Bush 631-600: Quantitative Methods

Lecture 10 (11.08.2022): Uncertainty vol. I

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Fall 2022

What is today's plan?

- Calculating uncertainty: detecting 'real' findings.
- From r.v.s. to estimators.
- Types of estimators: data, surveys, experiments.
- Simulations.
- Confidence intervals
- R work: table(), loops, simulations, plots.

Final project

- Research proposal lessons:
  - Clear objective / research question.
  - Literature - what do we know?
  - Focus on concepts, not measures.

Final project

Data report:

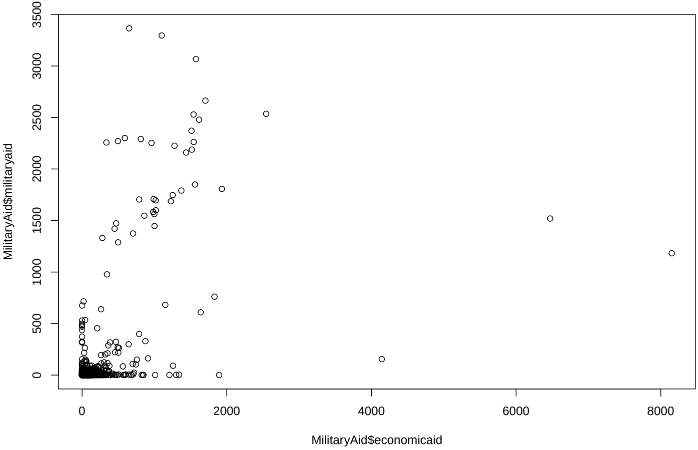
- Succinct description of topic and importance.
- What are my central arguments?
- Clear variable names.
- Variable values.

Visuals?

- Labels (axis, ticks).
- Title.
- Attention grabbing - use colors and add relevant text.

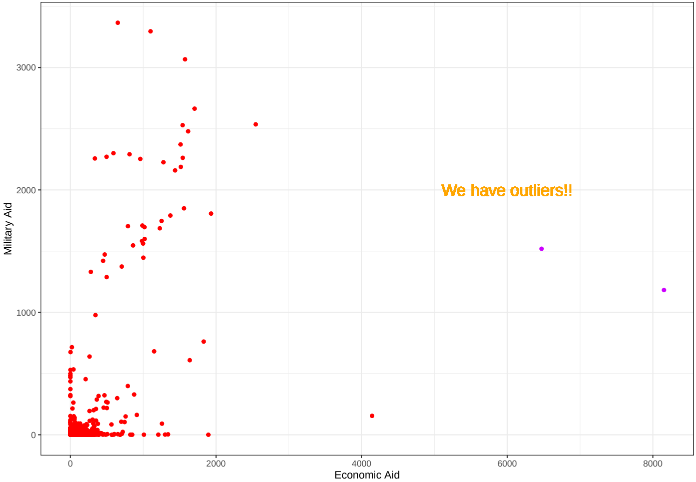
Useful visuals

What looks better?



Useful visuals

Or this...



We have findings!!!

- Data patters are systematic? Or noise?
- Our estimates $\rightarrow$ real relationship or random?
- Using probability calculations.

Events to numbers

- Random variables: map outcomes to numbers.
- Assess quantities in population $\rightarrow$ we cannot.
- Use sample: r.v.s and the values of concepts.
- Define a random variable X:
  - X=1 if 'random' person supports president, 0 otherwise.
  - $\bar{X} = \mathrm{E}[\bar{X}] = \mu$??
  - Yes!!
  - Large samples to the rescue.

Uncertainty on the diamond

2022 MLB Predictions

Updated after every game.

Standings Games Pitchers

Teams

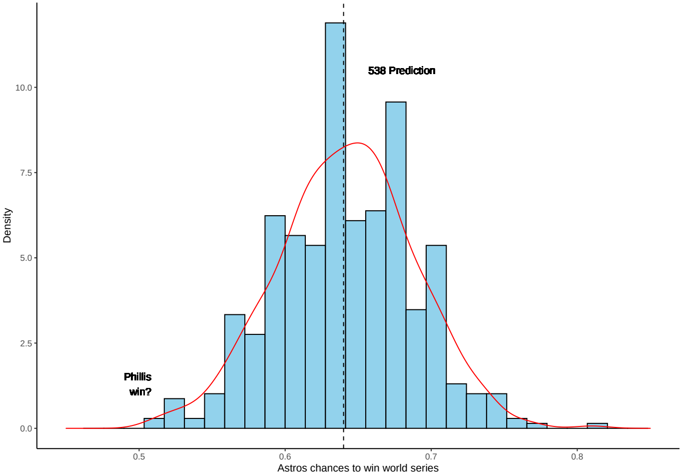
TEAM	DIVISION	TEAM RATING	1-WEEK CHANGE	CHANCE OF MAKING ... DIV. SERIES	LEAGUE CHAMP.	WORLD SERIES	WIN WORLD SERIES
Astros 3-0, 3-0	AL West	1588	+7	✓	✓	95%	64%
Phillies 3-1, 4-1	NL East	1547	+6	✓	✓	✓	33%

- No shot for the Phils?
- What does 64% mean?

Uncertainty on the diamond

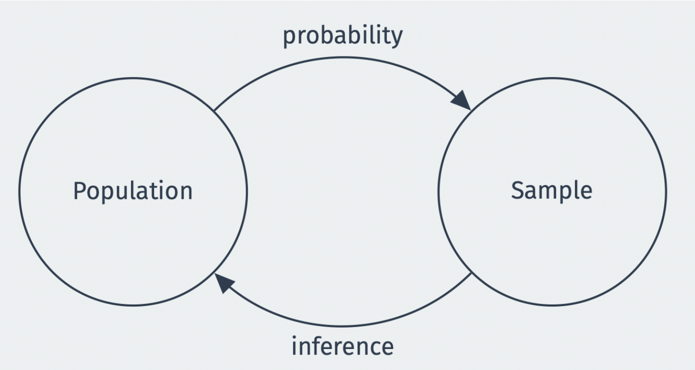
```
# Simulate the World series
sim <- 500
draws <- rbinom(sim, size = 100, prob = 0.64) / 100
mean(draws)
```

```
## [1] 0.64176
```



Our data - our research interests

- Making inferences from data to population



Uncertainty

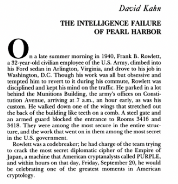
- Research questions:
  1. President's gender and FP actions?
  2. Regime type and frequency of terrorism?
  3. Regional trade zone and countries trade balance?
- Treatment / Factor has an effect:
  - Women are more aggressive in defense spending and public threats.
  - Democratic regimes experience more terror incidents.
  - Regional trade zone increased the trade balance with neighbors.
- Are these effects real or just noise?

Uncertainty in data: US and WW II

Pearl Harbor (December 7, 1941)

“Signals to noise ratio”
(Wohlstetter 1962)

Diplomatic .vs. military intelligence
(Kahn, 1991)

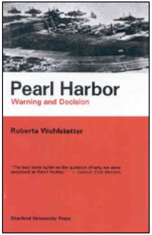


David Kahn

THE INTELLIGENCE FAILURE OF PEARL HARBOR

On a late summer morning in 1940, Frank B. Rowlett, a 32-year-old civilian employee of the U.S. Army, climbed into his Ford sedan in Arlington, Virginia, and drove to his job in Washington, D.C. Though his work was all but obsessive and tempted him to revert to it during his commute, Rowlett was disciplined and kept his mind on the traffic. He parked in a lot behind the Munitions Building, the army's offices on Constitution Avenue, arriving at 7 a.m., an hour early, as was his custom. He walked down one of the wings that stretched out the back of the building like teeth on a comb. A steel gate and an armed guard blocked the entrance to Rooms 3416 and 3418. They were among the most secure in the entire structure, and the work that went on in them among the most secret in the U.S. government.

Rowlett was a codebreaker; he had charge of the team trying to crack the most secret diplomatic cipher of the Empire of Japan, a machine that American cryptanalysts called PURPLE, and within hours on that day, Friday, September 20, he would be celebrating one of the greatest moments in American cryptology.

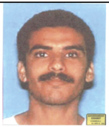


Uncertainty in data: 9/11 Intelligence failure

MEMORANDUM FOR:

FROM:

OFFICE:



SUBJECT: Re: Khalid Al-Mihdhar

REFERENCE:

Original Text of
Original Text of

TO:
FROM:
OFFICE:
DATE: *08/21/2001 04:05:02 PM*
SUBJECT: *Re: Khalid Al-Mihdhar*

WHAT?:?? Same passport number? How interesting. I know his fellow travelers made one or two trips to the US in the same January time frame, yes? Probably would be useufl to memorialize the US visits of the party in a cable.....

as I was reviewing all the cables on Khalid Al-Mihdhdar, I noticed he had a U.S. Visa in his passport. I asked INS to check and they just came back and said he entered the U.S. on 15 January 2000 and listed the Los Angeles as his destination. He departed the U.S. on 10 June 2000. I looked through traffic and could not find anything else.

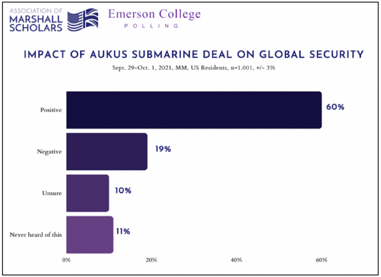
I'll be sending to FBI to pass what we know of Khalid Al-Mihdhar and that he entered the U.S. on 15 January. Maybe there is something they can do -- perhaps run his name by Ressam? I will be here in the morning, and will then be meeting with in the early afternoon to talk about the U.S.S. Cole and will give her a head's up. Let me know if you need me to do anything.

Estimation

- *Quantity of interest* in population.
- *Point estimation* $\rightarrow$ a ‘best guess’.
- Many possible point estimators:
  - Population mean (μ): elections turnout.
  - ‘Special population’ mean (μ): likelihood of joining international treaty.
  - Variance of a r.v. (σ^2): variation in support for sanctioning China/Russia.
  - Population ATE ($\mu_1 - \mu_0$): difference b-w treatment and control groups.

Estimation

Estimator θ



How to estimate public opinion?


MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV WATCHLIST CRAMER PRO

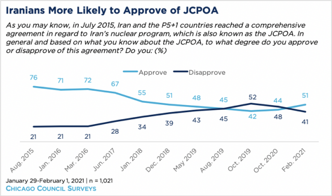
POLITICS

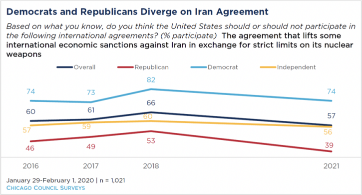
Iran will return to nuclear talks after months of stalled negotiations, State Department says

PUBLISHED WED, NOV 3 2021·4:29 PM EDT

Amanda Macias
@AMANDA_M_MACIAS

SHARE





- Random sample of respondents.

Estimating with public samples

- Assume: $X_1 ... X_n$ iid *Bernoulli distributed* random variables.
- Proportion of support for deal $\rightarrow$ p.
- An estimate: one realization of estimator (random variables right??)
- Estimate:

$$\hat{\theta} = \bar{X}_n \rightarrow \text{population p}$$

Still, an estimation...

- Is our estimate good?
- **Estimation error**: difference with 'true value'.
- Error $= \bar{X}_n - p$
- p is unknown, now what?
- Calculate average magnitude of estimation error.
- Hypothetical repetition of sampling:
  - Multiple estimate values ($\hat{\theta}$)
  - Multiple estimation error values.

Estimation

- Repetition $\rightarrow$ sampling distribution of $\hat{\theta}$
- Estimation error / bias using *expectations*
- $bias = E(est.Error) = E(Estimate - truth) = E(\bar{X}_n) - p = p - p = 0$

- **Unbiasedness**: Sample proportion is on average equal to the population proportion.
- Accuracy over multiple samples (not a single-shot survey)
- Estimator is *unbiased*

Estimators in experiments

- Treatment(s) and control groups.
- *Estimator* $\rightarrow$ diff-in-means.
- Sample Average Treatment Effect (SATE):

$$SATE = \frac{1}{n} * \sum_{i=1}^{n} [Y_i(1) - Y_i(0)]$$

Diff-in-means estimator

- Random sampling of population.
- Random assignment into treatment(s).
- Population Average Treatment Effect (PATE)
- $PATE = E[Y(1) - Y(0)]$
- Diff-in-means estimator is *unbiased*

Unbiased estimator

- Monte-Carlo simulations

```
# Create Sample, Control and treatment groups (means and SDs)
n <- 500
mu0 <- 0
sd0 <- 1
mu1 <- 1
sd1 <- 1

# Create sampling distributions
y0 <- rnorm(n, mean = mu0, sd = sd0)
head(y0)
```

```
## [1] -0.99274116 -1.32054615 -2.79817571  0.48616276 -0.03684791  1.41178990
```

```
y1 <- rnorm(n, mean = mu1, sd = sd1)
```

```
# calculate diff-in-means (SATE)
tau <- y1 - y0
head(tau)
```

```
## [1]  2.0929783  4.1806412  2.9382569  0.6697544  1.5358116 -0.2688237
```

```
SATE <- mean(tau)
SATE
```

```
## [1] 1.014856
```

Increasing the sample

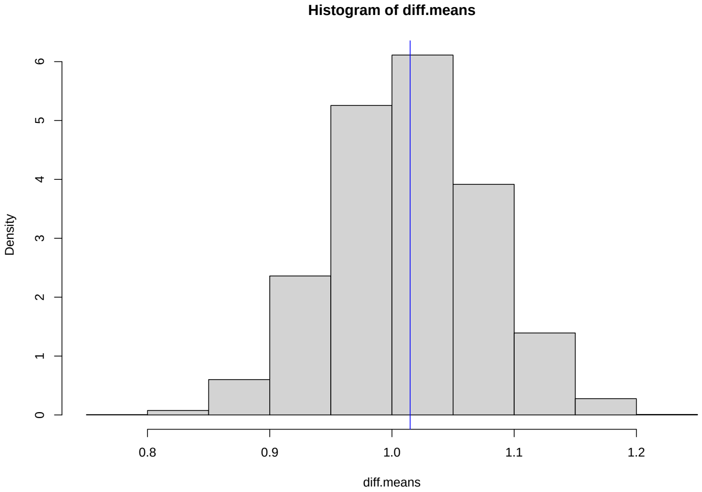
- Simulate & randomly assign treatment

```
# Repeat
sims <- 5000
diff.means <- rep(NA,sims)

for (i in 1:sims){
  treat <- sample(c(rep(1, n/2), rep(0, n/2)), size = n, replace = FALSE)
  diff.means[i] <- mean(y1[treat == 1] - mean(y0[treat == 0]))
}

est.error <- diff.means - SATE
summary(est.error)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.2459536 -0.0418603 -0.0015130 -0.0008485  0.0408982  0.2020634
```

SATE estimator (large sample simulation)

```
hist(diff.means, freq = FALSE)
abline(v=SATE, col = "blue")
```



Estimator distribution

- Calculate variation with SD (estimator)

```
# SD of estimator
sd(diff.means)
```

```
## [1] 0.06236986
```

```
sqrt(mean((diff.means - SATE)^2))
```

```
## [1] 0.0623694
```

- Calculate SD - only with a simulation.
- Reality $\rightarrow$ one sample, SD is unknown.

SD of sample

- **Standard error**: estimated degree of deviation from expected value
- Variability of our (single!) sample
- $\sqrt{\hat{V}(\hat{Y})} = \sqrt{\frac{\bar{Y}*(1-\bar{Y})}{n}}$

```
# Simulate and add SE calculate
sims2 <- 5000
diff.means2 <- rep(NA,sims)
diff.se <- rep(NA, sims)

for (i in 1:sims){
  Y0 <- rnorm(n, mean = mu0, sd = sd0)
  Y1 <- rnorm(n, mean = mu1, sd = sd1)
  treat <- sample(c(rep(1, n/2), rep(0, n/2)), size = n, replace = FALSE)
  diff.means2[i] <- mean(Y1[treat == 1] - mean(Y0[treat == 0]))
  diff.se[i] <- sqrt(var(Y1[treat == 1])/(n/2) + var(Y0[treat == 0])/(n/2))
}

sd(diff.means2)
```

```
## [1] 0.08911718
```

```
mean(diff.se)
```

```
## [1] 0.08936297
```

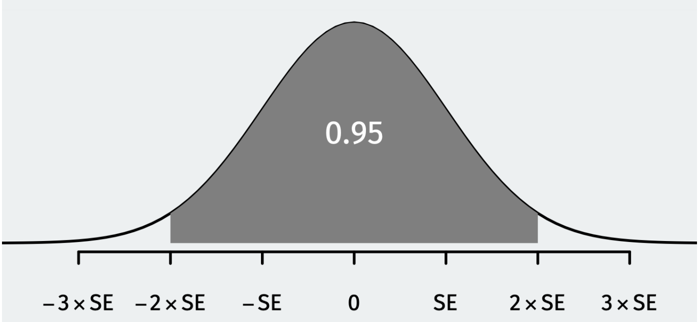
Broader approach to estimator distribution

- Quantities beyond means and SD.

Confidence Intervals

- Range of true values of estimator.
- Range of plausible values.
- Rest on assuming repeated sampling.

Chance errors intervals



- Normal distribution empirical rule.
- 95% of values within 2 SD, in sample $\rightarrow$ 2 SEs.
- Range of possible values $\rightsquigarrow$ $\pm$ 1.96 SEs

BYO CIs

- Constructing confidence intervals.
- (1) What *confidence level?*
- Conventional: 95%.
- Defined using $\alpha(0-1)=?$
- (2) CI: $100*(1-\alpha)\% = \bar{Y} \pm z_{\alpha/2} * SE$
- $\alpha = 0.05 \rightarrow 95\%$ CI.

Confidence Intervals

- Formal CI:

$$CI(\alpha) = (\bar{X}_n - z_{\alpha/2} * SE, \bar{X}_n + z_{\alpha/2} * SE)$$

- *Critical value* $= (1 - \alpha/2)$

α	*Confidence level*	*Critical value* $z_{\alpha/2}$	*R expression*
0.01	99%	2.58	`qnorm(0.995)`
0.05	95%	1.96	`qnorm(0.975)`
0.1	90%	1.64	`qnorm(0.95)`

Confidence intervals

- Finding the critical values
- qnorm() function: define lower.tail = FALSE

```
# find critical values
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.644854
```

```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.959964
```

```
qnorm(0.005, lower.tail = FALSE)
```

```
## [1] 2.575829
```

CIs in R

- CIs for our JCPOA survey

```
# Sample, Mean support and SE
n <- 2000
x.bar <- 0.6
Iran.se <- sqrt(x.bar * (1-x.bar)/n)
Iran.se
```

```
## [1] 0.01095445
```

```
# CIs
c(x.bar - qnorm(0.995) * Iran.se, x.bar + qnorm(0.995) * Iran.se)   #99%
```

```
## [1] 0.5717832 0.6282168
```

```
c(x.bar - qnorm(0.975) * Iran.se, x.bar + qnorm(0.975) * Iran.se)   #95%
```

```
## [1] 0.5785297 0.6214703
```

```
c(x.bar - qnorm(0.95) * Iran.se, x.bar + qnorm(0.95) * Iran.se)   #90%
```

```
## [1] 0.5819815 0.6180185
```

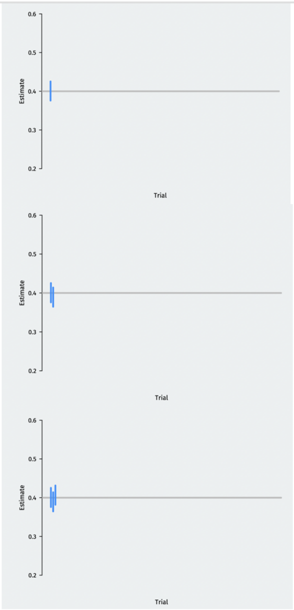
Interpretation

- How to interpret CIs?
- NO $\rightarrow$ 95% chance true value is within the interval.
- Why? Estimator is unknown (value is 0/1).
- YES $\rightarrow$ Interval contains true value 95% of the times in repeated random samples.
- Not the **Wait What?** pic again right???
- **One more time**:

Interval contains the true value 95% of the times in repeated random samples

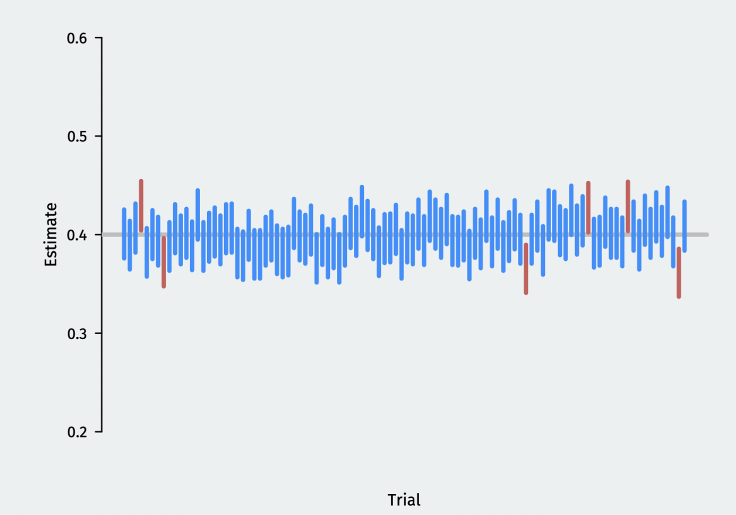
Simulate CIs

- Policy: Global Co_2 emissions reduction
- Sample = 1500 respondents
- p = 0.4 (assumed support)
- Calculate 95% CIs in multiple samples



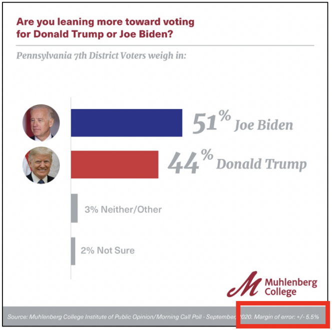
Simulate CIs

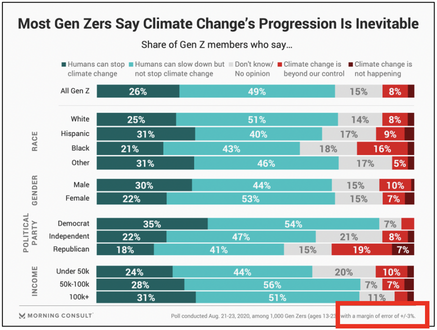
- How many overlap with ‘true’ support?



Polls: the 'fine print'

Margin of error





Margin of error

- MOE: half-width of a 95% CI.
- JCPOA sample proportion of support $= 0.6$
- JCPOA sample MOE $= \pm 3$ %
- JCPOA 95% CI: [57%,63%]

Margin of error

$$MOE = \pm z_{0.025} * SE \approx \pm 1.96 * \sqrt{\frac{\bar{X}_n * (1 - \bar{X}_n)}{n}}$$

- What is the minimum sample size?
- Popular stage in research design.
- Conduct **before** fielding the survey.

MOE and Sample size

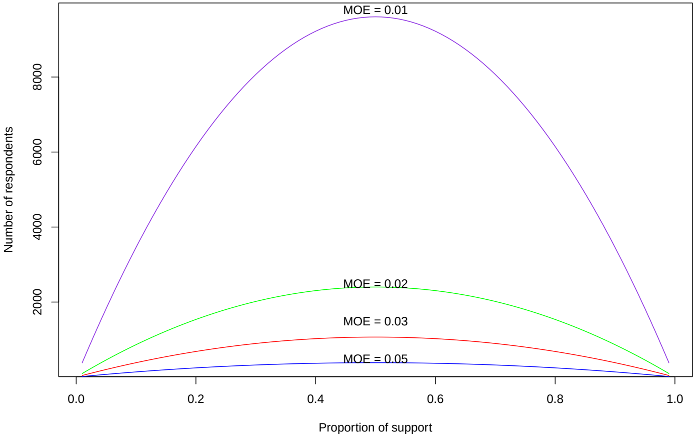
- Calculate multiple proportions of support.
- Define your MOE: 1% , 2%, 3%, 5%
- Possible sample sizes

```
# Define MOEs
moe <- c(0.01, 0.02, 0.03, 0.05)

# Define vector of proportion of support (0-100 by 1%)
prop <- seq(from = 0.01, to = 0.99, by = 0.01)

# Using MOE and proportion for possible sample sizes
num <- 1.96^2 * prop * (1-prop) / moe[1]^2
head(num, n=10)
```

```
##  [1]  380.3184  752.9536 1117.9056 1475.1744 1824.7600 2166.6624 2500.8816
##  [8] 2827.4176 3146.2704 3457.4400
```

MOE and Sample size

- Plotting our analysis
- CLT, SE and sample size...



CIs & Experiments

- Quantify uncertainty for causal effect analysis.
- JCPOA support among Americans $\rightarrow$ good!
- Variations of JCPOA support among groups $\rightarrow$ even better!
  - Men $\longleftrightarrow$ Women.
  - Young $\longleftrightarrow$ Old.
  - Vets (military) $\longleftrightarrow$ no military background.
- Estimator: *population* ATE
- $\mu_T - \mu_C$

Let's reduce plastics

- Environmental policy: 'fighting-back' against plastic bags.
- Policy, main aspects - financial incentives:
  1. Financial incentives: cash back.
  2. Financial incentives: fee for plastic bags.
- Define outcome: $X_i = 1$ if support policy, 0 otherwise.
- Sample mean (treatment: cash back), $\bar{X_T} = 0.43$
- Sample mean (control: plastics fee), $\bar{X_C} = 0.32$

$$A\hat{T}E = \hat{X_T} - \hat{X_C} = 0.11$$

Simulating policy support

- Sample diff-in-means on average equal to population diff-in-means
- Still, some variation

```
# Simulate our experiment in population
xt.sims <- rbinom(1000, size = 1000, prob = 0.43) / 1000
head(xt.sims)
```

```
## [1] 0.423 0.438 0.414 0.459 0.428 0.429
```

```
xc.sims <- rbinom(1000, size = 1000, prob = 0.32) / 1000
head(xc.sims)
```

```
## [1] 0.326 0.321 0.314 0.304 0.309 0.324
```

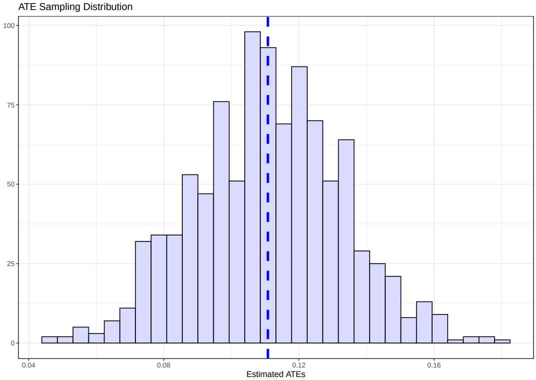
```
# Mean
mean(xt.sims-xc.sims)
```

```
## [1] 0.110806
```

ATE distribution

- How our $\hat{ATE} \approx 0.11$ looks like?

```
# Plot with tidyverse
hp <- data.frame(mn = (xt.sims-xc.sims))
ggplot(hp, aes(mn)) +
  geom_histogram(fill="#D6D7FF", color="black", alpha=0.9) +
  geom_vline(xintercept = mean(hp$mn), color = "blue", linetype = "dashed", size = 1.5) +
  xlab("Estimated ATEs") + ylab("") + ggtitle("ATE Sampling Distribution") +
  theme_bw()
```



Simulating policy support

- $\hat{ATE} \approx 0.11 \rightarrow$ makes a difference?
- Use SEs to learn of variation of estimator

```
# Calculate SE
x.se <- sqrt((0.43*0.57)/1000 + (0.32*0.68)/1100)
x.se
```

```
## [1] 0.02104562
```

```
# 95% CIs for meaningful results
c(0.43 - qnorm(0.975) * x.se, 0.43 + qnorm(0.975) * x.se)
```

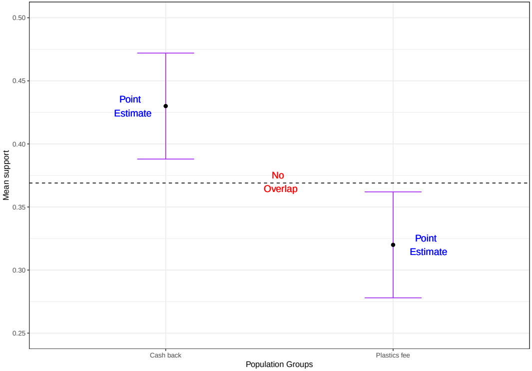
```
## [1] 0.3887513 0.4712487
```

```
c(0.32 - qnorm(0.975) * x.se, 0.32 + qnorm(0.975) * x.se)
```

```
## [1] 0.2787513 0.3612487
```

Plot and check effect

```r
# plot with tidyverse
ggplot(se_plot, aes(x,y)) +
  geom_errorbar(aes(ymin = y-2*se, ymax = y+2*se), width = 0.25, color = "purple") +
  geom_point(size = 2) + ylim(0.25,0.5) +
  geom_hline(yintercept = 0.369, linetype = "dashed") +
  geom_text(x=1.5,y=0.37,label = "No \n Overlap", color = "red", size = 4.5) +
  geom_text(x=0.85,y=0.43,label = "Point \n Estimate", color = "blue", size = 4.5) +
  geom_text(x=2.15,y=0.32,label = "Point \n Estimate", color = "blue", size = 4.5) +
  ylab("Mean support") + xlab("Population Groups") +
  theme_bw()
```



More simulations and data

- Create our own experimental data
- library(fabricatr): Random data generator
- Steps:
  1. Create treatments (assign sample size and probabilities).
  2. Create binary outcome variables.
  3. Create continuous outcome variables.
- Join all variables into one large data set.
- Focus on treatment 1 and cont. outcome variable:
  - Regime of aid recipient (democracy or not).
  - Extent of aid provided.

Create random data

- Code for treatments and all variables

```r
## Create data
# Set seed for randomizer
set.seed(12345)

# Create treatments (sample size of 1000)
exp.dat <- fabricate(
  N = 1000,
  trt1 = draw_binary(N = 1000, prob = 0.5),
  trt2 = draw_binary(N = 1000, prob = 0.5))

# Create Binary & Continuous outcome variables
random_vars <- fabricate(
  N = 1000,
  dv_cor1 = correlate(given = exp.dat$trt1, rho = 0.8,
                      draw_binary, N = 1000, prob = 0.65),
  dv_cor2 = correlate(given = exp.dat$trt2, rho = 0.65,
                      draw_binary, N = 1000, prob = 0.35),
  cont_cor1 = correlate(given = exp.dat$trt1, rho = 0.55,
                        rnorm, mean = 1500, sd = 30),
  cont_cor2 = correlate(given = exp.dat$trt2, rho = 0.75,
                        rnorm, mean = 1450, sd = 45))
```

Create random data

- Join variables and final data output

```r
# Tidyverse approach to join columns
exp.dat <- left_join(exp.dat, random_vars, by = "ID")

# Our random experimental data
head(exp.dat, n=8)
```

```
##     ID trt1 trt2 dv_cor1 dv_cor2 cont_cor1 cont_cor2
## 1 0001    1    0       1       0  1523.100  1395.533
## 2 0002    1    1       1       1  1492.402  1466.578
## 3 0003    1    0       1       0  1500.165  1431.904
## 4 0004    1    0       1       0  1510.011  1406.666
## 5 0005    0    1       0       1  1515.649  1442.158
## 6 0006    0    0       0       0  1512.053  1430.640
## 7 0007    0    0       0       1  1474.265  1451.380
## 8 0008    1    1       1       0  1498.759  1443.719
```

Exploring the experimental data

- Random assignment of 'respondents'?
- Calculate mean outcome for treatment 1 and ATE.

```
# How many 'respondents' assigned per treatment?
n.zero <- sum(exp.dat$trt1 == 0)
n.zero
```

```
## [1] 469
```

```
n.one <- sum(exp.dat$trt1 == 1)
n.one
```

```
## [1] 531
```

```
# Mean outcome variable by treatment 1
est.zero <- mean(exp.dat$cont_cor1[exp.dat$trt1 == 0])
est.zero
```

```
## [1] 1489.333
```

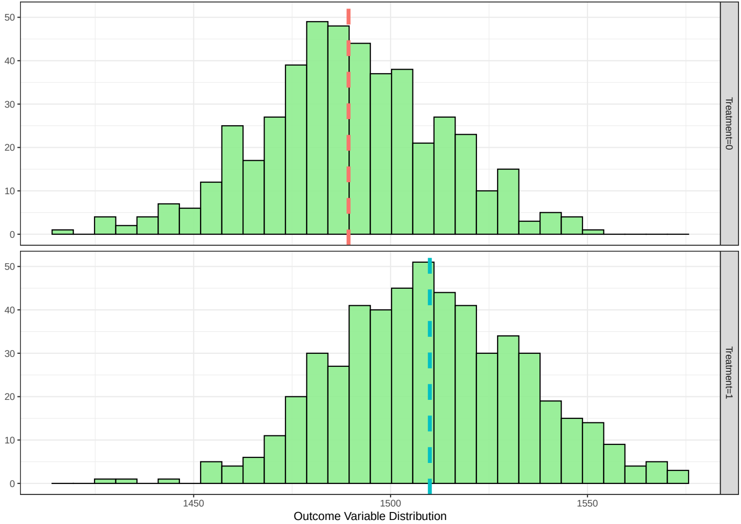
```
est.one <- mean(exp.dat$cont_cor1[exp.dat$trt1 == 1])
est.one
```

```
## [1] 1509.973
```

```
# calculate ATE (Y(1) - Y(0))
est.one - est.zero
```

```
## [1] 20.6396
```

How does it look?



Regime treatment matters?

- Calculate margin of error $\rightarrow$ SEs
- Calculate CIs (define $\alpha = 0.05$)

```
# SEs for treatment 1 results
se.zero <- sd(exp.dat$cont_cor1[exp.dat$trt1 == 0]) / sqrt(n.zero)
se.zero
```

```
## [1] 1.068058
```

```
se.one <- sd(exp.dat$cont_cor1[exp.dat$trt1 == 1]) / sqrt(n.one)
se.one
```

```
## [1] 1.06291
```

```
# Define alpha
alpha <- 0.05

# CIs
ci.zero <- c(est.zero - qnorm(1-alpha / 2) *
               se.zero, est.zero + qnorm(1-alpha / 2) * se.zero)
ci.zero
```

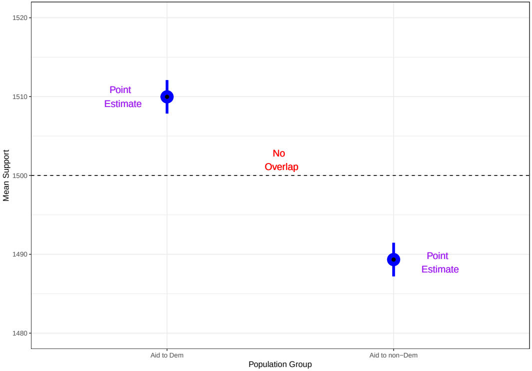
```
## [1] 1487.240 1491.427
```

```
ci.one <- c(est.one - qnorm(1-alpha / 2) *
              se.one, est.one + qnorm(1-alpha / 2) * se.one)
ci.one
```

```
## [1] 1507.890 1512.056
```

How does our effect looks? matters?

```
# plot with tidyverse
ggplot(se_plot2, aes(x,y)) +
  geom_pointrange(aes(ymin = y-2*se, ymax = y+2*se), color = "blue", size = 1.75) +
  geom_point(size = 2) + ylim(1480,1520) +
  geom_hline(yintercept = 1500, linetype = "dashed") +
  geom_text(x=1.5,y=1502,label = "No \n Overlap", color = "red", size = 4.5) +
  geom_text(x=0.8,y=1510,label = "Point \n Estimate", color = "purple", size = 4.5) +
  geom_text(x=2.2,y=1489,label = "Point \n Estimate", color = "purple", size = 4.5) +
  ylab("Mean Support") + xlab("Population Group") +
  theme_bw()
```



Clarifying objectives

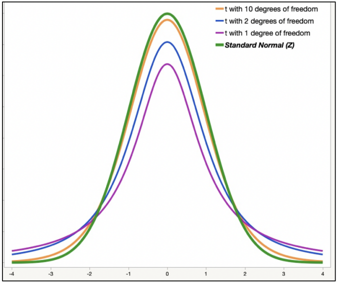
- What's with all the simulations?
- Real world: 1 sample, 1 mean...
- Research supported by simulations: public policy
  - Support for government policy: expand anecdotal findings.
  - Lobbying in the senate: women representatives example.
- Research supported by simulations: business world
  - Product design and development: expand A/B testing.

Estimation approaches

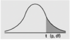
- Estimation thus far $\rightarrow$ CLT
- ATE & CIs are based on CLT assumption
- Alternative: outcome variable $\sim$ N (μ, σ^2)
- Use *student's t-distribution*:
  - Also describes DOF (degrees of freedom).
  - normal z-score $==$ student's t-statistic.
  - Distribution has 'heavier tails'.

student's t-distribution

- $DOF = (n - k)$, (n= observations; k=model parameters).
- Critical value: t-statistic



Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	——	——	80%	90%	95%	98%	99%	99.9%

t-distribution in R

- CIs ar wider, more conservative
- Use qt() function

```
# CI: CLT vs. t-distribution
# Treatment = 0
ci.zero
```

```
## [1] 1487.240 1491.427
```

```
ci.zeroT <- c(est.zero - qt(0.975,df = n.zero - 1) * se.zero,
              est.zero + qt(0.975,df = n.zero - 1) * se.zero)

ci.zeroT
```

```
## [1] 1487.234 1491.432
```

```
# Treatment = 1
ci.one
```

```
## [1] 1507.890 1512.056
```

```
ci.oneT <- c(est.one - qt(0.975,df = n.one - 1) * se.one,
             est.one + qt(0.975,df = n.one - 1) * se.one)
ci.oneT
```

```
## [1] 1507.885 1512.061
```

Wrapping up Week 10

- Summary:
  - The challenge of uncertainty: Separating signals and noise.
  - Estimation using sample mean or diff-in-means.
  - Simulations and estimators probability distributions.
  - SD, SEs and margin of errors.
  - Constructing CI - how to interpret 95% CI?
  - Estimators are uncertain, but meaningful?
  - Estimating with the t-distribution.

Data report next week!