Week 3: Learning from Random Samples

Brandon Stewart[1]

Princeton

September 14-18, 2020

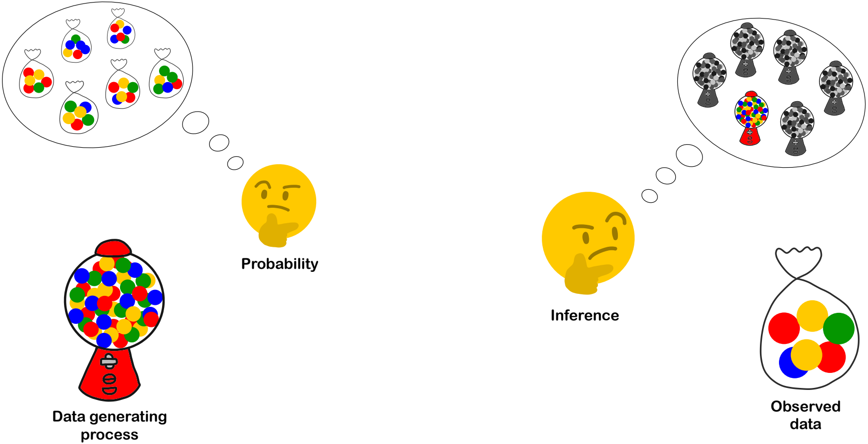
[1] These slides are heavily influenced by Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller, Erin Hartman and Matt Salganik. Some illustrations by Shay O'Brien.

Where We've Been and Where We're Going...

- Last Week
  - random variables
  - joint distributions
- This Week
  - estimators and sampling distributions
  - estimator properties (bias, variance, consistency)
  - confidence intervals
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability → inference → regression → causal inference

1. Estimation
    - Populations and Samples
    - Estimators
    - Analytical
2. Weak Law of Large Numbers and the Central Limit Theorem
    - Chebychev's Inequality
    - Weak Law of Large Numbers
    - The Central Limit Theorem
3. Properties of Estimators
    - Four Desirable Properties
    - Example
4. Interval Estimation
    - Intervals
    - Large Sample Intervals for a Mean
    - Small Sample Intervals for a Mean
    - Comparing Two Groups
    - Interval Estimation for a Proportion
5. Plug-In Principle

Where We've Been and Where We're Going...



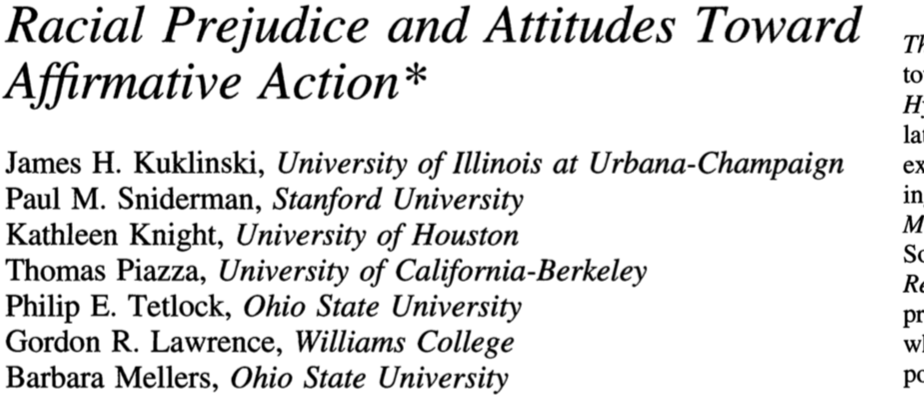
Probability

Data generating process

Inference

Observed data

Primary Goal for This Week



*Racial Prejudice and Attitudes Toward Affirmative Action**

James H. Kuklinski, *University of Illinois at Urbana-Champaign*
Paul M. Sniderman, *Stanford University*
Kathleen Knight, *University of Houston*
Thomas Piazza, *University of California-Berkeley*
Philip E. Tetlock, *Ohio State University*
Gordon R. Lawrence, *Williams College*
Barbara Mellers, *Ohio State University*

Th
to
Hy
la
ex
in
M
So
Re
pr
wl
po

https://www.jstor.org/stable/2111770

Primary Goal for This Week

We want to be able to interpret the numbers in this table (and a couple of numbers that can be derived from these numbers).

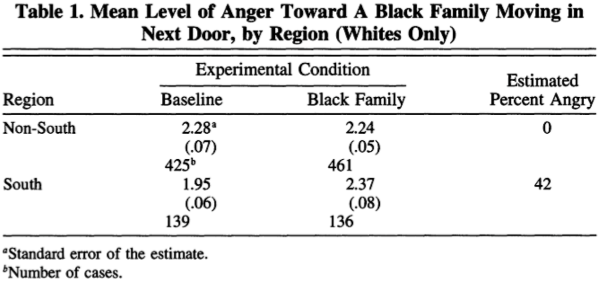


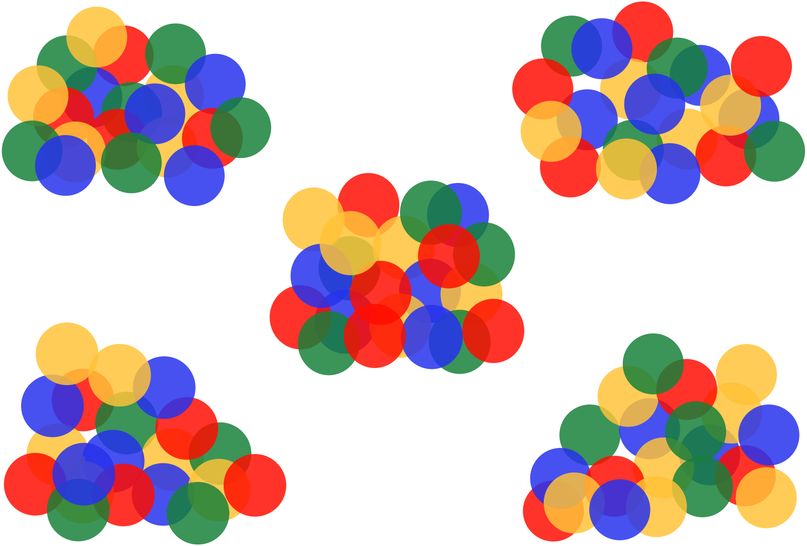
Table 1. Mean Level of Anger Toward A Black Family Moving in Next Door, by Region (Whites Only)

	Experimental Condition		
Region	Baseline	Black Family	Estimated Percent Angry
Non-South	2.28[a] (.07) 425[b]	2.24 (.05) 461	0
South	1.95 (.06) 139	2.37 (.08) 136	42

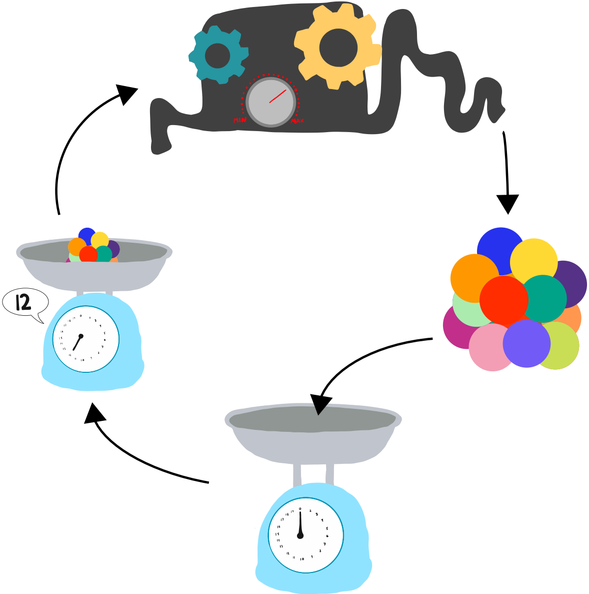
[a] Standard error of the estimate.

[b] Number of cases.

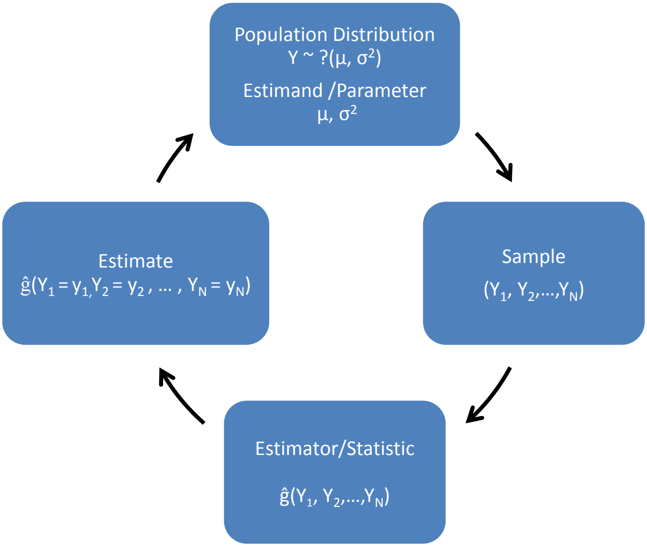
An Overview



An Overview



An Overview



Population Distribution
$Y \sim ?(\mu, \sigma^2)$

Estimand /Parameter
μ, σ^2

Sample
$(Y_1, Y_2, \ldots, Y_N)$

Estimator/Statistic
$\hat{g}(Y_1, Y_2, \ldots, Y_N)$

Estimate
$\hat{g}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N)$

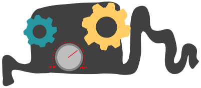
Populations

- Typically, we want to learn about the distribution of random variables for a population of interest.
- We will sometimes call the population distribution the data generating process and represent it with a pmf or pdf, $f_X(x; \theta)$.
- The population can be:
  - finite: as in all residents of a country
  - or infinite: as in all possible television ads.
- With either a finite or infinite population our main goal in inference is to learn about the population distribution f_X via summaries, like $E[X]$ or $V[X]$, which we call a population parameter (or just parameter).
- Ideally we assume as little as possible about the form of f_X.

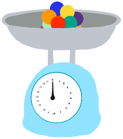
Nomenclature: Estimands, Estimators, and Estimates

The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by parameters.

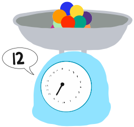
- Estimands are the parameters to estimate. Often written with greek letters (e.g. μ).



- Estimators are functions which map our data to guesses about the estimand. Often denoted with a "hat" (e.g. $\hat{\mu}$)



- Estimates are particular values of estimators that are realized in a given sample (e.g. 12)



Independent and Identically Distributed Samples

Statistical inference is learning about features of some population through a sampling mechanism.

- We will base most of our inferential machinery on the idea of random sampling.
- We will leverage the powerful assumption that we are observing IID—independent and identically distributed—samples of the random variable of interest.

Plain language:
Data are sampled IID when each observation is drawn from the same distribution, and the way an observation is drawn does not depend on the values of any other draw.

IID Formal Definition

Definition (Independent and Identically Distributed)

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be random variables with CDFs $F_1, F_2, \ldots, F_n$, respectively. Let F_A denote the joint CDF of the random variables with indices in the set A. Then $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are independently and identically distributed if they satisfy the following:

- Mutually independent: $\forall A \subseteq \{1, 2, \ldots, n\}, \forall (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n, F_A\left((x_i)_{i \in A}\right) = \prod_{i \in A} F_i(x_i)$
- Identically distributed: $\forall i, j \in \{1, 2, \ldots, n\}$ *and* $\forall x \in \mathbb{R}, F_i(x) = F_j(x)$

(Aronow and Miller Definition 3.1.1)

Sample Notation

- Under IID, we take a draw from a random variable X and then take another draw such that the outcome doesn't depend on the first.
- For a collection of samples of size n, we collect each of these units into a vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$.
- IID tells us that each one is produced under the same random process. This is how we get leverage to do estimation!
- We we will usually use unsubscripted capital letters, X, to refer to properties that all these draws share.
  e.g. $E[X] = E[X_1] = E[X_2] = \cdots = E[X_n]$

Independent and Identically Distributed

IID is an assumption that only approximates the truth.
What happens if it doesn't hold?

1) Observations may be dependent.
    - e.g. students in the same class who share a teacher.
    - when dependence is weak in the sense that we are still adding more information as we add more units then results generally carry through but likely to understate uncertainty.
2) Observations may not be identically distributed.
    - e.g. observations may be changing over time.
    - here the consequence is primarily for interpretability, limiting us to a pseudopopulation which aggregates over distinct distributions.

We want to avoid assumptions where we can to maintain credible inferences, but this is a relatively mild bedrock assumption.

We will return to these issues more in later videos and in future weeks.

Sampling from Finite Populations

- When we take a sample from a population it can be done with or without replacement.
- If the unit is replaced such that it can be sampled again, each draw is taken from the same population governed by the f_X.
- If the unit is not replaced, then each subsequent draw depends on the units previously sampled.
- If the population is very large relative to the sample, this won't turn out to matter much (because removing each unit doesn't change the overall distribution f_X).
- If the population is small relative to the sample size, it will be necessary to think carefully through the implications (see e.g. the challenge problem in problem set 3).

Sampling in R

```
## draw a sample of size 10 from our population
## drawn without replacement
my_sample <- dplyr::sample_n(my_data, size = 10,
                             replace = FALSE)
## this is a wrapper around sample.int()
my_sample <- my_data[sample.int(nrow(my_data),
                             size = 10, replace = FALSE), ]
```

What is An Estimator?

- An estimator $\hat{\theta}$ for some parameter θ, is a function of the sample $\hat{\theta} = h(Y_1, \ldots, Y_n)$.
- Because it is a function of the sample, the estimator is a random variable.
- We will study the properties of the estimator towards two goals:
  1. Inference: How much uncertainty do we have in this estimate?
  2. Evaluate Estimators: How do we choose which estimator to use?
- We study estimators by considering their behavior across an infinite number of hypothetical samples of size n that could be drawn. The resulting distribution of estimates is the sampling distribution.
- In real applications, we cannot draw repeated samples, so we approximate the sampling distribution.

The Sampling Distribution of the Sample Mean

Say we have the following population:

```
pop <- c(4, 2, 3, 6, 9, 2, 3, 6, 8, 5, 2, 9, 6, 3,
         4, 7, 6, 1, 2, 6, 9, 3, 1, 1, 1, 5, 7, 9)
```

We are going to take samples of size 10. How many possible samples are there?

```
choose(length(pop), 10)
```

```
## [1] 13123110
```

The Sampling Distribution

If we could draw each possible sample, we could calculate the sample mean in each one. This would form the full sampling distribution. We will simulate this by drawing 10,000 samples.

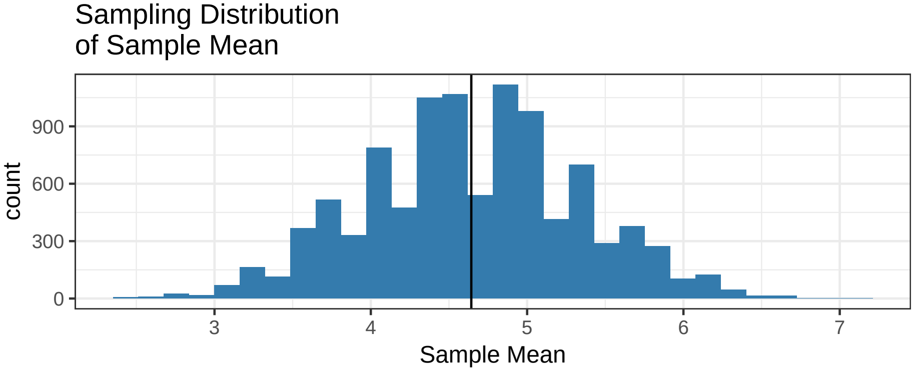
```
sim_res <- replicate(10000, {
  mean(pop[sample.int(length(pop), 10)])
}) %>% tibble(sample_mean = .) %>%
  rownames_to_column(var = "replicate")

sim_res[1:5, ]
```

```
## # A tibble: 5 x 2
##   replicate sample_mean
##       <chr>       <dbl>
## 1         1         5.4
## 2         2         4.9
## 3         3         3.7
## 4         4         3.6
## 5         5         5.3
```

The Sampling Distribution

And we can plot this sampling distribution

```
true_pop_mean = mean(pop)
ggplot(sim_res, aes(x = sample_mean)) +
  geom_histogram(fill = blue) +
  geom_vline(xintercept = true_pop_mean) +
  ggtitle("Sampling Distribution\nof Sample Mean") +
  xlab("Sample Mean") + theme_bw()
```



An Analytical Approach to the Sampling Distributions

- The sampling distribution tells us how the estimator performs over many hypothetical samples.
- Unfortunately in real-world analysis we don't get to see the whole distribution, just one draw!
- Because the estimator is a random variable (remember it is a function of the statistic!) we can characterize the sampling distribution using the same tools from last week.
- We will start with a common estimator, the sample mean, $\overline{X}_n = \frac{1}{n}\sum_{i=1} X_i$.
- Under the identically and independently distributed assumption we can characterize properties of the distribution like the expectation and variance.

Describing the Sampling Distribution for the Sample Mean

We would like a full description of the sampling distribution for the sample mean estimator, but it will be useful to separate this description into three parts.

If we assume that $X_1, \dots X_n \sim_{i.i.d} \; ?(\mu, \sigma^2)$, then we would like to identify the following things about $\overline{X}_n$.

- $E[\overline{X}_n]$
- $V[\overline{X}_n]$
- $f_{\overline{X}_n} \sim ?$

Expectation of $\overline{X}_n$

Let $X_1, X_2, \ldots X_n$ be identically and independently distributed from a population distribution with mean ($E[X_i] = \mu$) and variance ($V[X_i] = \sigma^2$).

$$\begin{aligned}
E[\overline{X}_n] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n}E[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n \times \mu \\
&= \mu
\end{aligned}$$

Variance of $\overline{X}_n$

Let $X_1, X_2, \ldots X_n$ be identically and independently distributed from a population distribution with mean ($E[X_i] = \mu$) and variance ($V[X_i] = \sigma^2$).

$$\begin{aligned}
V[\overline{X}_n] &= V[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2} V[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} V[X_i] \text{(because i.i.d)} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n \times \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Note the n in the denominator: as we have more observations, the variance of the sampling distribution will shrink.

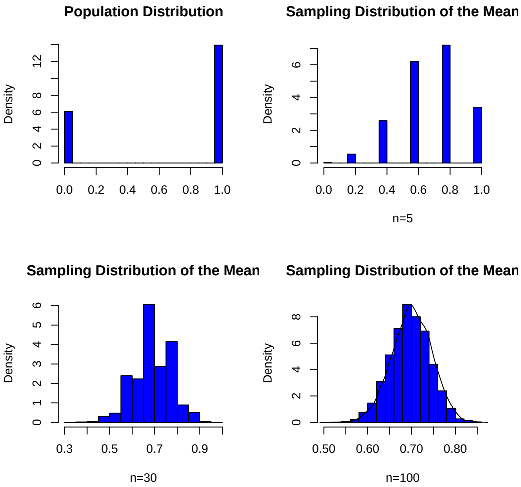
What about the “?”

If $X_1, \ldots, X_n \sim_{i.i.d.} N(\mu, \sigma^2)$, then

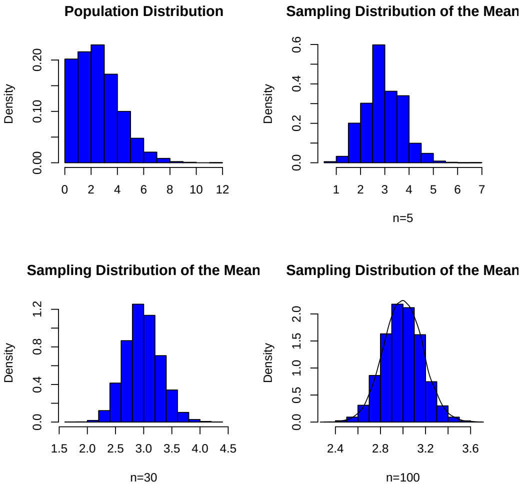
$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

What if $X_1, \ldots, X_n$ are not normally distributed?

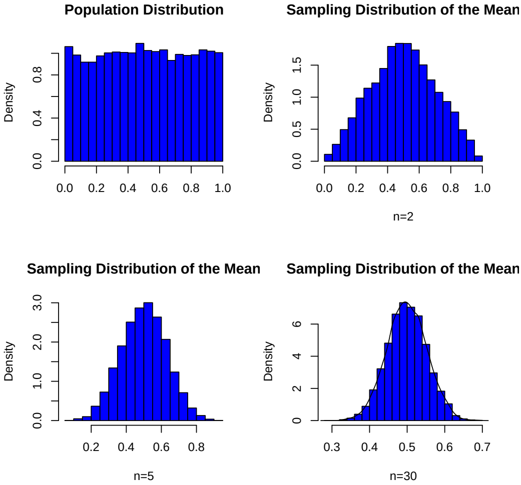
Bernoulli (Coin Flip) Distribution



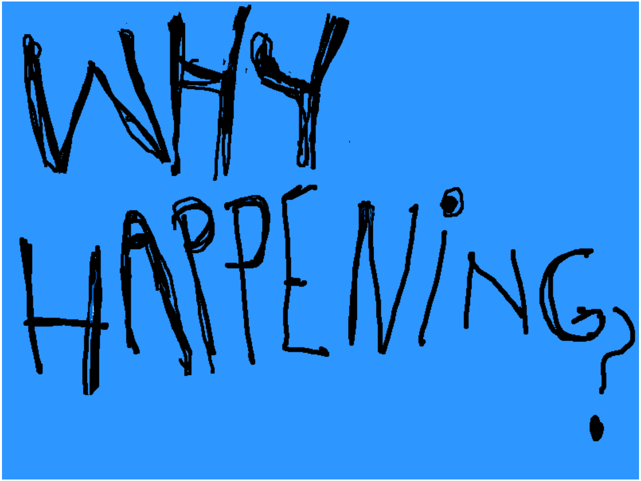
Poisson (Count) Distribution



Uniform Distribution



Why would this be true?



Images from *Hyperbole and a Half* by Allie Brosh.

We Covered. . .

- Populations and samples.
- Estimators, estimands and estimates.
- Sampling distributions.

Next time: the answer to 'why happening?' and the most important theorem in statistics.

Where We've Been and Where We're Going...

- Last Week
  - random variables
  - joint distributions
- This Week
  - estimators and sampling distributions
  - estimator properties (bias, variance, consistency)
  - confidence intervals
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability → inference → regression → causal inference

1. Estimation
   - Populations and Samples
   - Estimators
   - Analytical
2. Weak Law of Large Numbers and the Central Limit Theorem
   - Chebychev's Inequality
   - Weak Law of Large Numbers
   - The Central Limit Theorem
3. Properties of Estimators
   - Four Desirable Properties
   - Example
4. Interval Estimation
   - Intervals
   - Large Sample Intervals for a Mean
   - Small Sample Intervals for a Mean
   - Comparing Two Groups
   - Interval Estimation for a Proportion
5. Plug-In Principle

Cliffhanger

We started last time on a statistical cliffhanger: why did the sampling distribution of the mean look the same as the number of observations increased regardless of the population distribution?

The answer has to do with the asymptotic behavior of the estimator—what happens as the sample size n increases.

In order to characterize this formally though, we are going to have to set up some more probabilistic infrastructure.

The key thing to remember is that the sample mean is itself a random variable.

Warning: This video is a little bit mathier. At the end I'll wrap up with things you need to know so don't stress out your first watch through.

Bounding a Random Variable

Let's start by seeing how much we can learn about an estimator using only the things we've calculated so far (the expectation and variance).

Theorem (Chebychev's Inequality)

Let X be a random variable with finite $\sigma[X] > 0$. Then $\forall \epsilon > 0$,

$$P\big[|X - E[X]| \geq \epsilon\sigma[X]\big] \leq \frac{1}{\epsilon^2}$$

(Aronow and Miller Theorem 2.1.18)

This allows us to put an upper bound on the probability that a draw from the distribution will be more than a given number of standard deviations from the mean.

This let's us bound the behavior of a random variable knowing only the expectation and variance (regardless of distributional shape!).

Chebychev's Inequality for the Sample Mean

To apply this to the sample mean, we plug in the expectation and variance to the Chebychev's inequality and re-arranging terms, we get the following handy result.

Theorem (Chebychev's Inequality for the Sample Mean)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite variance $V[X] > 0$. Then, $\forall \epsilon > 0$,

$$P\left[|\overline{X}_n - E[X]| \geq \epsilon\right] \leq \frac{V[X]}{\epsilon^2 n}$$

(Aronow and Miller Theorem 3.2.5)

This allows us to put an upper bound on the probability that the sample mean for a given sample size and known variance will be some arbitrary distance from the true mean.

Planning a Survey

- Problem: planning a survey to estimate Biden support in the 2020 election.
  - How many people should we survey so that our estimator has no less than a .95 probability of being within .02 of the true population proportion?
  - i.e. how do we get a margin of ± 2 percentage points?
- Notation
  - π is proportion of voters expressing support for Biden (the estimand).
  - $X_1, X_2, \dots X_n \sim \text{Bernoulli}(\pi)$ be the iid random variables for each respondent.
  - $\hat{\pi} = \overline{X}_n$ is our sample mean estimator.
- What do we know?
  - $E[\overline{X}_n] = E[X]$ and $V[\overline{X}_n] = \frac{V[X]}{n}$
  - We know Bernoulli has variance of $\pi(1-\pi)$ which is maximized at $\pi = .5$.

Planning a Survey

What does n have to be to maintain $P(|\overline{X}_n - E[X]| \geq 0.02) \leq .05$?

$$\begin{aligned}
P\left[|\overline{X}_n - E[X]| \geq \epsilon\right] &\leq \frac{V[X]}{\epsilon^2 n} \\
P\left[|\overline{X}_n - E[X]| \geq 0.02\right] &\leq \frac{V[X]}{(0.02^2 n} \\
&\leq \frac{\pi(1-\pi)}{0.0004n} \\
&\leq \frac{0.25}{0.0004n} \\
&\leq \frac{1}{0.0016n}
\end{aligned}$$

We want to bound the probability by 0.05 which requires $\frac{1}{0.0016n} \leq 0.05$ which means... we need $n \geq 12,500$ respondents!

Planning a Survey

That's an expensive survey! Do we really need that many people?

No! Chebyshev provides a bound that is guaranteed to hold (and in the worst case variance), but actual probabilities are much smaller.

We can use simulation to assess. Let's simulate a survey with $\pi = .55$ and 12,500 respondents and see how many are far away?

```
nsims <- 10000
holder <- vector(mode="numeric", length=nsims)
for (i in 1:nsims) {
  my.sample <- rbinom(n=12500, size=1, prob=.55)
  holder[i] <- mean(my.sample)
}
mean(abs(holder - .55) > 0.02)
```

None were outside the range!

Taking Stock

- We have IID random variables $X_1, \ldots, X_n$ with unknown distribution.
- Current knowledge about distribution of $\overline{X}_n$
  - Expectation is $E[\overline{X}_n] = E[X]$
  - Variance is $V[\overline{X}_n] = \frac{V[X]}{n}$
  - Tail probabilities using the above and Chebyshev
- We still want to know more about the distribution of $\overline{X}_n$.
- We can think about behavior of $\overline{X}_n$ as n gets large by thinking of the estimators with increasing sample sizes as a sequence of random variables

$$\overline{X}_1, \overline{X}_2, ..., \overline{X}_n \; = \; X_1, \frac{X_1 + X_2}{2}, ..., \frac{X_1 + \cdots X_n}{n}$$

- What does this sequence converge to?

What do we mean by 'converge'?

Definition (Convergence in Probability)

Let $(T_{(1)}, T_{(2)}, T_{(3)}, \dots)$ be a sequence of random variables and let $c \in \mathbb{R}$. Then $T_{(n)}$ converges in probability to c if for all accuracy levels satisfying $\epsilon > 0$,

$$\lim_{n\to\infty} P\left[|T_{(n)} - c| \geq \epsilon\right] = 0$$

We will write this as

$$T_{(n)} \xrightarrow{p} c \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} T_{(n)} = c.$$

(Aronow and Miller Theorem 3.2.6)

Intuition: the probability that the random variable $T_{(n)}$ lies outside a super tiny interval around c approaches zero as n approaches infinity.

NB: Any continuous function of the sequence itself convergence to the value of the function at the probability limit by the Continuous Mapping Theorem (Aronow and Miller Theorem 3.2.7)

(Weak) Law of Large Numbers (WLLN)

Definition (Weak Law of Large Numbers)

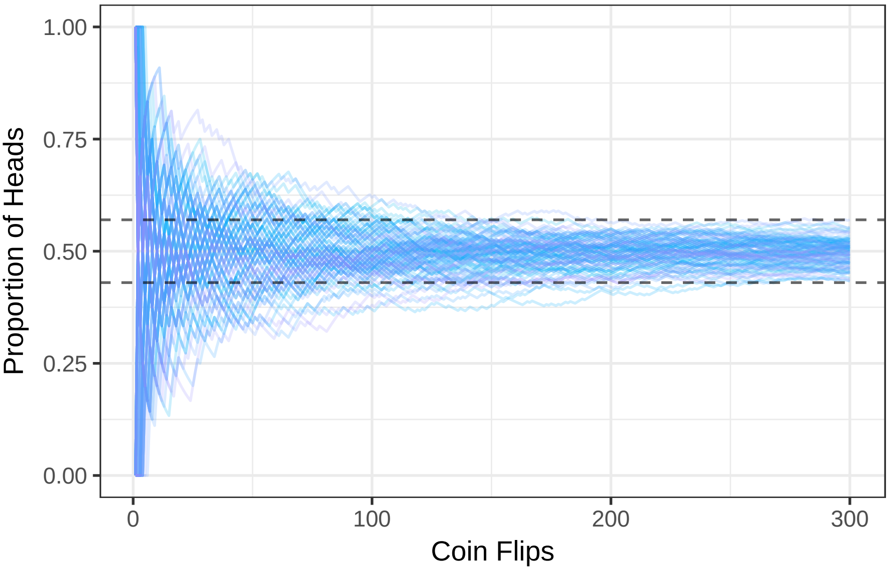
Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with finite variance $V[X] > 0$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\overline{X}_n \xrightarrow{p} E[X]$$

(Aronow and Miller Theorem 3.2.8) Proof is based on Chebychev's inequality for the mean plus a result called the Squeeze Theorem for Sequences.

- Intuition: The probability of the sample mean being far away from the expectation of X goes to zero as the sample size gets big.
- The distribution of $\overline{X}_n$ collapses on $E[X]$.
- No assumptions necessary about the distribution of X beyond i.i.d. sampling and a finite variance!

Weak Law of Large Numbers



Weak Law of Large Numbers

- This is an incredibly useful result!
- As the sample mean gets large it approximates the expectation to any arbitrary degree of precision.
- An implication of the Weak Law of Large Numbers is that the CDF of X can be estimated to arbitrary precision with random iid samples from X. We will return to this result in two videos.

Okay that's pretty cool, but we are almost ready to state the coolest result in statistics.

Convergence in Distribution

We want to know what form the sampling distribution will have asymptotically. For this we need a notion of what it means for a distribution to converge.

Definition (Convergence in Distribution)

Let $(T_{(1)}, T_{(2)}, T_{(3)}, \dots)$ be a sequence of random variables with CDFs $(F_{(1)}, F_{(2)}, F_{(3)}, \dots)$ and let T be a random variable with CDF F_T. Then $T_{(n)}$ converges in distribution to T if for all $t \in \mathbb{R}$at which F_T is continuous

$$\lim_{n\to\infty} F_{(n)}(t) = F_T(t).$$

We write this as

$$T_{(n)} \overset{d}{\to} T.$$

- Intuition: when n is big, the distribution of $T_{(n)}$ is very similar to F_T, the distribution of T.
- We will call this the asymptotic distribution or the limit distribution.
- NB: convergence in probability is a special case of convergence in distribution with a degenerate distribution.

Standardizing the Sample Mean

Last prerequisite!

Definition (Standardizing a Random Variable)

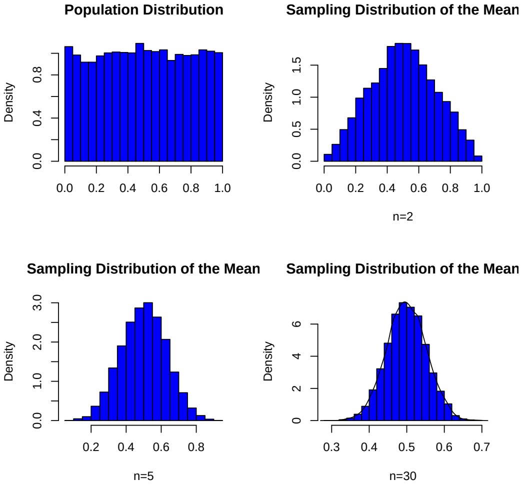
For i.i.d. random variables $X_1, X_2, \dots, X_n$ with finite $E[X] = \mu$ and finite $V[X] = \sigma^2 > 0$, the standardized sample mean is

$$Z = \frac{\left(\overline{X} - E[\overline{X}]\right)}{\sigma[\overline{X}]} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$$

(Aronow and Miller Definition 3.2.23)

- For any X this will have $E[Z] = 0$ and $V[Z] = 1$.
- This is often called the Z-score.

The Puzzle



Population Distribution
Density
0.0
0.4
0.8
0.0 0.2 0.4 0.6 0.8 1.0
Sampling Distribution of the Mean
Density
0.0
0.5
1.0
1.5
0.0 0.2 0.4 0.6 0.8 1.0
n=2
Sampling Distribution of the Mean
Density
0.0
1.0
2.0
3.0
0.2 0.4 0.6 0.8
n=5
Sampling Distribution of the Mean
Density
0
2
4
6
0.3 0.4 0.5 0.6 0.7
n=30

The Central Limit Theorem

Definition (Lindeberg-Lévy Central Limit Theorem)

Let $X_1, ..., X_n$ be i.i.d. random variables each with (finite) $E[X] = \mu$ and finite variance $\sigma^2 > 0$. Then, for *any* population distribution of X,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

- CLT also implies that the standardized sample mean converges to a standard normal random variable:

$$Z_n \equiv \frac{\overline{X}_n - E\left[\overline{X}_n\right]}{\sqrt{V\left[\overline{X}_n\right]}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

- This is free of distribution assumptions on X!
- This makes it easy to characterize the sampling distribution of the sample mean for large n.
- NB: the equivalence of the two forms is due to Slutsky's Theorem (see e.g. Aronow and Miller Theorem 3.2.25).

Question:

As the number of observations in a dataset increases, which of the following is true?

A) The distribution of X becomes more normally distributed.

B) The distribution of $\overline{X}$ becomes more normally distributed.

C) Both statements are true.

Replanning that Survey

Recall we wanted to find n such that,

$$P(|\overline{X}_n - \pi| > 0.02) \leq 0.05$$

By the CLT, for large n, then

$$\overline{X}_n - \pi \sim \mathcal{N}(0, \sigma^2/n)$$

Plugging in our conservative variance of 0.25 we get

$$\overline{X}_n - \pi \sim \mathcal{N}\left(0, \frac{1}{4n}\right)$$

Standardizing, we get

$$Z = \frac{\overline{X}_n - \pi}{1/\sqrt{4n}} = 2\sqrt{n}(\overline{X}_n - \pi) \sim \mathcal{N}(0,1)$$

It is easier to work with this standardized variable so:

$$P(|Z| > 0.02(2\sqrt{n})) \leq 0.05$$

Replanning that Survey

$$\begin{aligned} P(|Z| > 0.04\sqrt{n}) &\leq 0.05 \\ P(Z < -0.04\sqrt{n}) + P(Z > 0.04\sqrt{n}) &\leq 0.05 \end{aligned}$$

The standard normal is symmetric around 0, so we can equivalently say,

$$\begin{aligned} 2P(Z < -0.04\sqrt{n}) &\leq 0.05 \\ P(Z < -0.04\sqrt{n}) &\leq 0.025 \end{aligned}$$

To solve for n we plug in the quantile $P(Z \leq q) = 0.025$ which we can get from the inverse CDF of the standard Normal.
Typing `qnorm(0,025, mean=0, sd=1)` in `R` gets us -1.96.
We need $-0.04\sqrt{n} \leq -1.96$ which is $n > 2401$ respondents.
This is much lower than the 12,500 from Chebyshev, but that makes sense here because we used more information.

Planning a Survey

We can use simulation to assess again. Let's simulate a survey with $\pi = .55$ and 2,401 respondents and see how many our outside our prescribed margin of error?

```
nsims <- 10000
holder <- vector(mode="numeric", length=nsims)
for (i in 1:nsims) {
  my.sample <- rbinom(n=2401, size=1, prob=.55)
  holder[i] <- mean(my.sample)
}
mean(abs(holder - .55) > 0.02)
```

We get 0.0485!

Real Talk: this has been a mathy video.

What you need to know:

- two types of stochastic convergence
  - Convergence in probability: values in the sequence eventually take a constant value
    (i.e. the limiting distribution is a point mass)
  - Convergence in distribution: values in the sequence continue to vary, but the variation eventually comes to follow an unchanging distribution
    (i.e. the limiting distribution is a well characterized distribution)
- intuition for the weak law of large numbers
- means will asymptotically have normal sampling distributions due to the central limit theorem
- what asymptotic properties are

It is okay if you didn't follow all the math here. We will keep coming back to these ideas.

The Central Limit Theorem is deep and amazing. Want to learn more about Central Limit Theorem?

Watch this video (Joe Blitzstein):
https://www.youtube.com/watch?v=0prNqnHsVIA&list=PLLVplP8OIVc8EktkrD3Q8td0GmId7DjW0&index=31&t=0s

There are many CLT variants that deal with non-iid random variables as well!

We Covered. . .

- Chebychev's Inequality
- Weak Law of Large Numbers
- Central Limit Theorem

Next time: properties of estimators.

Where We've Been and Where We're Going...

- Last Week
  - random variables
  - joint distributions
- This Week
  - estimators and sampling distributions
  - estimator properties (bias, variance, consistency)
  - confidence intervals
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability → inference → regression → causal inference

1. Estimation
   - Populations and Samples
   - Estimators
   - Analytical

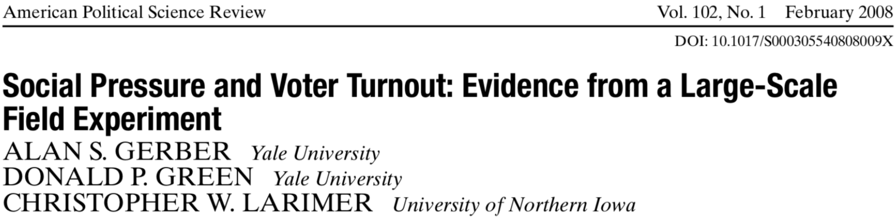
2. Weak Law of Large Numbers and the Central Limit Theorem
   - Chebychev's Inequality
   - Weak Law of Large Numbers
   - The Central Limit Theorem

3. Properties of Estimators
   - Four Desirable Properties
   - Example

4. Interval Estimation
   - Intervals
   - Large Sample Intervals for a Mean
   - Small Sample Intervals for a Mean
   - Comparing Two Groups
   - Interval Estimation for a Proportion

5. Plug-In Principle

Example: The scariest pieces of mail ever!



American Political Science Review Vol. 102, No. 1 February 2008

DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER *University of Northern Iowa*

`https://doi.org/10.1017/S000305540808009X`

Basic Analysis

They make their data available (https://isps.yale.edu/research/data/d001). We can analyze it.

```
load("gerber_green_larimer.RData")
## turn turnout variable into a numeric
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])
neigh.mean
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])
contr.mean
neigh.mean - contr.mean
```

$$.378 - .315 = .063$$

Is this a “real” effect? Is it big?

Desirable Properties of Estimators

Sometimes there are many possible estimators for a given parameter. Which one should we choose?

- We'd like an estimator that gets the right answer on average.
- We'd like an estimator that doesn't change much from sample to sample.
- We'd like an estimator that gets closer to the right answer (probabilistically) as the sample size increases.
- We'd like an estimator that has a known sampling distribution (approximately) when the sample size is large.

Properties of Estimators

Estimators are random variables, for which randomness comes from repeated sampling from the population.

The distribution of an estimator due to repeated sampling is called the sampling distribution.

The properties of an estimator refer to the characteristics of its sampling distribution.

Finite-sample Properties (apply for any sample size):

- Unbiasedness: Is the sampling distribution of our estimator centered at the true parameter value? $E[\hat{\mu}] = \mu$
- Efficiency: Is the variance of the sampling distribution of our estimator reasonably small? $V[\hat{\mu}_1] < V[\hat{\mu}_2]$

Asymptotic Properties (kick in when n is large):

- Consistency: As our sample size grows to infinity, does the sampling distribution of our estimator converge to the true parameter value?
- Asymptotic Normality: As our sample size grows large, does the sampling distribution of our estimator approach a normal distribution?

1: Bias

(not getting the right answer on average)

Definition

Bias is the expected difference between the estimator and the parameter. Over repeated samples, an unbiased estimator is right on average.

$$\begin{aligned} \text{Bias}(\hat{\mu}) &= E\left[\hat{\mu} - E[X]\right] \\ &= E\left[\hat{\mu}\right] - \mu \end{aligned}$$

Bias is not the difference between a particular estimate and the parameter. For example,

$$\text{Bias}(\overline{X}_n) \neq E\left[\overline{x}_n - E[X]\right]$$

An estimator is unbiased if and only if:

$$\text{Bias}(\hat{\mu}) = 0$$

Example: Estimators for Population Mean

Candidate estimators:

1. $\hat{\mu_1} = Y_1$ (the first observation)
2. $\hat{\mu_2} = \frac{1}{2}(Y_1 + Y_n)$ (average of the first and last observation)
3. $\hat{\mu_3} = 42$
4. $\hat{\mu_4} = \overline{Y}_n$ (the sample average)

How do we choose between these estimators?

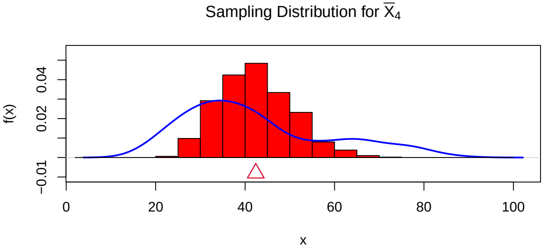
Bias of Example Estimators

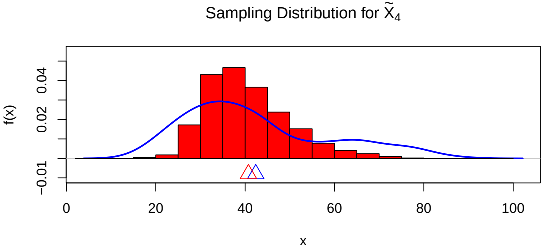
Which of these estimators are unbiased?

1. $E[Y_1 - \mu] = \mu - \mu = 0$
2. $E[\frac{1}{2}(Y_1 + Y_n) - \mu] = \frac{1}{2}(E[Y_1] + E[Y_n]) - \mu = \frac{1}{2}(\mu + \mu) - \mu = 0$
3. $E[42 - \mu] = 42 - \mu$
4. $E[\overline{Y}_n - \mu] = \frac{1}{n}\sum_1^n E[Y_i] - \mu = \mu - \mu = 0$

- Estimators 1,2, and 4 are unbiased because they get the right answer on average.
- Estimator 3 is biased.

Age population distribution in blue, sampling distributions in red





2: Efficiency

(doesn't change much sample to sample)

- All else equal, we prefer estimators that have a sampling distribution with smaller variance.

Definition (Efficiency)

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators of θ, then $\hat{\theta}_1$ is more efficient relative to $\hat{\theta}_2$ iff

$$V[\hat{\theta}_1] < V[\hat{\theta}_2]$$

- Under repeated sampling, estimates based on $\hat{\theta}_1$ are likely to be closer to θ
- Note that this does **not** imply that a particular estimate is always close to the true parameter value
- The standard deviation of the sampling distribution of an estimator, $\sqrt{V[\hat{\theta}]}$, is often called the standard error of the estimator

Aronow and Miller discuss efficiency in terms of MSE (more on this in a second).

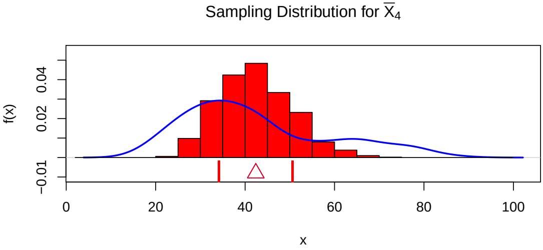
Variance of Example Estimators

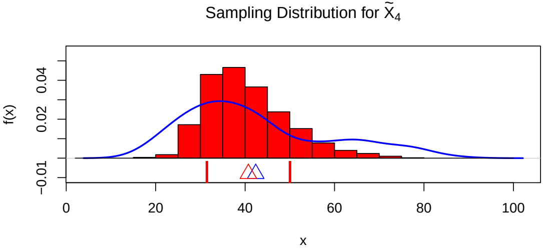
What is the variance of our estimators?

1. $V[Y_1] = \sigma^2$
2. $V[\frac{1}{2}(Y_1 + Y_n)] = \frac{1}{4}V[Y_1 + Y_n] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2$
3. $V[42] = 0$
4. $V[\overline{Y}_n] = \frac{1}{n^2}\sum_1^n V[Y_i] = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2$

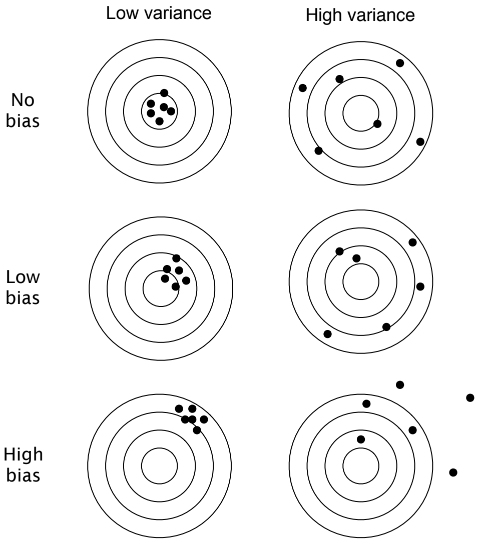
Among the unbiased estimators, the sample average has the smallest variance. This means that Estimator 4 (the sample average) is likely to be closer to the true value μ, than Estimators 1 and 2.

Age population distribution in blue, sampling distributions in red





Trading Off Bias and Variance



Salganik (2018), Figure 3.1

Mean Squared Error

How can we choose between an unbiased estimator and a biased, but lower variance estimator?

Definition (Mean Squared Error)

To compare estimators in terms of both efficiency and unbiasedness we can use the Mean Squared Error (MSE), the expected squared difference between $\hat{\theta}$ and θ:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Bias(\hat{\theta})^2 + V(\hat{\theta}) = \left[E[\hat{\theta}] - \theta\right]^2 + V(\hat{\theta})$$

Sometimes (as in Aronow and Miller Deinition 3.2.16) efficiency is defined as having lower MSE.

3-4: Asymptotic Evaluations: A Brief Review

(what happens as sample size increases)

- Unbiasedness and efficiency are finite-sample properties of estimators, which hold regardless of sample size
- Estimators also have asymptotic properties, i.e., the characteristics of sampling distributions when sample size becomes infinitely large
- To define asymptotic properties, consider a sequence of estimators at increasing sample sizes:

$$\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$$

- For example, the sequence of sample means ($\bar{X}_n$) is defined as:

$$\bar{X}_1, \bar{X}_2, ..., \bar{X}_n \;=\; X_1, \frac{X_1 + X_2}{2}, ..., \frac{X_1 + \cdots X_n}{n}$$

- Asymptotic properties of an estimator are defined by the behavior of $\hat{\theta}_1, ...\hat{\theta}_n$ when n goes to infinity.

3: Consistency

(does it get closer to the right answer as sample size increases)

Definition

An estimator $\hat{\theta}_n$ is consistent if the sequence $\hat{\theta}_1, ..., \hat{\theta}_n$ converges in probability to the true parameter value θ as sample size n grows to infinity:

$$\hat{\theta}_n \xrightarrow{p} \theta \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} \hat{\theta}_n = \theta$$

- Often seen as a minimal requirement for estimators
- A consistent estimator may still perform badly in small samples
- Two ways to verify consistency:
  1. Analytic: Often easier to check if $E[\hat{\theta}_n] \to \theta$ and $V[\hat{\theta}_n] \to 0$
  2. Simulation: Increase n and see how the sampling distribution changes
- Does unbiasedness imply consistency?
- Does consistency imply unbiasedness?

Deriving Consistency of Estimators

Our candidate estimators:

1. $\widehat{\mu}_1 = Y_1$
2. $\widehat{\mu}_2 = 4$
3. $\widehat{\mu}_3 = \overline{Y}_n \equiv \frac{1}{n}(Y_1 + \cdots + Y_n)$
4. $\widehat{\mu}_4 = \widetilde{Y}_n \equiv \frac{1}{n+5}(Y_1 + \cdots + Y_n)$

Which of these estimators are consistent for μ?

1. $E[\widehat{\mu}_1] = \mu$ and $V[\widehat{\mu}_1] = \sigma^2$
2. $E[\widehat{\mu}_2] = 4$ and $V[\widehat{\mu}_2] = 0$
3. $E[\widehat{\mu}_3] = \mu$ and $V[\widehat{\mu}_3] = \frac{1}{n}\sigma^2$
4. $E[\widehat{\mu}_4] = \frac{n}{n+5}\mu$ and $V[\widehat{\mu}_4] = \frac{n}{(n+5)^2}\sigma^2$

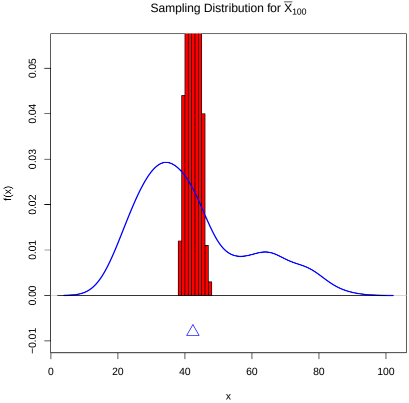
Consistency

The sample mean is a consistent estimator for μ.

$$\overline{X}_n \sim_{approx} N\left(\mu, \frac{\sigma^2}{n}\right)$$

As n increases, $\frac{\sigma^2}{n}$ approaches 0.

$$n = 125100$$



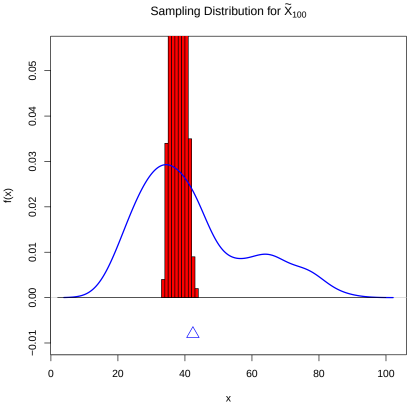
Inconsistency

An estimator can be inconsistent in several ways:

- The sampling distribution collapses around the wrong value
- The sampling distribution never collapses around anything

Inconsistency

Consider the median estimator: $\tilde{X}_n = \text{median}(Y_1, ..., Y_n)$ Is this estimator consistent for the expectation?
$n = 125100$



4: Asymptotic Distribution

(known sampling distribution for large sample size)

We are also interested in the shape of the sampling distribution of an estimator as the sample size increases.

Due to the central limit theorem, the sampling distributions of many estimators converge towards a normal distribution such that,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{V[\hat{\theta_n}]}} \xrightarrow{d} \mathcal{N}(0,1)$$

This will play a crucial role in our ability to form confidence intervals.

Summary of Properties

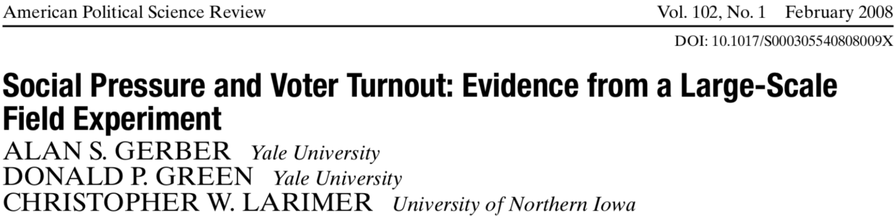
Concept	Criteria	Intuition
Unbiasedness	$E[\hat{\mu}] = \mu$	Right on average
Efficiency	$V[\hat{\mu}_1] < V[\hat{\mu}_2]$	Low variance
Consistency	$\hat{\mu}_n \xrightarrow{p} \mu$	Converge to estimand as $n \to \infty$
Asymptotic Normality	$\hat{\mu}_n \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n})$	Approximately normal in large n

How do we learn about an estimator?

Repeating theme in this class—how to characterize an estimator.

1. Define estimand of interest (causal quantity, survey outcome, model parameter)
2. Find an estimator for the quantity of interest
3. Is this estimator consistent?
4. Is this estimator unbiased?
5. What is the variance of the estimator?
6. Can we find an unbiased estimator for the variance (and is it consistent)?
7. What are the finite sample properties of the estimator?
8. What are the asymptotic properties of the estimator?

Back to the Example



American Political Science Review Vol. 102, No. 1 February 2008

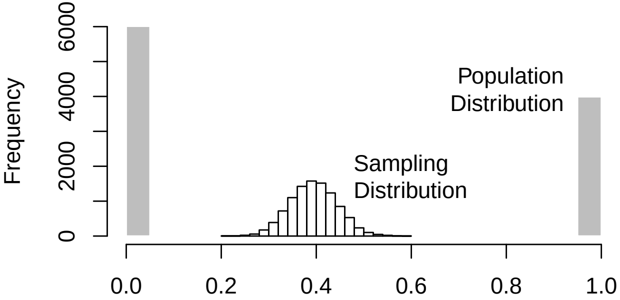
DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER *University of Northern Iowa*

Population vs. Sampling Distribution

We want to think about the sampling distribution of the estimator.



But remember that we only get to see one draw from the sampling distribution. Thus ideally we want an estimator with good properties.

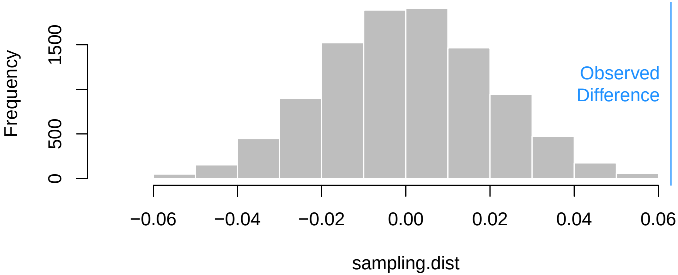
Asymptotic Normality

Going back to the Gerber, Green, and Larimer result...

- The estimator is difference in means
- The estimate is 0.063
- Suppose we have an estimate of the estimator's standard error $\hat{\text{SE}}(\hat{\theta}) = 0.02$.
- What if there was no difference in means in the population $(\mu_y - \mu_x = 0)$?
- By asymptotic Normality $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$
- By the properties of Normals, we know that this implies that $\hat{\theta} \sim \mathcal{N}(0, \text{SE}(\hat{\theta}))$

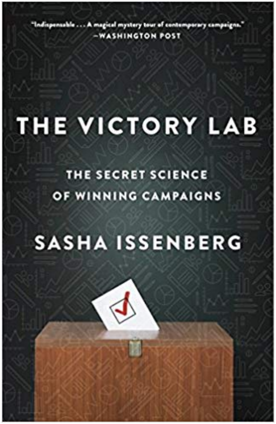
Asymptotic Normality

We can plot this to get a feel for it.



Does the observed difference in means seem plausible if there really were no difference between the two groups in the population?

The scariest pieces of mail ever! continued



Summarizes the relationships between political science research and campaigns. Also, attempts to weaponize the results of Gerber et al (2008).

We Covered...

- Four properties of estimators: bias, efficiency, consistency and asymptotic normality.
- A brief example of how we can use asymptotic normality in an example that will return!

Next Time: interval estimation

Where We've Been and Where We're Going...

- Last Week
  - random variables
  - joint distributions
- This Week
  - estimators and sampling distributions
  - estimator properties (bias, variance, consistency)
  - confidence intervals
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability → inference → regression → causal inference

1. Estimation
   - Populations and Samples
   - Estimators
   - Analytical

2. Weak Law of Large Numbers and the Central Limit Theorem
   - Chebychev's Inequality
   - Weak Law of Large Numbers
   - The Central Limit Theorem

3. Properties of Estimators
   - Four Desirable Properties
   - Example

4. Interval Estimation
   - Intervals
   - Large Sample Intervals for a Mean
   - Small Sample Intervals for a Mean
   - Comparing Two Groups
   - Interval Estimation for a Proportion

5. Plug-In Principle

What is Interval Estimation?

- A point estimator $\hat{\theta}$ estimates a scalar population parameter θ with a single number.
- However, because we are dealing with a random sample, we might also want to report uncertainty in our estimate.
- An interval estimator for θ takes the following form:

  $$[\hat{\theta}_{lower}, \hat{\theta}_{upper}]$$

  where $\hat{\theta}_{lower}$ and $\hat{\theta}_{upper}$ are random quantities that vary from sample to sample.
- The interval represents the range of possible values within which we estimate the true value of θ to fall.
- An interval estimate is a realized value from an interval estimator. The estimated interval typically forms what we call a confidence interval, which we will define shortly.

Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from with $E[X] = \mu$, $V[X] = 1$.

From previous lecture, we know that the sampling distribution of the sample average in large samples is:

$$\overline{X}_n \ \sim \ N(\mu, \sigma^2/n) = N(\mu, 1/n)$$

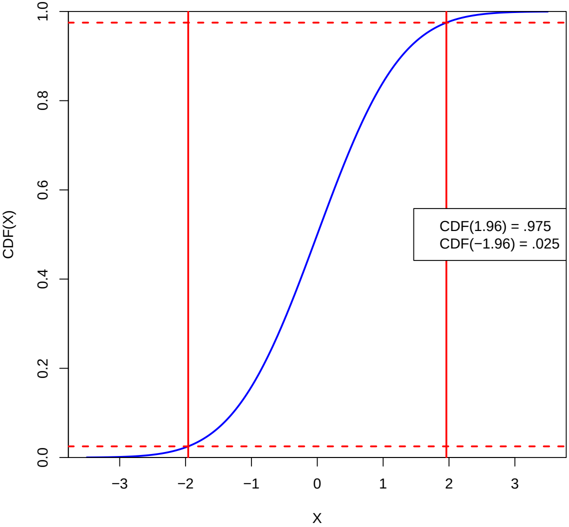
Therefore, the standardized sample average is distributed as follows:

$$\frac{\overline{X}_n - \mu}{1/\sqrt{n}} \ \sim \ N(0,1)$$

This implies

$$P\left(-1.96 < \frac{\overline{X}_n - \mu}{1/\sqrt{n}} < 1.96\right) \ = \ .95$$

CDF of the Standard Normal Distribution



Constructing a Confidence Interval with Known σ^2

So we know that:

$$P\left(-1.96 < \frac{\overline{X}_n - \mu}{1/\sqrt{n}} < 1.96\right) \;=\; .95$$

Rearranging yields:

$$P\left(\overline{X}_n - 1.96/\sqrt{n} \;<\; \mu \;<\; \overline{X}_n + 1.96/\sqrt{n}\right) \;=\; .95$$

This implies that the following interval estimator

$$\left[\,\overline{X}_n - 1.96/\sqrt{n}\,,\, \overline{X}_n + 1.96/\sqrt{n}\,\right]$$

contains the true population mean μ with probability 0.95.

We call this estimator a 95% confidence interval for μ.

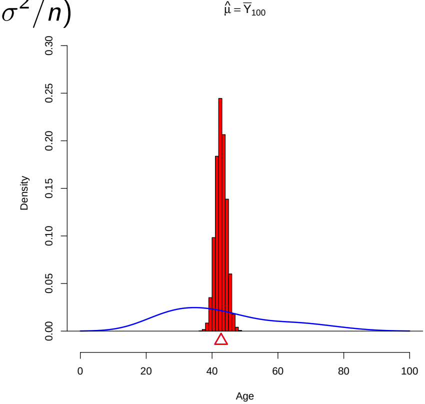
Kuklinski Example

$$\overline{Y} \sim_{approx} ?(?,?)?(\mu,?)?(\mu,\sigma^2/n)N(\mu,\sigma^2/n)$$

Suppose the 1,161 respondents in the Kuklinski data set were the population, with $\mu = 42.7$ and $\sigma^2 = 257.9$.

If we sampled 100 respondents, the sampling distribution of $\overline{Y}_{100}$ is:

$$\overline{Y}_{100} \sim_{approx} N(42.7, 2.579)$$

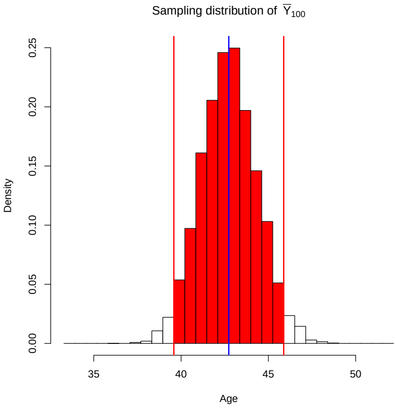


The standard error of $\overline{Y}$

The standard error of the sample mean is the standard deviation of the sampling distribution for $\overline{Y}$:

$$SE(\overline{Y}) = \sqrt{V(\overline{Y})} = \frac{\sigma}{\sqrt{n}}$$

What is the probability that $\overline{Y}$ falls within 1.96 SEs of μ?



Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2 and our previous example relied on that knowledge.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ where σ^2 is unknown. Then, as before,

$$\overline{X}_n \sim N(\mu, \sigma^2/n) \quad \text{and so} \quad \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Previously, we then constructed the interval:

$$\left[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}\right]$$

But we can **not** directly use this now because σ^2 is unknown.

Instead, we need an estimator of σ^2, $\hat{\sigma}^2$.

Estimators for the Population Variance

Two possible estimators of population variance:

$$S_{0n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

$$S_{1n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Which do we prefer? Let's check properties of these estimators.

1. Unbiasedness: We can show (after some algebra) that

$$E[S_{0n}^2] = \frac{n-1}{n}\sigma^2 \quad \text{and} \quad E[S_{1n}^2] = \sigma^2$$

2. Consistency: We can show that

$$S_{0n}^2 \ \overset{p}{\to}\ \sigma^2 \quad \text{and} \quad S_{1n}^2 \ \overset{p}{\to}\ \sigma^2$$

S_{1n}^2 (unbiased and consistent) is commonly called the sample variance.

Estimating σ and the SE

Returning to Kulinski et. al...

We will use the sample variance:

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

and thus the sample standard deviation can be written as

$$S = \sqrt{S^2}$$

We will plug in S for σ and our estimated standard error will be

$$\widehat{SE}[\hat{\mu}] = \frac{S}{\sqrt{n}}$$

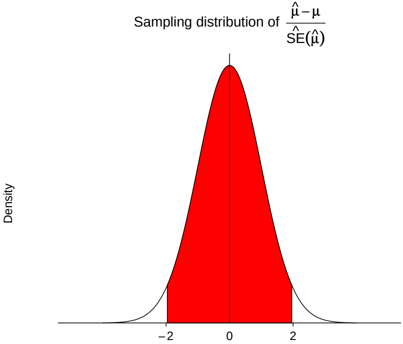
95% Confidence Intervals

If $X_1, ..., X_n$ are i.i.d. and n is large, then

$$\begin{aligned}\widehat{\mu} &\sim N(\mu, (\widehat{SE}[\hat{\mu}])^2) \\ \widehat{\mu} - \mu &\sim N(0, (\widehat{SE}[\hat{\mu}])^2) \\ \frac{\widehat{\mu} - \mu}{\widehat{SE}[\hat{\mu}]} &\sim N(0, 1)\end{aligned}$$

We know that

$$P\left(-1.96 \leq \frac{\widehat{\mu} - \mu}{\widehat{SE}[\hat{\mu}]} \leq 1.96\right) = 95\%$$



95% Confidence Intervals

We can work backwards from this:

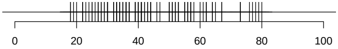
$$\begin{aligned}
P\left(-1.96 \leq \frac{\widehat{\mu} - \mu}{\widehat{SE}[\hat{\mu}]} \leq 1.96\right) &= 95\% \\
P\left(-1.96\widehat{SE}[\hat{\mu}] \leq \widehat{\mu} - \mu \leq 1.96\widehat{SE}[\hat{\mu}]\right) &= 95\% \\
P\left(\widehat{\mu} - 1.96\widehat{SE}[\hat{\mu}] \leq \mu \leq \widehat{\mu} + 1.96\widehat{SE}[\hat{\mu}]\right) &= 95\%
\end{aligned}$$

The random quantities in this statement are $\widehat{\mu}$ and $\widehat{SE}[\hat{\mu}]$.
Once the data are observed, nothing is random!

What does this mean?

We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

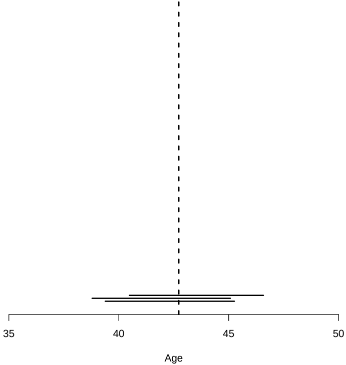


2) Calculate $\hat{\mu}$ and $\widehat{SE}[\hat{\mu}]$:

$$\hat{\mu} = 43.53 \quad \widehat{SE}[\hat{\mu}] = 1.555$$

3) Construct the 95% CI:

$$(40.5, 46.6)$$

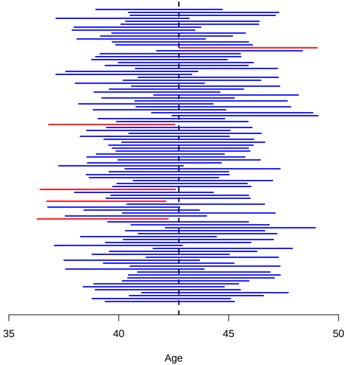


What does this mean?

By repeating this process, we generate the sampling distribution of the 95% CIs.

Most of the CIs cover the true μ; some do not.

In the long run, we expect 95% of the CIs generated to contain the true value.



Interpreting a Confidence Interval

This can be tricky, so let's break it down.

- Imagine we implement the interval estimator $\overline{X}_n \pm 1.96/\sqrt{n}$ for a particular sample and obtain the estimate of [2.5, 4].
- Does this mean that there is a .95 probability that the true parameter value μ lies between these two particular numbers? No!
- Confidence intervals are easy to construct, but difficult to interpret:
  - Each confidence interval estimat<u>e</u> from a particular sample either contains μ or not.
  - The probability statement is a property of the procedure. If we were to repeatedly calculate the interval estima<u>tor</u> over many random samples from the same population, 95% of the time the constructed confidence intervals would cover μ
  - Therefore, we refer to .95 as the coverage probability

What makes a good confidence interval?

1. The coverage probability: how likely it is that the interval covers the truth.
2. The length of the confidence interval:
    - Infinite intervals $(-\infty, \infty)$ have coverage probability 1
    - For a probability, a confidence interval of $[0, 1]$ also have coverage probability 1
    - Zero-length intervals, like $[\bar{Y}, \bar{Y}]$, have coverage probability 0

- You want the the shortest confidence interval with the desired coverage probability.

Is 95% all there is?

Our 95% CI had the following form: $\hat{\mu} \pm 1.96\widehat{SE}[\hat{\mu}]$

Remember where 1.96 came from?

$$P\left(-1.96 \leq \frac{\widehat{\mu} - \mu}{\widehat{SE}[\hat{\mu}]} \leq 1.96\right) = 95\%$$

What if we want a different percentage?

$$P\left(-z \leq \frac{\widehat{\mu} - \mu}{\widehat{SE}[\hat{\mu}]} \leq z\right) = (1 - \alpha)\%$$

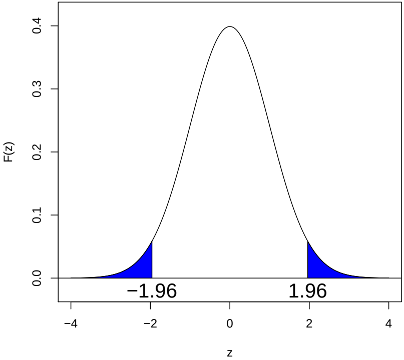
How can we find z?

Normal PDF

We know that z comes from the probability in the tails of the standard normal distribution.

When $(1 - \alpha) = 0.95$, we want to pick z so that 2.5% of the probability is in each tail.

This gives us a value of 1.96 for z.

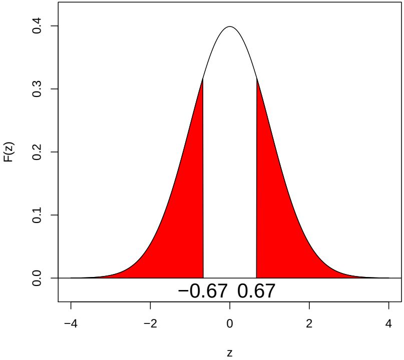


Normal PDF

What if we want a 50% confidence interval?

When $(1-\alpha) = 0.50$, we want to pick z so that 25% of the probability is in each tail.

This gives us a value of 0.67 for z.



$(1-\alpha)\%$ Confidence Intervals

In general, let $z_{\alpha/2}$ be the value associated with $(1-\alpha)\%$ coverage:

$$P\left(-z_{\alpha/2} \leq \frac{\widehat{\mu}-\mu}{\widehat{SE}[\hat{\mu}]} \leq z_{\alpha/2}\right) = (1-\alpha)\%$$

$$P\left(\widehat{\mu} - z_{\alpha/2}\widehat{SE}[\hat{\mu}] \leq \mu \leq \widehat{\mu} + z_{\alpha/2}\widehat{SE}[\hat{\mu}]\right) = (1-\alpha)\%$$

We usually construct the $(1-\alpha)\%$ confidence interval with the following formula.

$$\hat{\mu} \pm z_{\alpha/2}\widehat{SE}[\hat{\mu}]$$

Statistical problems emerge from real science

Comparing different methods of growing barley (Full history: `https://www.jstor.org/stable/2245613`)

`https://en.wikipedia.org/wiki/Guinness#/media/File:Guinness.jpg`

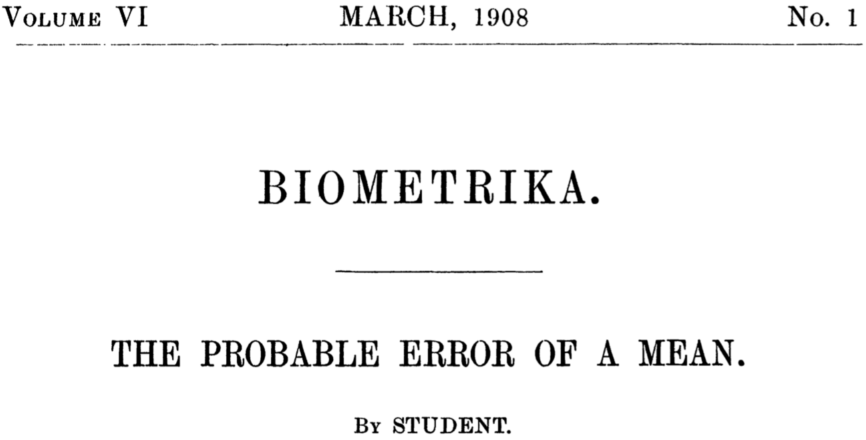
The problem with small samples

Up to this point, we have relied on large sample sizes to construct confidence intervals.

If the sample is large enough, then the sampling distribution of the sample mean follows a normal distribution.

If the sample is large enough, then the sample standard deviation (S) is a good approximation for the population standard deviation (σ).

When the sample size is small, we need to know something about the distribution in order to construct confidence intervals with the correct coverage (because we can't appeal to the CLT or assume that S is a good approximation of σ).



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

https://www.jstor.org/stable/2331554

Canonical Small Sample Example

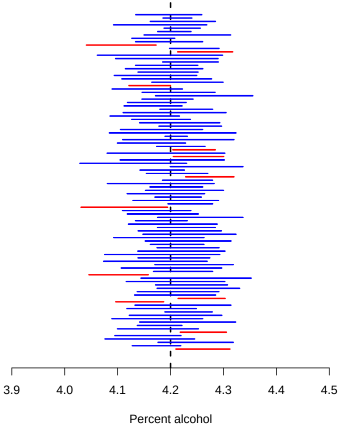
What happens if we use the large-sample formula?

The percent alcohol in Guinness beer is distributed $N(4.2, 0.09)$.

Take 100 six-packs of Guinness and construct CIs of the form

$$\hat{\mu} \pm 1.96\widehat{SE}[\hat{\mu}]$$

In this sample, only 88 of the 100 CIs cover the true value.



The t distribution

If X is normally distributed, then $\overline{X}$ is normally distributed even in small samples. Assume

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

If we know σ, then

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \sim \quad N(0,1)$$

We rarely know σ and have to use an estimate instead:

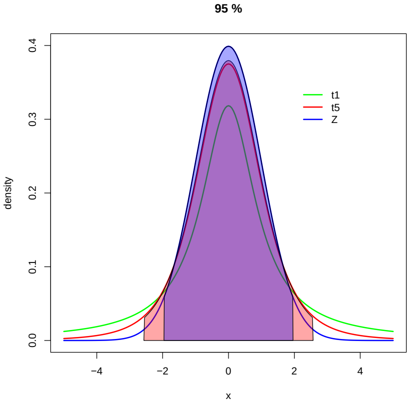
$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim ??t_{n-1}$$

The t distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but is more spread out.

As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

Eventually the t distribution converges to the standard normal.



$(1-\alpha)\%$ Confidence Intervals

In general, let $t_{\alpha/2}$ be the value associated with $(1-\alpha)\%$ coverage:

$$P\left(-t_{\alpha/2} \leq \frac{\widehat{\mu}-\mu}{\widehat{SE}[\hat{\mu}]} \leq t_{\alpha/2}\right) = (1-\alpha)\%$$

$$P\left(\widehat{\mu} - t_{\alpha/2}\widehat{SE}[\hat{\mu}] \leq \mu \leq \widehat{\mu} + t_{\alpha/2}\widehat{SE}[\hat{\mu}]\right) = (1-\alpha)\%$$

We usually construct the $(1-\alpha)\%$ confidence interval with the following formula.

$$\hat{\mu} \pm t_{\alpha/2}\widehat{SE}[\hat{\mu}]$$

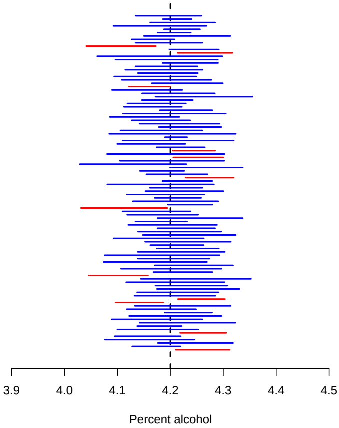
Small Sample Example

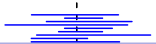
When we generated 95% CIs with the large sample formula

$$\hat{\mu} \pm 1.96\widehat{SE}[\hat{\mu}]$$

only 88 out of 100 intervals covered the true value.

When we use the correct small-sample formula





Another Rationale for the t-Distribution

Does $\overline{X}_n \sim N(\mu, S_n^2/n)$, which would imply $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim N(0,1)$?
No, because S_n is a random variable instead of a parameter (like σ).

Thus, we need to derive the sampling distribution of the new random variable. It turns out that T_n follows Student's t-distribution with $n-1$ degrees of freedom.

Theorem (Distribution of t-Value from a Normal Population)

Suppose we have an i.i.d. random sample of size n from $N(\mu, \sigma^2)$. Then, the sample mean $\overline{X}_n$ standardized with the estimated standard error $S_n/\sqrt{n}$ satisfies,

$$T_n \equiv \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim \tau_{n-1}$$

Kuklinski Example Returns

The Kuklinski et al. (1997) article compares responses to the baseline list with responses to the treatment list.

- How should we estimate the difference between the two groups?
- How should we obtain a confidence interval for our estimate?

Comparing Two Groups

We will often assume the following when comparing two groups,

- $X_{11}, X_{12}, ..., X_{1n_1} \sim_{i.i.d.} ?(\mu_1, \sigma_1^2)$
- $X_{21}, X_{22}, ..., X_{2n_2} \sim_{i.i.d.} ?(\mu_2, \sigma_2^2)$
- The two samples are independent of each other.

We will usually be interested in comparing μ_1 to μ_2, although we will sometimes need to compare σ_1^2 to σ_2^2 in order to make the first comparison.

Sampling Distribution for $\overline{X}_1 - \overline{X}_2$

What is the expected value of $\overline{X}_1 - \overline{X}_2$?

$$
\begin{aligned}
E[\overline{X}_1 - \overline{X}_2] &= E[\overline{X}_1] - E[\overline{X}_2] \\
&= \frac{1}{n_1}\sum E[X_{1i}] - \frac{1}{n_2}\sum E[X_{2j}] \\
&= \frac{1}{n_1}\sum \mu_1 - \frac{1}{n_2}\sum \mu_2 \\
&= \mu_1 - \mu_2
\end{aligned}
$$

Sampling Distribution for $\overline{X}_1 - \overline{X}_2$

What is the variance of $\overline{X}_1 - \overline{X}_2$?

$$
\begin{aligned}
Var[\overline{X}_1 - \overline{X}_2] &= Var[\overline{X}_1] + Var[\overline{X}_2] \\
&= \frac{1}{n_1^2} \sum Var[X_{1i}] + \frac{1}{n_2^2} \sum Var[X_{2j}] \\
&= \frac{1}{n_1^2} \sum \sigma_1^2 + \frac{1}{n_2^2} \sum \sigma_2^2 \\
&= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
\end{aligned}
$$

Sampling Distribution for $\overline{X}_1 - \overline{X}_2$

What is the distributional form for $\overline{X}_1 - \overline{X}_2$?

- $\overline{X}_1$ is distributed $\sim N(\mu_1, \frac{\sigma_1^2}{n_1})$.
- $\overline{X}_2$ is distributed $\sim N(\mu_2, \frac{\sigma_2^2}{n_2})$.
- $\overline{X}_1 - \overline{X}_2$ is distributed $\sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.

CIs for $\mu_1 - \mu_2$

Using the same type of argument that we used for the univariate case, we write a $(1-\alpha)\%$ CI as the following:

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Interval estimation of the population proportion

- Let's say that we have a sample of iid Bernoulli random variables, $Y_1, \ldots, Y_n$, where each takes $Y_i = 1$ with probability π. Note that this is also the population proportion of ones. We have shown in previous weeks that the expectation of one of these variable is just the probability of seeing a 1: $E[Y_i] = \pi$.
- The variance of a Bernoulli random variable is a simple function of its mean: $\text{Var}(Y_i) = \pi(1 - \pi)$.
- **Problem** Show that the sample proportion, $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, of the above iid Bernoulli sample, is unbiased for the true population proportion, π, and that the sampling variance is equal to $\frac{\pi(1-\pi)}{n}$.
- Note that if we have an estimate of the population proportion, $\hat{\pi}$, then we also have an estimate of the sampling variance: $\frac{\hat{\pi}(1-\hat{\pi})}{n}$.
- Given the facts from the previous problem, we just apply the same logic from the population mean to show the following confidence interval:

$$P\left(\hat{\pi} - z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = (1 - \alpha)$$

Gerber, Green, and Larimer experiment

Let's go back to the Gerber, Green, and Larimer experiment from last class. Here are the results of their experiment:

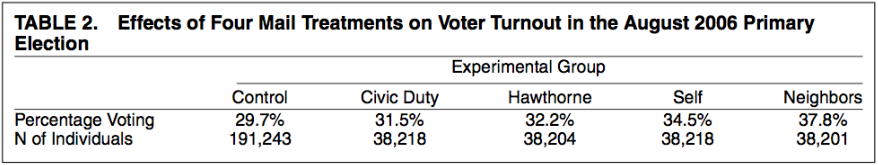


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Let's use what we have learned up until now and the information in the table to calculate a 95% confidence interval for the difference in proportions voting between the Neighbors group and the Civic Duty group.
- You may assume that the samples with in each group are iid and the two samples are independent.

Calculating the CI for social pressure effect

- We know distribution of sample proportion turned among Civic Duty group $\hat{\pi}_C \sim N(\pi_C, (\pi_C(1-\pi_C))/n_C)$
- Sample proportions are just sample means, so we can do difference in means:

$$\hat{\pi}_N - \hat{\pi}_C \sim N\left(\pi_N - \pi_C, \sqrt{SE_N^2 + SE_C^2}\right)$$

- Replace the variances with our estimates:

$$\hat{\pi}_N - \hat{\pi}_C \sim N\left(\pi_N - \pi_C, \sqrt{\widehat{SE}_N^2 + \widehat{SE}_C^2}\right)$$

- Apply usual formula to get 95% confidence interval:

$$(\hat{\pi}_N - \hat{\pi}_C) \pm 1.96 \times \sqrt{\widehat{SE}_N^2 + \widehat{SE}_C^2}$$

- Remember that we can calculate the sample variance for a sample proportion like so: $(\hat{\pi}_C(1-\hat{\pi}_C))/n_C$

Gerber, Green, and Larimer experiment

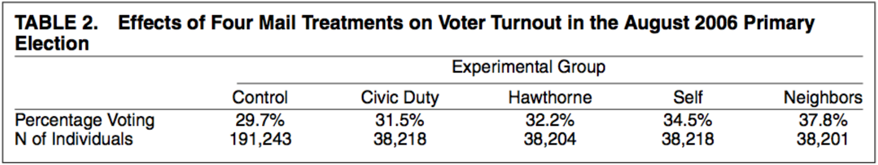


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Now, we calculate the 95% confidence interval:

$$(\hat{\pi}_N - \hat{\pi}_C) \pm 1.96 \times \sqrt{\frac{\hat{\pi}_N(1-\hat{\pi}_N)}{n_N} + \frac{\hat{\pi}_C(1-\hat{\pi}_C)}{n_C}}$$

```
n.n <- 38201
samp.var.n <- (0.378 * (1 - 0.378))/n.n
n.c <- 38218
samp.var.c <- (0.315 * (1 - 0.315))/n.c
se.diff <- sqrt(samp.var.n + samp.var.c)
## lower bound
(0.378 - 0.315) - 1.96 * se.diff
## [1] 0.05626701
## upper bound
(0.378 - 0.315) + 1.96 * se.diff
## [1] 0.06973299
```

Thus, the confidence interval for the effect is $[0.056267, 0.069733]$.

Review

We can use our analytic samples to find a confidence interval

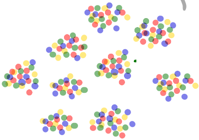
Our estimate

$$CI(\alpha) = [r - z_{\alpha/2} * SE, r + z_{\alpha/2} * SE]$$

Alpha

Standard error of our estimate

Critical value

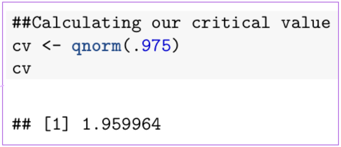


α/2 because we're looking for a two-sided interval

Review

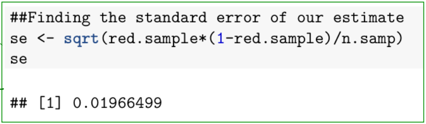
To use the confidence interval formula,
we need to find:

1. The distribution
2. Confidence level
   - Alpha
3. Sidedness
4. Critical value(s)



```
##Calculating our critical value
cv <- qnorm(.975)
cv

## [1] 1.959964
```

5. Standard error of our estimate



```
##Finding the standard error of our estimate
se <- sqrt(red.sample*(1-red.sample)/n.samp)
se

## [1] 0.01966499
```

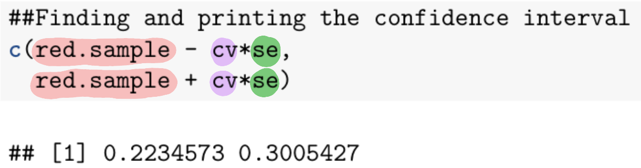
for a proportion, the formula is:

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

Review

Calculating the confidence interval

$$CI(\alpha) = [r - z_{\alpha/2} * SE, r + z_{\alpha/2} * SE]$$



```
##Finding and printing the confidence interval
c(red.sample - cv*se,
  red.sample + cv*se)
```

```
## [1] 0.2234573 0.3005427
```

Our results

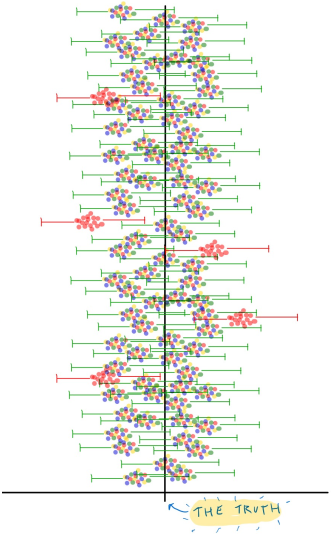
26.2% red with a 95 percent confidence interval of **[22.3, 30.1]**

Review

We hope our sample is in the 95%

samples whose confidence intervals contain the truth (95%)

samples whose confidence intervals do not contain the truth (5%)



We Covered. . .

- Interval estimates provide a means of assessing uncertainty.
- Interval estimators have sampling distributions.
- Interval estimates should be interpreted in terms of repeated sampling.

Next Time: The plug-in principle!

Where We've Been and Where We're Going...

- Last Week
  - random variables
  - joint distributions
- This Week
  - estimators and sampling distributions
  - estimator properties (bias, variance, consistency)
  - confidence intervals
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability → inference → regression → causal inference

1. Estimation
   - Populations and Samples
   - Estimators
   - Analytical
2. Weak Law of Large Numbers and the Central Limit Theorem
   - Chebychev's Inequality
   - Weak Law of Large Numbers
   - The Central Limit Theorem
3. Properties of Estimators
   - Four Desirable Properties
   - Example
4. Interval Estimation
   - Intervals
   - Large Sample Intervals for a Mean
   - Small Sample Intervals for a Mean
   - Comparing Two Groups
   - Interval Estimation for a Proportion
5. **Plug-In Principle**

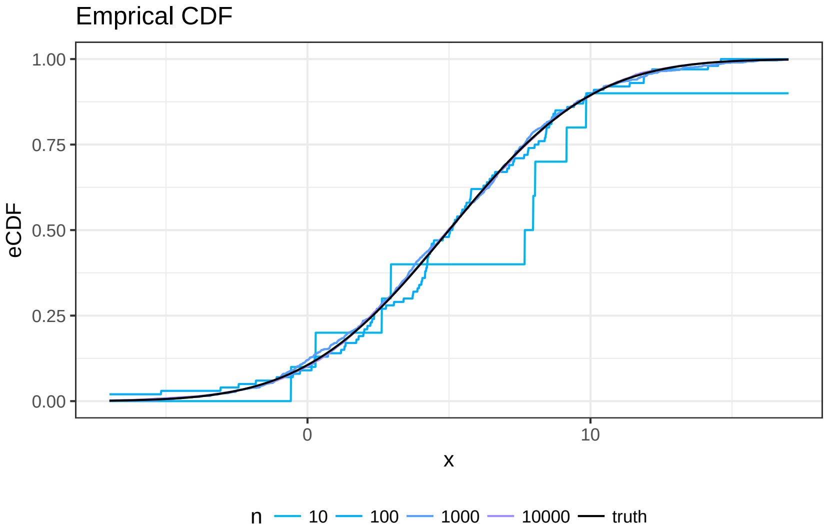
Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

- The simplest way is to use the sample analog.
- Ex: If we're interested in the population mean, we use the sample mean
- This is justified because of the plug-in principle.
- The Weak Law of Large Numbers tells us that the empirical CDF is a good sample analog of the true CDF (which fully describes a distribution).

The Plug-in Principle in Action

Say we have a $\mathcal{N}(5, 4)$ distribution



The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

we define the empirical CDF (eCDF) as:

$$\hat{F}(x) = \overline{\mathbb{I}(X \leq x)}, \forall x \in \mathbb{R}$$

- WLLN tells us that the eCDF will be unbiased and consistently estimated. Any given sample will, on average, look representative of the true distribution.

For iid random variables $X_1, X_2, \ldots, X_n$ with common CDF F, the plug-in estimator of $\theta = T(F)$ is:

$$\hat{\theta} = T(\hat{F})$$

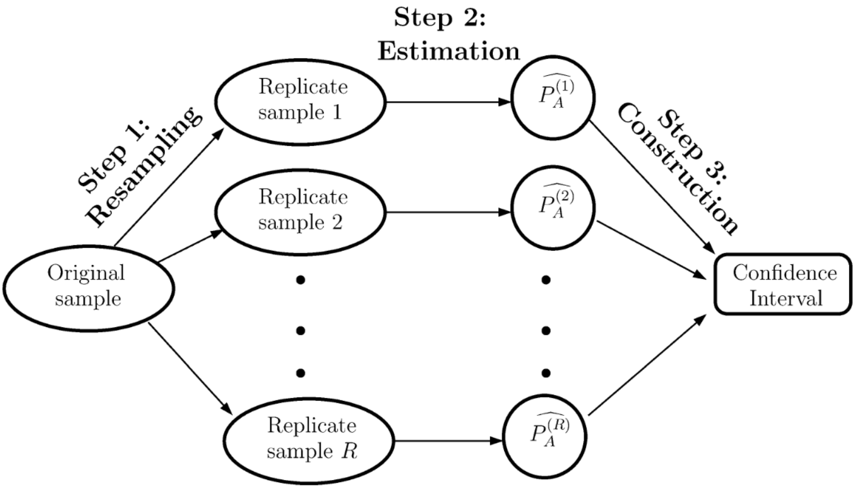
if T is well-behaved, then $\hat{\theta}$ is also asymptotically normal.

Bootstrapped Sampling Distributions

What if there was a way to replace thinking with computers?

What if there was a way to replacing analytical derivations, which can be hard, with computer simulations which are easy?

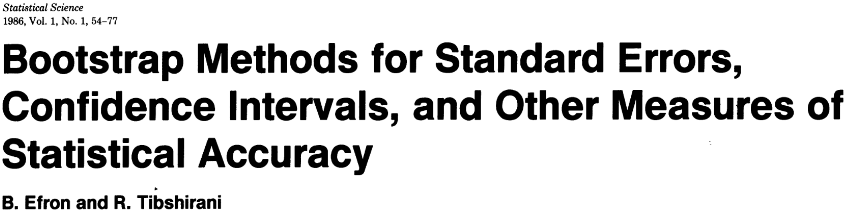
The plug-in principle gives us a way forward.



Source: Salganik (2006)

This works for almost* any estimator

*basically it works when plug-in estimation works



Statistical Science
1986, Vol. 1, No. 1, 54–77

Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy

B. Efron and R. Tibshirani

Efron and Tibshirani (1986), `http://www.jstor.org/stable/2245500`

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size *n* from our sample.
2. Calculate our would-be estimate using this bootstrap sample.
3. Repeat steps 1 and 2 many (B) times.
4. Using the resulting collection of bootstrap estimates, calculate the standard deviation of the bootstrap distribution of our estimator. This serves our estimate of the standard deviation of the sampling distribution

Example of a Bootstrap

```r
samp <- c(9.7, 4.99, 5.9, 3.58, 8.15, 5.54, 4.77, 5.01, 4.89,
          3.42, 8.63, 7.17, 8.93, 7.5, 4.93, 8.6, 6.26, 7.31,
          8.96, 3.95)

obs_mean = mean(samp)

obs_mean
```

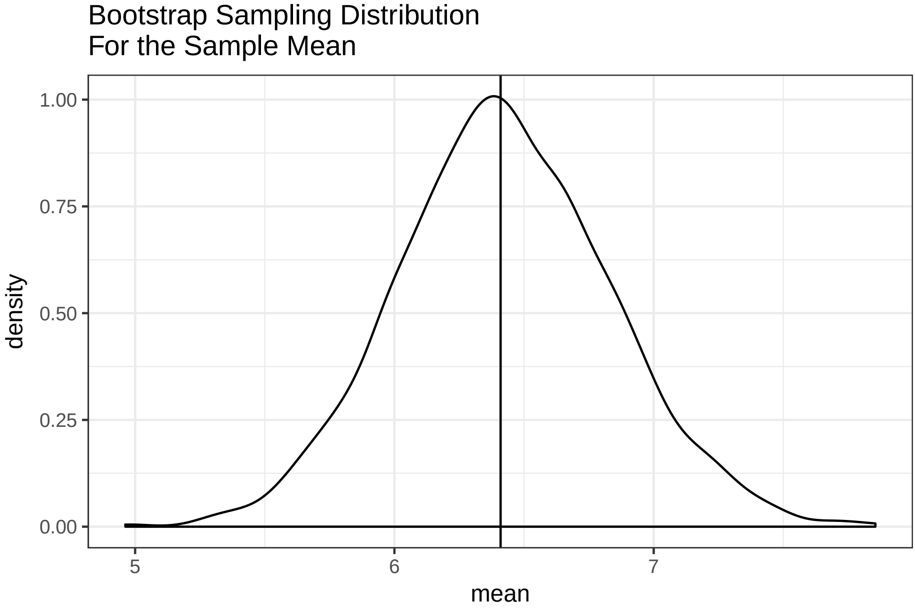
```
## [1] 6.4095
```

Example of a Bootstrap

```r
# resample WITH REPLACEMENT reps times
# recalculate the mean within each bootstrap replicate
boot_samp_dist <- replicate(2000, {
      mean(samp[sample.int(length(samp), replace = TRUE)])
    })
```

```r
ggplot(tibble(boot_samp_dist = boot_samp_dist),
       aes(x = boot_samp_dist)) +
  geom_density() +
  geom_vline(xintercept = obs_mean) +
  theme_bw() + ggtitle("Bootstrap Sampling Distribution
                       For the Sample Mean") +
  xlab("mean")
```

Example of a Bootstrap



Two ways to calculate bootstrap intervals

1) Using normal approximation intervals, use the estimates from step 4.

$$\left[\overline{X} - \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}, \overline{X} + \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}\right]$$

- Note here that the standard error is just the standard deviation of the boostrap replicates. There is no square root of n. Why?

2) Percentile method for the CI: Sort B bootstrap estimates from smallest to largest. α interval is constructed as

$$CI_{1-\alpha} = \left[\alpha/2 * B \text{ sample}, (1-\alpha/2) * B \text{ sample}\right]$$

- Percentile method does not rely on normal approximation, and behaves better with small n.

We covered

- The plug-in principle.
- The bootstrap.
- We will return to both in future weeks.

This Week in Review

- Estimation!
- Central Limit Theorem!
- Properties of Estimators!
- Intervals!
- Plug-In Principle!

Going Deeper:

Aronow and Miller (2019) *Foundations of Agnostic Statistics*. Cambridge University Press. Chapter 3.

Next week: hypothesis testing and regression!