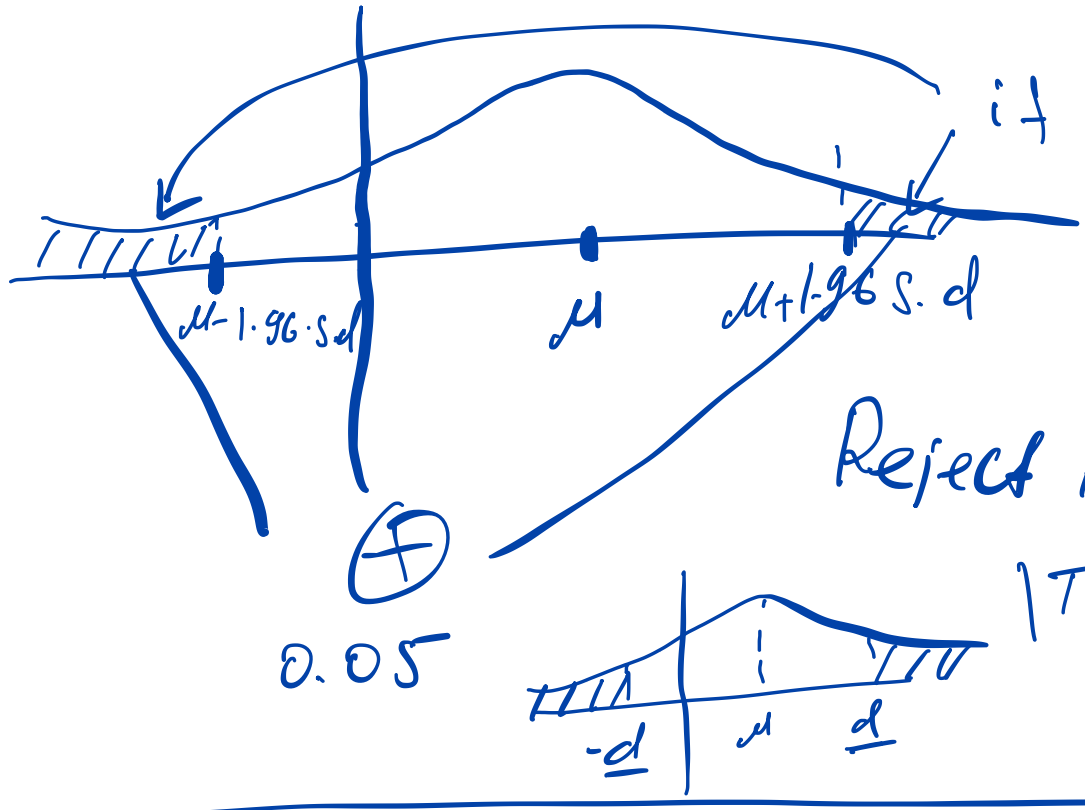
& Accept H_1: p-value ≤ 0.05; $|T\text{-score}| \geq 1.64$

Cannot Reject H_0: p-val ≥ 0.05 or $|T\ score| \leq 1.64$

Goog: H_0 olg Alg_1 is better than new Alg_2

Two-sided test:

$H_0: d = \mu \quad H_1: d \neq \mu$



Reject H_0: p-value ≤ 0.05

$|T\text{-score}| \geq 1.96$

Pepsi: $\hat{p} = 0.56$ proportion of Pepsi fans

$H_0: p \leq 1/2 \quad H_1: p > 1/2$

Revisit Google Algorithm

	Algo1	Algo2
success	1755	1818
failure	745	682
	2500	2500

The statistic we are interested is the difference of proportions

$$\hat{p}_1 - \hat{p}_2 = 1755/2500 - 1818/2500 \text{ and } \mu = 0$$

The search team claims that new algorithms (Algo2) is better. Thus

$$H_0 : p_1 - p_2 \geq 0, \quad H_1 : p_1 - p_2 \leq 0$$

```
prop.test(c(1755,1818),c(2500,2500), alternative = "less", correct = F)
```

Coke versus Pepsi

The most famous hypothesis test in history in whether people can decide the difference between Coke and Pepsi

Double Blind: neither the experimenter or subject know the allocation

- Pepsi claimed that more than $\frac{1}{2}$ of Diet Coke drinkers said they preferred to drink Diet Pepsi
- Suppose we take a random sample of 100 drinkers and find that 56 favor Pepsi?

Coke versus Pepsi

This is a hypothesis test about the proportion of drinkers who prefer Pepsi

$$H_0 : p \leq \frac{1}{2} \text{ and } H_1 : p > \frac{1}{2}$$

My best estimate of the true p

$$\hat{p} = X/n = 56/100 = 0.56.$$

The standard error of my statistic

$$\sqrt{\hat{p}(1-\hat{p})/n} = 0.0496$$

The 95% confidence interval is then

$$0.56 \pm 1.96(0.0496) = 0.56 \pm 0.098 = (0.463, 0.657)$$

Testing

The T-score now with $s_{\hat{p}} = \sqrt{p_0(1-p_0)/n} = 0.05$

$$Z = \frac{\hat{p} - p_0}{s_{\hat{p}}} = \frac{0.56 - 0.5}{0.05} = 1.2 < 1.64$$

Let's take the usual $\alpha = 0.05$. Don't reject H_0 for a one-sided test at 5% level.

We need a larger n to come to a more definitive conclusion.

Common to take $n = 1000$

As a Hypothesis test:

```
prop.test(56,100,alternative="greater", conf.level = 0.95)
```

Type I and II Errors

Marketing Camp: Show ads on weekends

H_0: Weekend ad have no effect

H_1: Have effect

There are two types of errors you can make when testing

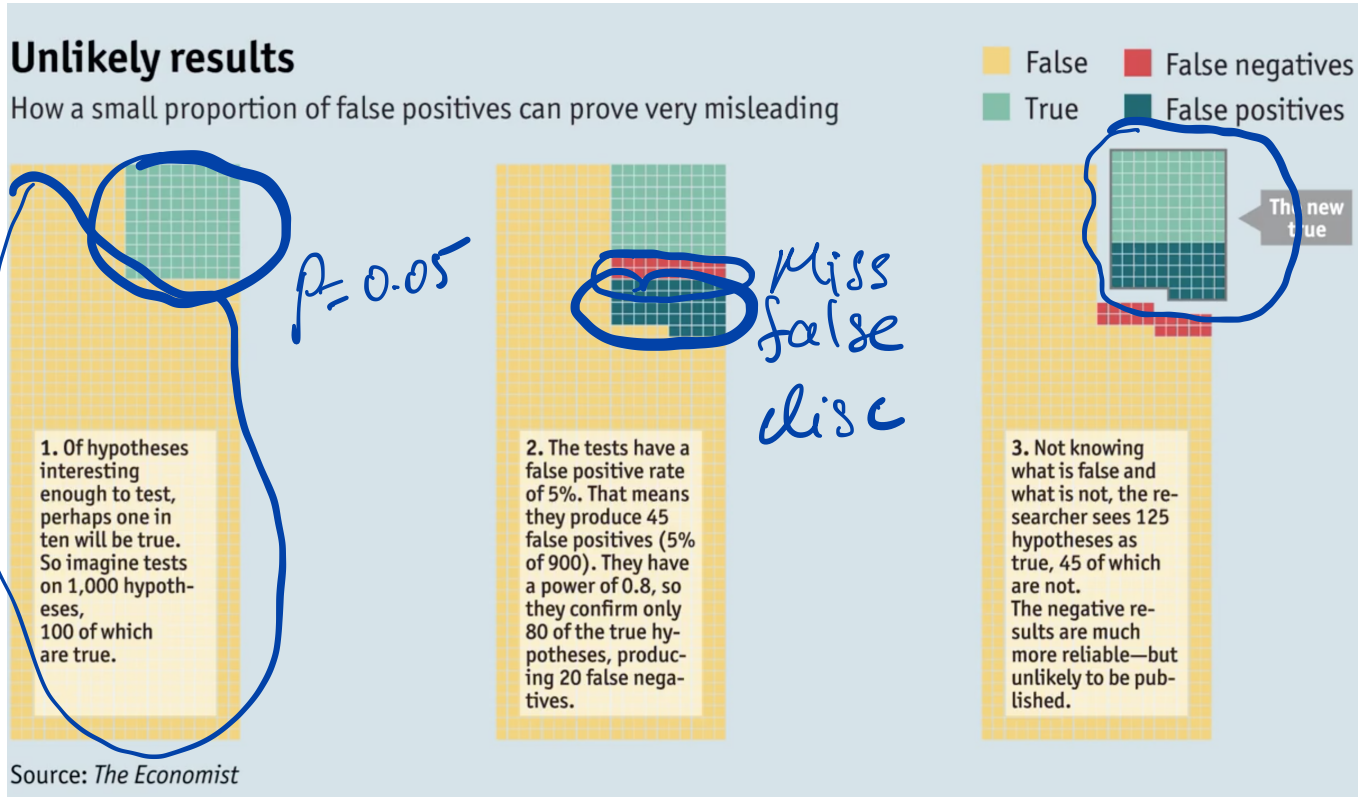
1. Type I Error: Rejecting a true H_0. False discovery
2. Type II Error: Not rejecting a false H_0. Miss discovery

- Significance Level: $P(\text{reject} H_0 \mid H_0 \text{ true}) = P(\text{type I error})$.

p-value = P(False discovery | H_0 true)

Type I and II Errors

Economist

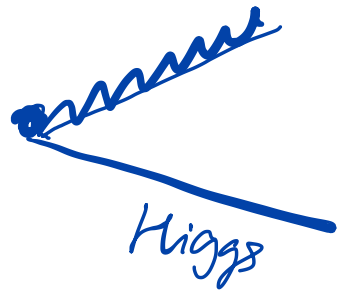


Source: http://www.economist.com/blogs/graphicdetail/2013/10/daily-chart-2

Scientific Reporting

Higgs Boson

2014 Nobel Prize



The five-sigma concept is somewhat counter-intuitive. It has to do with a one-in-3.5-million probability.

H_0: Observed Traj is not from Higgs Boson

H_1: Higgs Boson

That is not the probability that the Higgs boson doesn't exist. It is, rather, the inverse: If the particle doesn't exist, one in 3.5 million is the chance an experiment just like the one announced this week would nevertheless come up with a result appearing to confirm it does exist.

In other words, one in 3.5 million is the likelihood of finding a false positive a fluke produced by random statistical fluctuation that seems as definitive as the findings released by two teams of researchers at the CERN laboratory in Geneva.

p-value one-in-3.5-million and T-score $Z = 5$.

Pfizer

$H_0: \hat{P}_{pop} \leq \hat{P}_{pf}$ $H_1: \hat{P}_{pop} > \hat{P}_{pf}$

$$Var(\hat{P}_{pop}) = \frac{\hat{P}_{pop}(1-\hat{P}_{pop})}{1.5m}$$

$$Var(\hat{P}_{pop} - \hat{P}_{PF}) = Var(\hat{P}_{pop}) + Var(\hat{P}_{PF})$$

$$\hat{P}_{PF} = \frac{77}{6m} = 1.2 \cdot 10^{-5}$$

Pfizer introduced Viagra in early 1998

- During 1998 of the 6 million Viagra users 77 died from coronary problems such as heart attacks.
- Pfizer claimed that this rate is no more than the general population.
- A clinical study found 11 out of 1,500,000 men who were not on Viagra died of coronary problems during the same length of time as the 77 Viagra users who died in 1998.

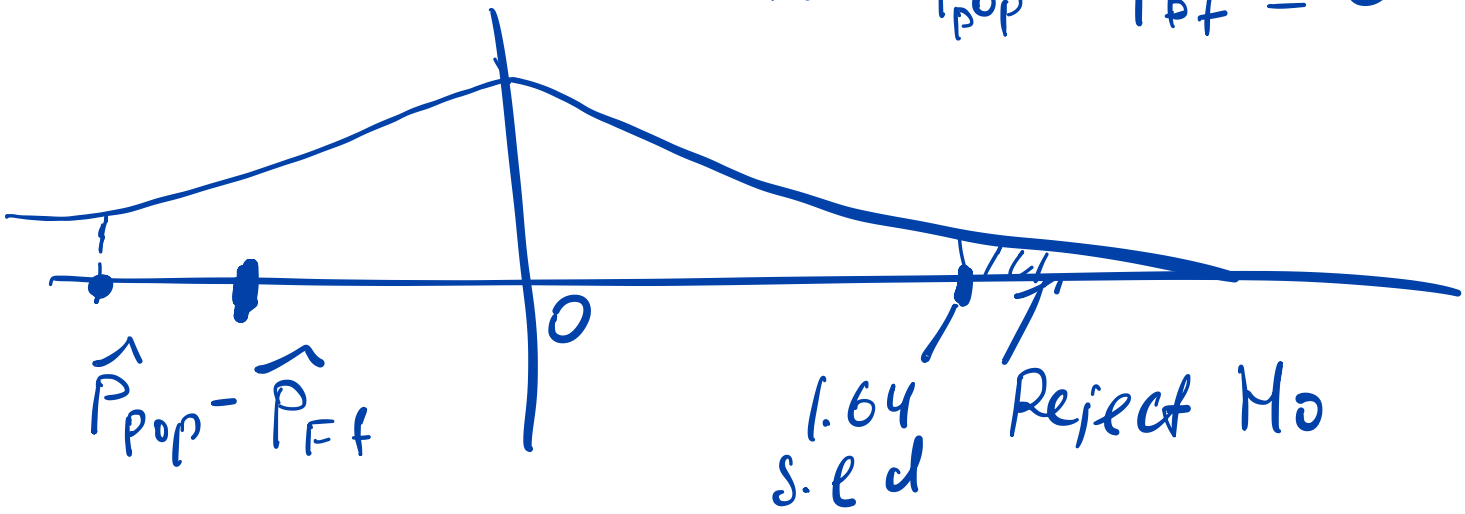
$$\hat{P}_{pop} = \frac{11}{1.5m} = 7 \cdot 10^{-6}$$

Is this statistically significant?

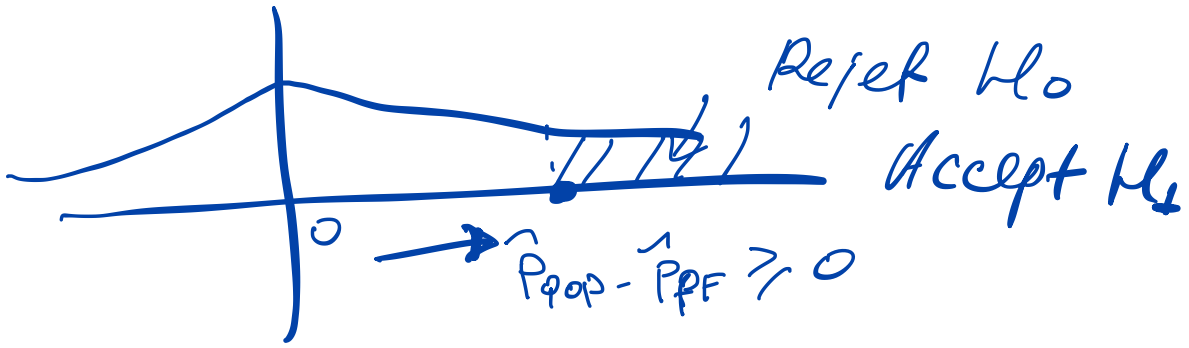
$d = \hat{P}_{pop} - \hat{P}_{PF}$: $H_0: d \leq 0$; $H_1: d > 0$

$\mu = 0$

$H_0: \hat{P}_{pop} - \hat{P}_{ft} \leq 0$ $H_1: \hat{P}_{pop} - \hat{P}_{ft} > 0$



Cannot Reject Ho Coef = 1%



$P_{pop} \leq P_{pf}$

Conclusion: $\hat{P}_{pop} \leq \hat{P}_{Ft}$ (Ho)

Confidence Interval

A 95% confidence interval for a difference in proportions $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

1. Can do a confidence interval or a T-score test.
2. With Viagra, $\hat{p}_1 = 77/6000000 = 0.00001283$ and without Viagra $\hat{p}_2 = 11/1500000 = 0.00000733$.
3. Need to test whether these are equal.

Confidence Interval

With a 95% confidence interval for $(p_1 - p_2)$ you get an interval

$$(1.06 \times 10^{-5}, 3.04 \times 10^{-7})$$

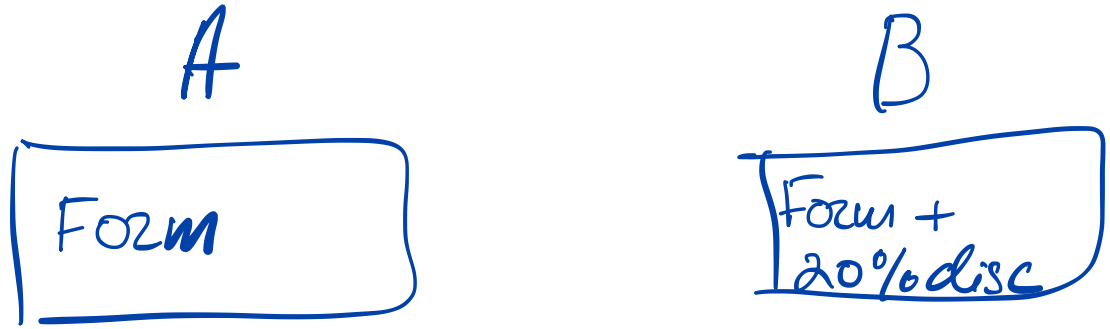
This doesn't contain zero. The evidence is that the proportion is higher.

1. Measured very accurately as n is large even though p is small.
2. With testing might use a one-sided test and an α of 0.01.

Difference of proportions:

```
> prop.test(x=c(11, 77), n=c(1500000, 6000000),
correct = F, alternative = "greater")
```

Electronic Arts



EA SimCity 5, one of EA's most popular video games, sold 1.1 million copies in the first two weeks of its launch last year. 50% of digital sales due to a A/B testing strategy.

Promotion banner or not?

Surprising results: The variation with no offer messaging whatsoever drove 43.4% more purchases. A

Confounding variables?

R: abtest

Examine whether a black or pink background results in more purchases Run experiment for one week:

- Pink background: 40% purchase rate with 500 visitors
- Black background: 30% purchase rate with 550 visitors

`abtest` to see which is more effective
Purchase rate for the pink background is significantly higher

```
> abtestfunc(site1, site2)
[1] 37.2 42.8
[1] 27.5 32.5
```

Google

Google offers A/B Testing

Create an A/B test

Follow these steps to create a simple A/B test.

An A/B test is a randomized experiment using two or more variants of the same web page (A and B). Variant A is the original and variant B through n each contain at least one element that is modified from the original.

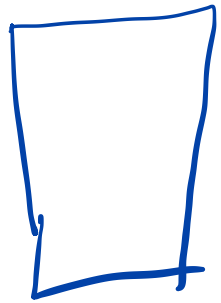
In this article:

Create a hypothesis
Create an A/B test
The variants card
The configuration card
Start your experiment
How long should your experiment run?
Experiment management options
Reports
Related resources

Figure: Steps to create A/B testing on Google

Netflix

Netflix



- Every product change Netflix considers goes through a rigorous A/B testing process before becoming the default user experience. Adaptive streaming and content delivery network algorithms.

> # It's All A/Bout Testing: The Netflix Experimentation Platform
>
> Netflix Technology Blog in Netflix TechBlog Follow
> Apr 29, 2016 · 11 min read
>
> Ever wonder how Netflix serves a great streaming experience with high-quality video and minimal playback interruptions? Thank the team of engineers and data scientists who constantly A/B test their innovations to our adaptive streaming and content delivery network algorithms. What about more obvious changes, such as the complete redesign of our UI layout or our new personalized homepage? Yes, all thoroughly A/B tested.

Discovering Argon

Nobel Prize

Lord Rayleigh won the Nobel Prize for discovering Argon.

This discovery occurred when he noticed a small discrepancy between two sets of measurements on nitrogen gas that he had extracted from the air and one he had made in the lab.

1. First, he removed all oxygen from a sample of air. He measured the density of the remaining gas in a fixed volume at constant temperature and pressure.
2. Second, he prepared the same volume of pure nitrogen by the chemical decomposition of nitrous oxide (N_2O) and nitric oxide NO.

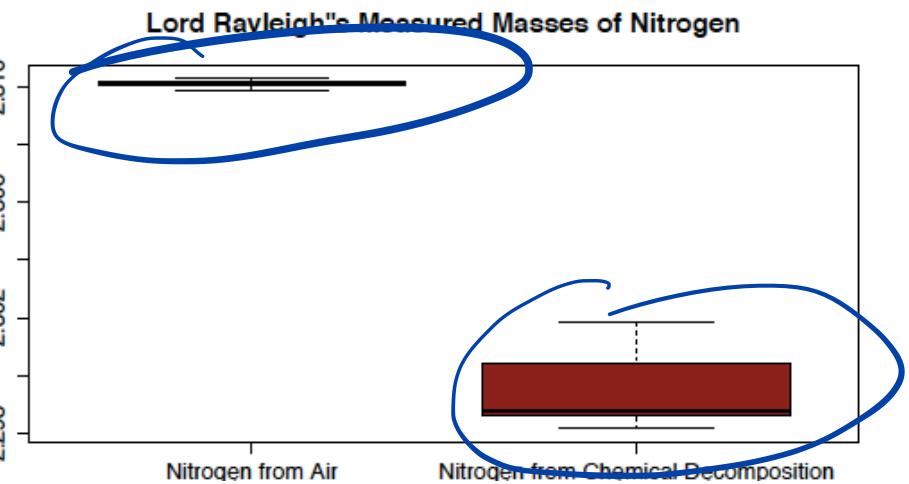
Here's the results

$Air = \cancel{O} + N + \varepsilon$

Data

Air N is heavier

Lord Rayleigh's data ...



From air	From chemical decomposition
2.31017	2.30143
2.30986	2.29890
2.31010	2.29816
2.31001	2.30182
2.31024	2.29869
2.31010	2.29940
2.31028	2.29849
—	2.29889
average	
2.31011	2.29947
st dev	
0.00014	0.00138

Discovering Argon in R

t.test(a,b,var.equal=T)

```
t = 20.2137, df = 13, p-value = 3.321e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.009499327 0.011772816
sample estimates:
mean of x mean of y
 2.310109  2.299473
```

- There's a $T = 20.21$! We've found Argon!

Cookie Cats


42
41
40
39
38
37
36
Area Locked
Collect 3 keys to unlock the gate
Play For Keys
Ask Friends For Keys
Unlock Gate Now
Gate Unlocked!
Welcome to
Suburbs
OK
Unlock Gate Now

Cookie Cats: AB Testing

		retention_1	retention_7	sum_gamerounds
A	gate_30	[0.444, 0.453]	[0.187, 0.194]	[50.076, 54.836]
B	gate_40	[0.438, 0.447]	[0.178, 0.186]	[50.350, 52.248]
	Observation	Overlap	No Overlap	Overlap
	Decision	No Difference	Difference	No Difference

Observational Studies vs Field Experiments

Booth ↗ ⇒ Bank
⇐ acct

Does owning a Tesla mean you have more money in the bank?

- I could go and buy a Tesla!

How do I test this? randomized experiment

- Randomly pick 100 and split into two groups 50/50
- Each person in the first group (treatment) gets a Tesla
- Each person in the second group (control) gets a Chevy
- In two years we compare average amount of money in each group

Science | Random
Business | Studies
Standard

How to you test for difference between Booth and Kellog?

Collect data: $H_0: Salary_B = Salary_K$

Randomized Controlled Trials: Nobel Prize 2019

UC 2020

On Oct 14, 2019 Duflo, Banerjee and Kremer got the prize for "for their experimental approach to alleviating global poverty."

- Experiment-based approach has transformed development economics
- Three ingredients: Right question + carefully designed experiment + statistical hypothesis testing
- Most of the traditional economic models are based on observational data
- Randomized experiments are the gold standard for understanding causality. Otherwise hard to distinguish correlation from causation.

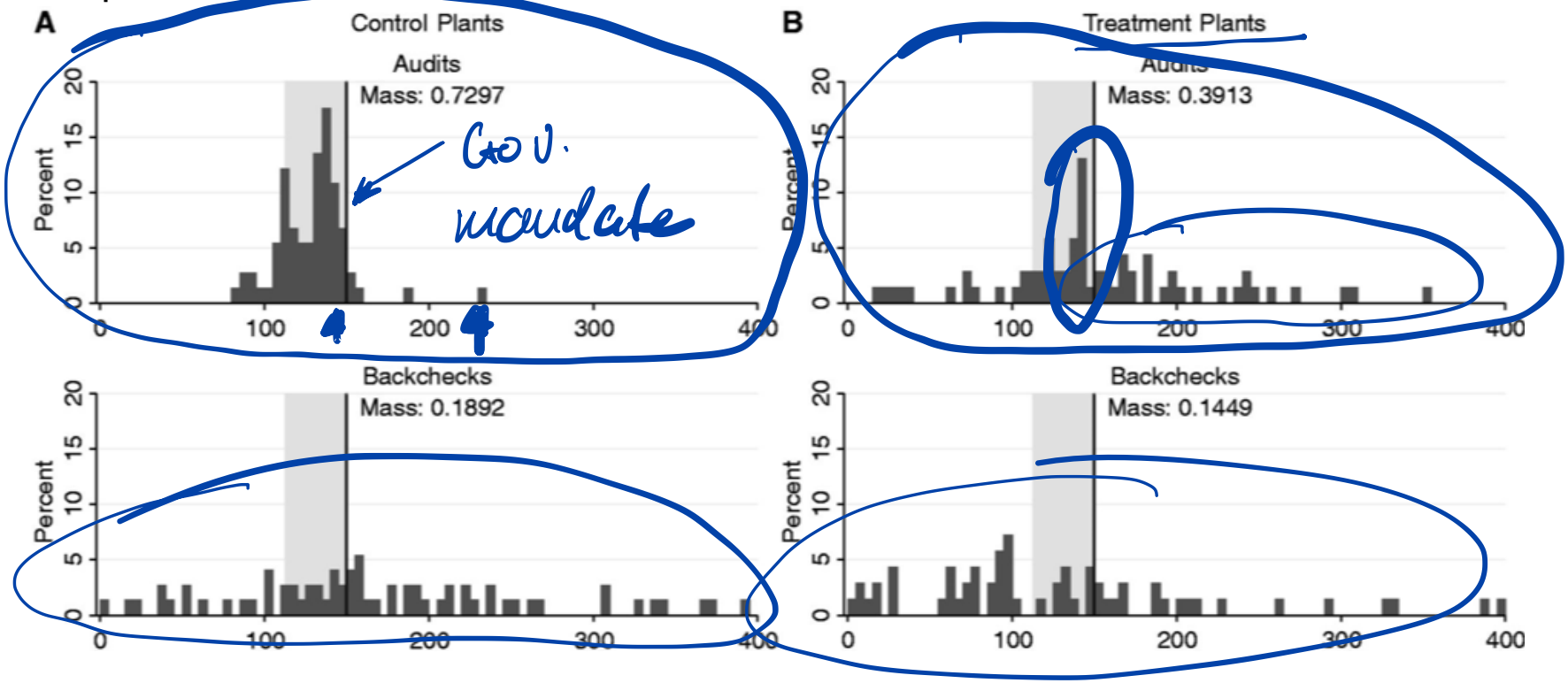
How do we collect data: Field Experiment

Field experiments (a.k.a randomized experiments) are standard in natural sciences (medicine, agriculture,...)

- They are standard in many business applications (AB testing)
- Now standard in policy analysis (the prize)
- What are the issues when we do not randomize?

Field Experiments: Pollution Reduction Policy in India

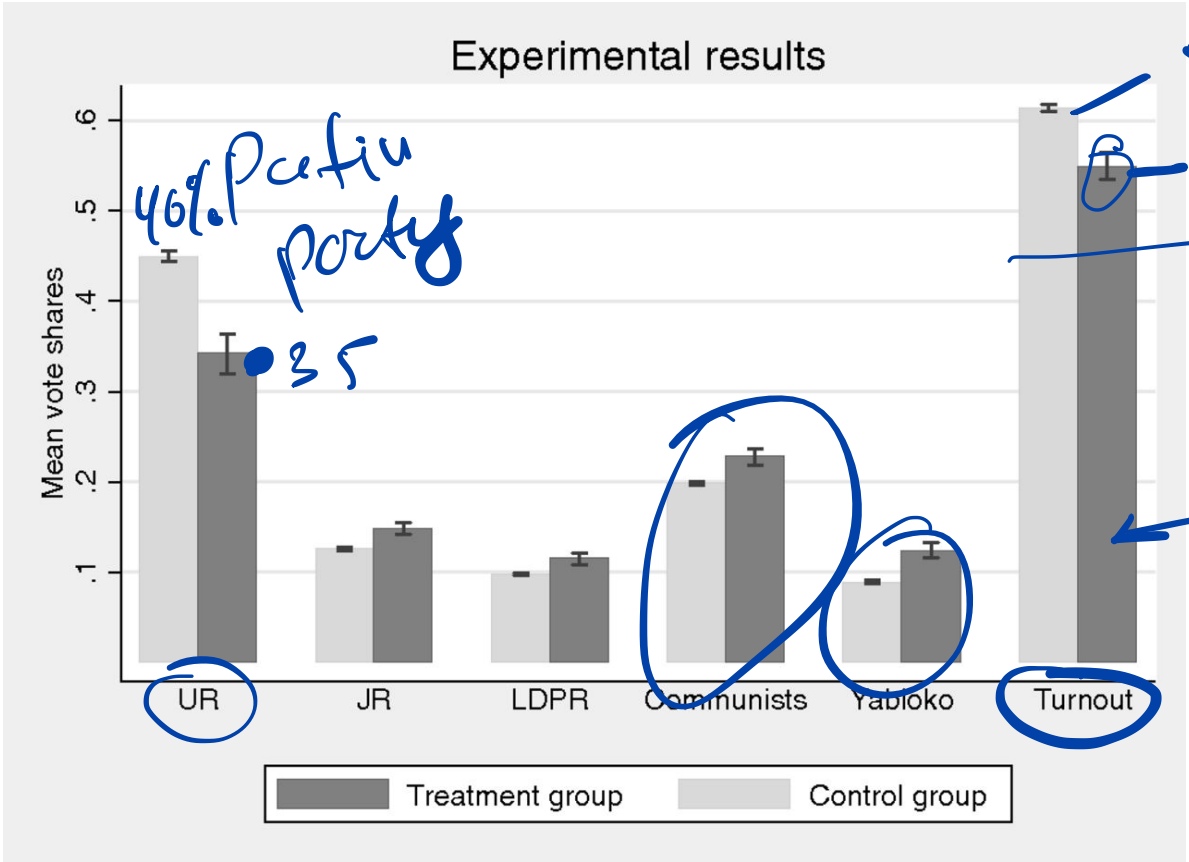
Suspended Particulate Matter concentrations



Source: Duflo 2013: Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India

Field Experiments: Russian Election Fraud

Independent observers were randomly assigned to 156 of 3,164 polling stations in the city of Moscow



Source: Field experiment estimate of electoral fraud in Russian parliamentary elections

Canibalization and Crossover

You test alternatives that lead to increased revenues, however, the total budget buyers is to spend is limited

- LinkedIn is testing match between ads and members want to improve ad impressions and revenue. What if advertisers always use 100% of ad budgets?
- Airbnb is testing new price suggestion algorithm. You see 10% improvement in control group. What happens when you implement new pricing everywhere?

B: Price suggested by Airbnb

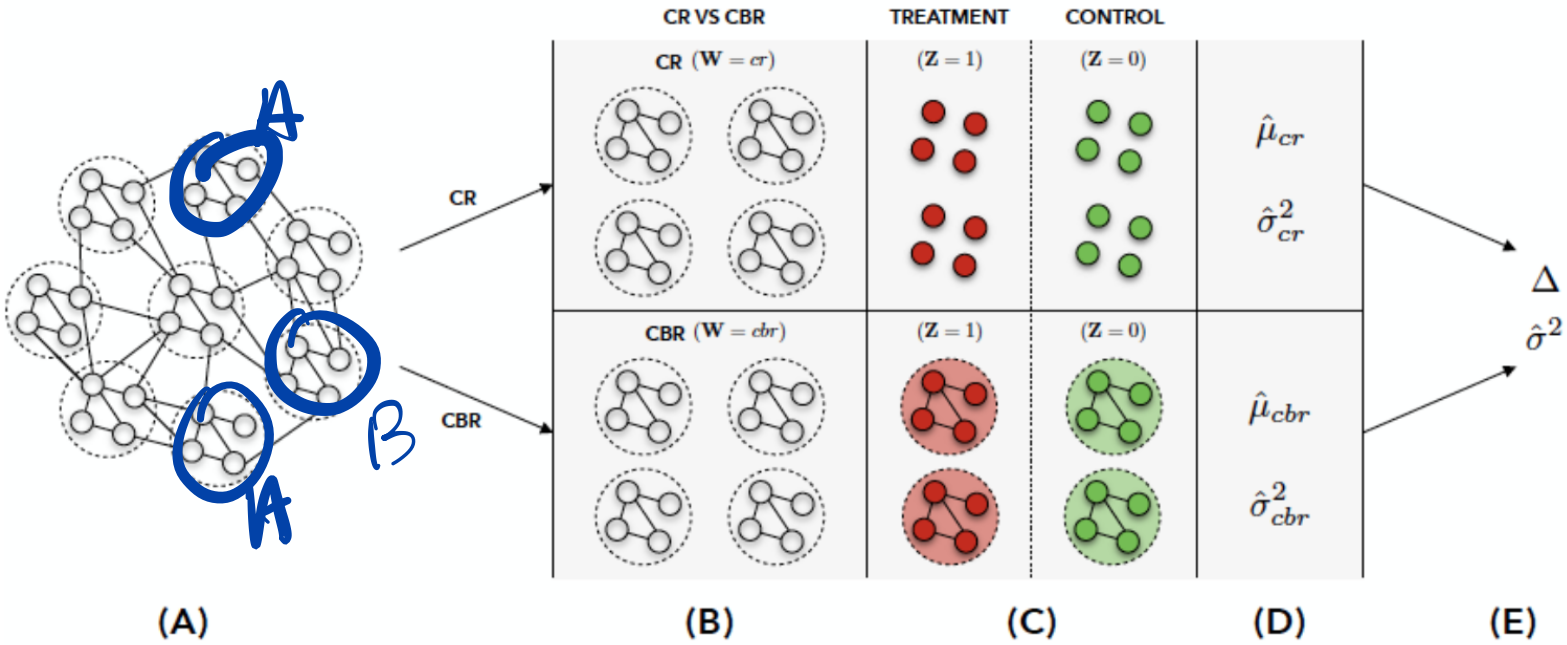
A: Price selected by owner

Biased rollouts

- When you recruit you control group, you will have a biased sample, not random!
- Twitter is testing 2.0 re-design and allowed users to opt in to the redesign
- You end up recruiting most engaged/motivated people

Interference

- What if treatment effects both control (A) and treatment (B) groups?
- Typically due to network effects
- Example A group of FB users are encouraged to write original content. Their friends in B group spend more time reading and have less time to write.
- David Puelz on how to analize interactions between groups



Source: LinedIn

Multiple Variants

When you test each k hypothesis separately, using some level of significance α, the probability of observing at least one significant by chance

$$1 - (1 - \alpha)^k$$

For $\alpha = 0.05$ and $k = 20$, you get 64% chance of observing at least one significant result, even if all of the tests are actually not significant.

- In 2009 Google could not decide which shades of blue for ad links
- They decided to test 41 shades
- At a 95% confidence level, the chance of getting a false positive was 88%.

If they had tested 10 shades, the chance of getting a false positive would have been 40

Coming Next: The Predictive Culture

Making a good prediction is arguably the most valuable application of machine learning and statistics in business problems.

- Instead of building a causal/interpretable/intuitive model the goal is to build a model that predicts well.
- This leads to very complex functions $f(x)$, e.g. Deep Learning!

Coming soon to economics!

Summary

- Estimating Parameters and Fitting Distributions
- Confidence and Prediction Intervals
- Means, Proportions, Differences
- A/B Testing