Week 3: Learning from Random Samples

Brandon Stewart[1]

Princeton

September 26/28, 2016

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

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

- Last Week

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

- Last Week
  - random variables

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

- Last Week
  - random variables
  - joint distributions

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

- Last Week
  - random variables
  - joint distributions
- This Week

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?

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

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

Inference: given that we observe something in the data, what is our best guess of what it will be in other samples?

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

Inference: given that we observe something in the data, what is our best guess of what it will be in other samples?

- For the last few classes we have talked about probability—that is, if we knew how the world worked, we are describing what kind of data we should expect.

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

Inference: given that we observe something in the data, what is our best guess of what it will be in other samples?

- For the last few classes we have talked about probability—that is, if we knew how the world worked, we are describing what kind of data we should expect.
- Now we want to move the other way. If we have a set of data, can we estimate the various parts of the probability distributions that we have talked about. Can we estimate the mean, the variance, the covariance, etc?

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

Inference: given that we observe something in the data, what is our best guess of what it will be in other samples?

- For the last few classes we have talked about probability—that is, if we knew how the world worked, we are describing what kind of data we should expect.
- Now we want to move the other way. If we have a set of data, can we estimate the various parts of the probability distributions that we have talked about. Can we estimate the mean, the variance, the covariance, etc?
- Moving forward this is going to be very important. Why? Because we are going to want to estimate the population conditional expectation in regression.

Primary Goals for This Week

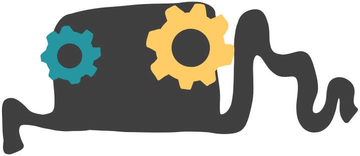
We want to be able to interpret the numbers in this table (and a couple of numbers that can be derived from these numbers).

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

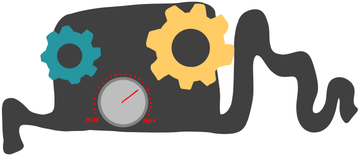
Region	Experimental Condition: Baseline	Experimental Condition: Black Family	Estimated Percent Angry
Non-South	2.28[a]	2.24	0
	(.07)	(.05)	
	425[b]	461	
South	1.95	2.37	42
	(.06)	(.08)	
	139	136	

[a]Standard error of the estimate.
[b]Number of cases.

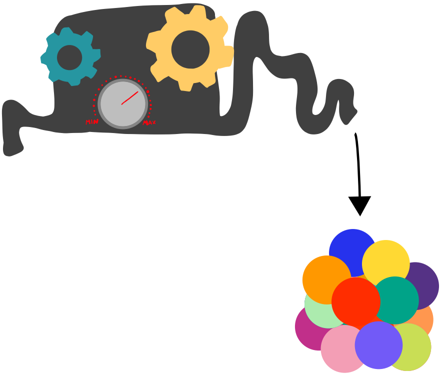
An Overview



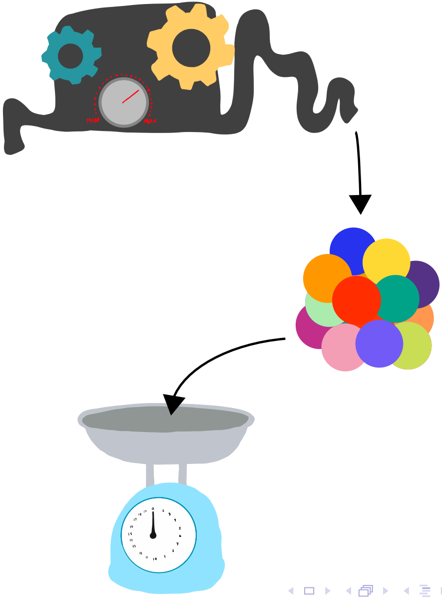
An Overview



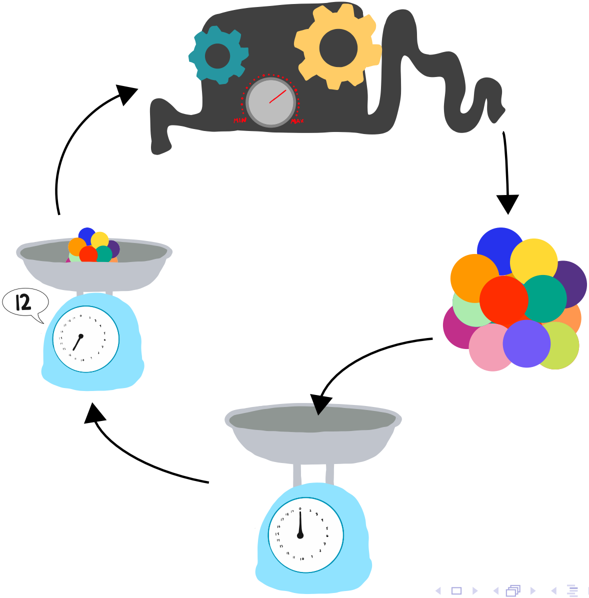
An Overview



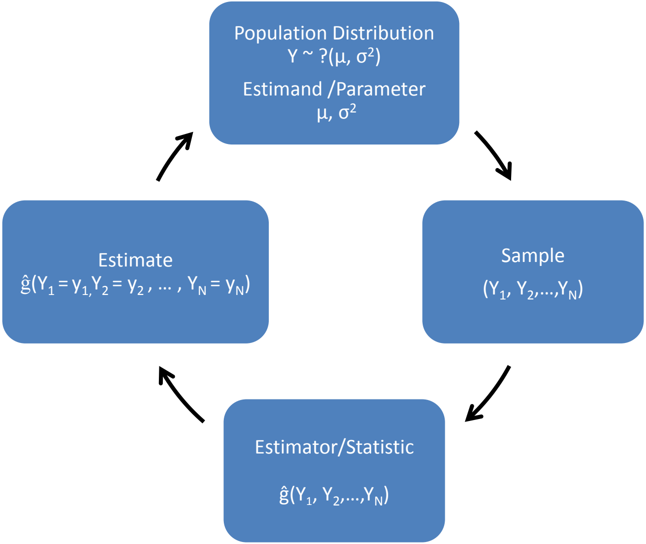
An Overview



An Overview



An Overview



Population Distribution
Y ~ ?(μ, σ2)
Estimand /Parameter
μ, σ2
Sample
(Y1, Y2,…,YN)
Estimator/Statistic
ĝ(Y1, Y2,…,YN)
Estimate
ĝ(Y1 = y1, Y2 = y2 , … , YN = yN)

Populations

Populations

- Typically, we want to learn about the distribution of random variable (or set of random variables) for a population of interest.

Populations

- Typically, we want to learn about the distribution of random variable (or set of random variables) for a population of interest.
- e.g. the distribution of votes for Hillary Clinton in the population of registered voters in the United States. This is an example of a finite population.

Populations

- Typically, we want to learn about the distribution of random variable (or set of random variables) for a population of interest.
- e.g. the distribution of votes for Hillary Clinton in the population of registered voters in the United States. This is an example of a finite population.
- Sometimes the population will be more abstract, such as the population of all possible television ads. This is an example of an infinite population.

Populations

- Typically, we want to learn about the distribution of random variable (or set of random variables) for a population of interest.
- e.g. the distribution of votes for Hillary Clinton in the population of registered voters in the United States. This is an example of a finite population.
- Sometimes the population will be more abstract, such as the population of all possible television ads. This is an example of an infinite population.
- With either a finite or infinite population our main goal in inference is to learn about the population distribution or particular aspects of that distribution, like the mean or variance, which we call a population parameter (or just parameter).

Population Distribution

Population Distribution

- We sometimes call the population distribution the data generating process and represent it with a pmf or pdf, $f(x; \theta)$.

Population Distribution

- We sometimes call the population distribution the data generating process and represent it with a pmf or pdf, $f(x;\theta)$.
- Ideally we would place no restrictions on f and learn everything we can about it from the data. This nonparametric approach is difficult due to the fact that the space of possible distributions is vast!

Population Distribution

- We sometimes call the population distribution the data generating process and represent it with a pmf or pdf, $f(x;\theta)$.
- Ideally we would place no restrictions on f and learn everything we can about it from the data. This nonparametric approach is difficult due to the fact that the space of possible distributions is vast!
- Instead, we will often make a parametric assumption and assume that the formula for f is known up to some unknown parameters.

Population Distribution

- We sometimes call the population distribution the data generating process and represent it with a pmf or pdf, $f(x; \theta)$.
- Ideally we would place no restrictions on f and learn everything we can about it from the data. This nonparametric approach is difficult due to the fact that the space of possible distributions is vast!
- Instead, we will often make a parametric assumption and assume that the formula for f is known up to some unknown parameters.
- Thus, f has two parts: the known part which is the formula for the pmf/pdf (sometimes called the parametric model and comes from the distributional assumptions) and the unknown part, which are the parameters, θ.

Population Distribution

Population Distribution

- For instance, suppose we have a binary r.v. such as intending to vote for Hillary Clinton ($X = 1$). Then we might assume that the population distribution is Bernoulli with unknown probability of $X = 1$, θ.

Population Distribution

- For instance, suppose we have a binary r.v. such as intending to vote for Hillary Clinton ($X = 1$). Then we might assume that the population distribution is Bernoulli with unknown probability of $X = 1$, θ.
- Thus we would have:

$$f(x;\theta) = \theta^x (1-\theta)^{1-x}$$

Our goal is to learn about the probability of someone voting for Hillary Clinton from a sample of draws from this distribution.

Population Distribution

- For instance, suppose we have a binary r.v. such as intending to vote for Hillary Clinton ($X = 1$). Then we might assume that the population distribution is Bernoulli with unknown probability of $X = 1$, θ.
- Thus we would have:

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}$$

  Our goal is to learn about the probability of someone voting for Hillary Clinton from a sample of draws from this distribution.
- Probability tells us what types of samples we should expect for different values of θ.

Population Distribution

- For instance, suppose we have a binary r.v. such as intending to vote for Hillary Clinton ($X = 1$). Then we might assume that the population distribution is Bernoulli with unknown probability of $X = 1$, θ.
- Thus we would have:

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}$$

  Our goal is to learn about the probability of someone voting for Hillary Clinton from a sample of draws from this distribution.
- Probability tells us what types of samples we should expect for different values of θ.
- For some problems, such as estimating the mean of a distribution, we actually won't need to specify a parametric model for the distribution allowing us to take an agnostic view of statistics.

Using Random Samples to Estimate Population Parameters

Let's look at possible estimators $\widehat{\mu}$ for the (unobserved) mean income in the population μ.

How can we use sample data to estimate μ?

Using Random Samples to Estimate Population Parameters

Let's look at possible estimators $\widehat{\mu}$ for the (unobserved) mean income in the population μ.

How can we use sample data to estimate μ?

If we think of a sample of size n as randomly sampled with replacement from the population, then $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.) random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$ for all $i \in \{1, ..., n\}$.

Using Random Samples to Estimate Population Parameters

Let's look at possible estimators $\widehat{\mu}$ for the (unobserved) mean income in the population μ.

How can we use sample data to estimate μ?

If we think of a sample of size n as randomly sampled with replacement from the population, then $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.) random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$ for all $i \in \{1, ..., n\}$.

In R, we can draw an i.i.d. random sample of size `n` from `population` by

```
sample(population, size=n, replace=TRUE)
```

where `population` is a vector that contains the Y_i values for all units in the population.

Using Random Samples to Estimate Population Parameters

Let's look at possible estimators $\widehat{\mu}$ for the (unobserved) mean income in the population μ.

How can we use sample data to estimate μ?

If we think of a sample of size n as randomly sampled with replacement from the population, then $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.) random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$ for all $i \in \{1, ..., n\}$.

In R, we can draw an i.i.d. random sample of size `n` from `population` by

```
sample(population, size=n, replace=TRUE)
```

where `population` is a vector that contains the Y_i values for all units in the population.

Our estimators, $\widehat{\mu}$, are functions of $Y_1, \ldots, Y_n$ and will therefore be random variables with their own probability distributions.

Why Samples?

Why Samples?

- The population is infinite (e.g., ballot order randomization process).

Why Samples?

- The population is infinite (e.g., ballot order randomization process).
- The population is large (e.g., all Southern adults with telephones).

Why Samples?

- The population is infinite (e.g., ballot order randomization process).
- The population is large (e.g., all Southern adults with telephones).
- The population includes counterfactuals.

Why Samples?

- The population is infinite (e.g., ballot order randomization process).
- The population is large (e.g., all Southern adults with telephones).
- The population includes counterfactuals.

  Even if our population is limited to the surveyed individuals in the Kuklinski et al. study, we might take the population of interest to include potential responses for each individual to both the treatment and baseline questions. For each individual we must sample one of these two responses (because we cannot credibly ask both questions).

Why Samples?

- The population is infinite (e.g., ballot order randomization process).
- The population is large (e.g., all Southern adults with telephones).
- The population includes counterfactuals.

  Even if our population is limited to the surveyed individuals in the Kuklinski et al. study, we might take the population of interest to include potential responses for each individual to both the treatment and baseline questions. For each individual we must sample one of these two responses (because we cannot credibly ask both questions).

  This occurs whenever we are interested in making causal inferences.

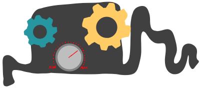
Estimands, Estimators, and Estimates

Estimands, Estimators, and Estimates

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

Estimands, Estimators, and Estimates

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

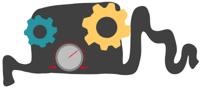


- Estimands are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$

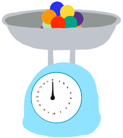
Estimands, Estimators, and Estimates

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

- Estimands are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



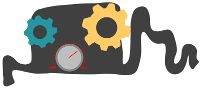
- Estimators are functions of sample data (i.e. statistics) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)



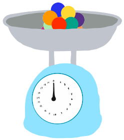
Estimands, Estimators, and Estimates

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

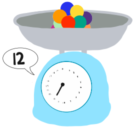
- Estimands are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N}\sum_{i=1}^{N} y_i$



- Estimators are functions of sample data (i.e. statistics) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)



- Estimates are particular values of estimators that are realized in a given sample (e.g. sample mean): $\frac{1}{n}\sum_{i=1}^{n} y_i$



What Are We Estimating?

What Are We Estimating?

- Our goal is to learn about the data generating process that generated the sample

What Are We Estimating?

- Our goal is to learn about the data generating process that generated the sample
- We might assume that the data $Y_1, \ldots, Y_n$ are i.i.d. draws from the population distribution with p.m.f. or p.d.f. of a certain form $f_Y()$ indexed by unknown parameters θ

What Are We Estimating?

- Our goal is to learn about the data generating process that generated the sample
- We might assume that the data $Y_1, \ldots, Y_n$ are i.i.d. draws from the population distribution with p.m.f. or p.d.f. of a certain form $f_Y()$ indexed by unknown parameters θ
- Even without a full probability model we can estimate particular properties of a distribution such as the mean $E[Y_i] = \mu$ or the variance $V[Y_i] = \sigma^2$
- An estimator $\hat{\theta}$ of some parameter θ, is a function of the sample $\hat{\theta} = h(Y_1, \ldots, Y_n)$ and thus is a random variable.

Why Study Estimators?

Why Study Estimators?

- Two Goals:
  1. Inference: How much uncertainty do we have in this estimate?

Why Study Estimators?

- Two Goals:
  1. Inference: How much uncertainty do we have in this estimate?
  2. Evaluate Estimators: How do we choose which estimator to use?

Why Study Estimators?

- 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 will consider the hypothetical sampling distribution of estimates we would have obtained if we had drawn repeated samples of size n from the population.

Why Study Estimators?

- 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 will consider the hypothetical sampling distribution of estimates we would have obtained if we had drawn repeated samples of size n from the population.
- In real applications, we cannot draw repeated samples, so we attempt to approximate the sampling distribution (either by resampling or by mathematical formulas)

Why Study Estimators?

- 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 will consider the hypothetical sampling distribution of estimates we would have obtained if we had drawn repeated samples of size n from the population.
- In real applications, we cannot draw repeated samples, so we attempt to approximate the sampling distribution (either by resampling or by mathematical formulas)

Sampling Distribution of the Sample Mean

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

Repeated Sampling Procedure:

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

Repeated Sampling Procedure:

1. Take a simple random sample of size $n = 4$.

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

Repeated Sampling Procedure:

1. Take a simple random sample of size $n = 4$.
2. Calculate the sample mean.

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

Repeated Sampling Procedure:

1. Take a simple random sample of size $n = 4$.
2. Calculate the sample mean.
3. Repeat steps 1 and 2 at least 10,000 times.

Sampling Distribution of the Sample Mean

Consider a toy example where we have the population and can construct sampling distributions.

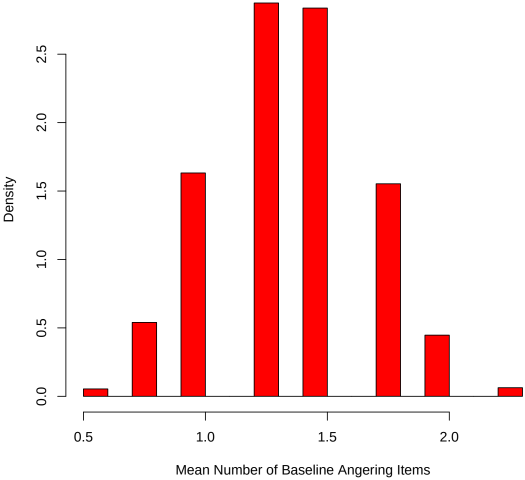
Repeated Sampling Procedure:

1. Take a simple random sample of size $n = 4$.
2. Calculate the sample mean.
3. Repeat steps 1 and 2 at least 10,000 times.
4. Plot the sampling distribution of the sample means (maybe as a histogram).

Repeated Sampling Procedure

```
# population data

ypop <- c(rep(0,0),rep(1,17),rep(2,10),rep(3,4))

# simulate the sampling distribution of the sample mean

SamDistMeans <- replicate(10000, mean(sample(ypop,size=4,repla
```

Sampling Distribution of the Sample Mean



Sampling Distribution of the Sample Standard Deviation

Sampling Distribution of the Sample Standard Deviation

We can consider sampling distributions for other sample statistics (e.g., the sample standard deviation).

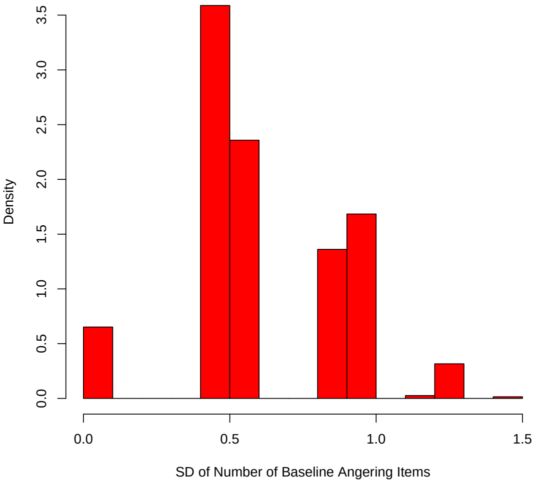
Sampling Distribution of the Sample Standard Deviation

We can consider sampling distributions for other sample statistics (e.g., the sample standard deviation).

Repeated Sampling Procedure:

1. Take a simple random sample of size $n = 4$.
2. Calculate the sample standard deviation.
3. Repeat steps 1 and 2 at least 10,000 times.
4. Plot the sampling distribution of the sample standard deviations (maybe as a histogram).

Sampling Distribution of the Sample Standard Deviation



Standard Error

We refer to the standard deviation of a sampling distribution as a standard error.

Standard Error

We refer to the standard deviation of a sampling distribution as a standard error.

Two Points of Potential Confusion:

- Each sampling distribution has its own standard deviation, and therefore its own standard error. (.35 for mean, .30 for sd)

Standard Error

We refer to the standard deviation of a sampling distribution as a standard error.

Two Points of Potential Confusion:

- Each sampling distribution has its own standard deviation, and therefore its own standard error. (.35 for mean, .30 for sd)
- Some people refer to an estimated standard error as the standard error.

Bootstrapped Sampling Distributions

In real applications, we cannot draw repeated samples from the population, so we must approximate the sampling distribution (either by resampling or by mathematical formulas).

Bootstrapped Sampling Distributions

In real applications, we cannot draw repeated samples from the population, so we must approximate the sampling distribution (either by resampling or by mathematical formulas).

The resampling procedure that we will use in this class is called bootstrapping (resamples with replacement of size n from the sample).

Bootstrapped Sampling Distributions

In real applications, we cannot draw repeated samples from the population, so we must approximate the sampling distribution (either by resampling or by mathematical formulas).

The resampling procedure that we will use in this class is called bootstrapping (resamples with replacement of size n from the sample).

Because we will not have actual sampling distributions, we cannot actually calculate standard errors (we can only approximate them).

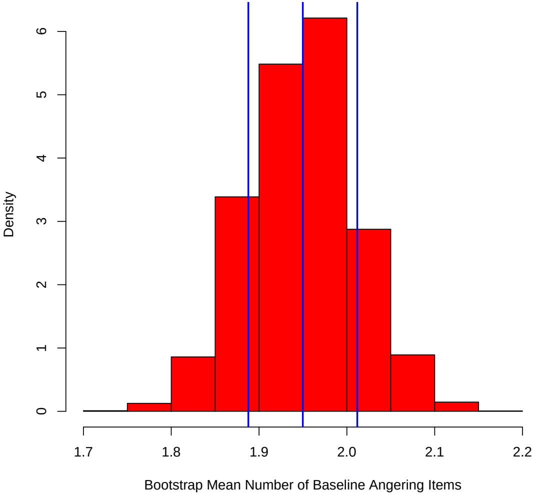
Southern Responses to Baseline List

# of angering items	0	1	2	3
# of responses	2	37	66	34

```
ysam <- c(rep(0,2),rep(1,37),rep(2,66),rep(3,34))

mean(ysam) # sample mean

# Bootstrapping
BootMeans <- replicate(
10000,
mean(sample(ysam,size=139,replace=TRUE))
)

sd(BootMeans) # estimated standard error
```

```
> ysam <- c(rep(0,2),rep(1,37),rep(2,66),rep(3,34))
>
> mean(ysam) # sample mean
[1] 1.949640
>
> BootMeans <- replicate(
+  10000,
+  mean(sample(ysam,size=n,replace=TRUE))
+  )
>
> sd(BootMeans) # estimated standard error
[1] 0.06234988
```

Bootstrapped Sampling Distribution of the Sample Mean



Independence for Random Variables

Independence for Random Variables

Recall:

Definition (Independence of Random Variables)

Two random variables Y and X are independent if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all x and y. We write this as $Y \perp\!\!\!\perp X$.

Independence for Random Variables

Recall:

Definition (Independence of Random Variables)

Two random variables Y and X are independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all x and y. We write this as $Y \perp\!\!\!\perp X$.

Independence implies

$$f_{Y|X}(y|x) = f_Y(y)$$

and thus

$$E[Y|X=x] = E[Y]$$

Notation for Sampling Distributions

Notation for Sampling Distributions

Suppose we took a simple random sample with replacement from the population.

Notation for Sampling Distributions

Suppose we took a simple random sample with replacement from the population.

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

Then we write $X_1, X_2, \ldots X_n \sim_{i.i.d} \quad ?(\mu, \sigma^2)$

Describing the Sampling Distribution for the Mean

We would like a full description of the sampling distribution for the mean, 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 $\overline{X}_n$.

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

Expectation of $\overline{X}_n$

Again, let $X_1, X_2, \ldots X_n$ be identically and independently distributed from a population distribution with a mean ($E[X_1] = \mu$) and a variance ($V[X_1] = \sigma^2$). Using the properties of expectation, calculate

$$E[\overline{X}_n] = E[\frac{1}{n}\sum_{i=1} X_i]$$
$$= ?$$

Variance of $\overline{X}_n$

Again, let $X_1, X_2, \ldots X_n$ be identically and independently distributed from a population distribution with a mean ($E[X_1] = \mu$) and a variance ($V[X_1] = \sigma^2$). Using the properties of variances, calculate

$$V[\overline{X}_n] = V[\frac{1}{n}\sum_{i=1}^{n} X_i]$$
$$= ?$$

Question:

What is the standard deviation of $\overline{X}_n$, also known as the standard error of $\overline{X}_n$?

A) $\frac{\sigma}{n}$
B) $\frac{\sigma}{n-1}$
C) $\frac{\sigma}{\sqrt{n}}$
D) $\frac{\sigma}{\sqrt{n-1}}$

What about the "?"

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

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

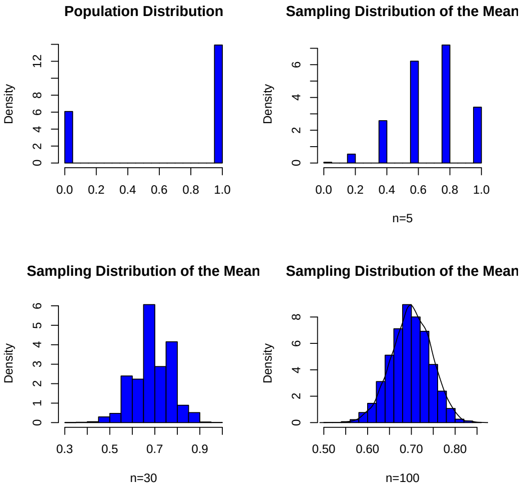
What about the "?"

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

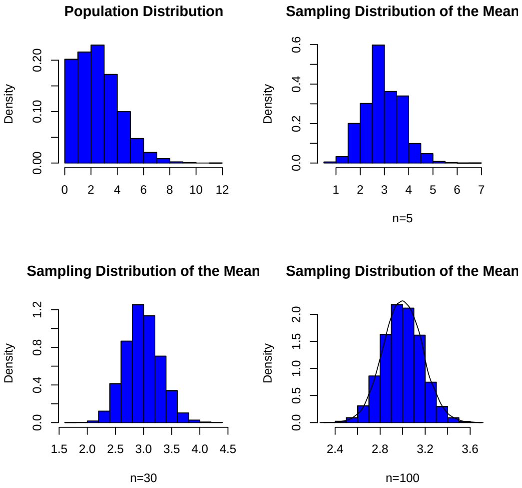
$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

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

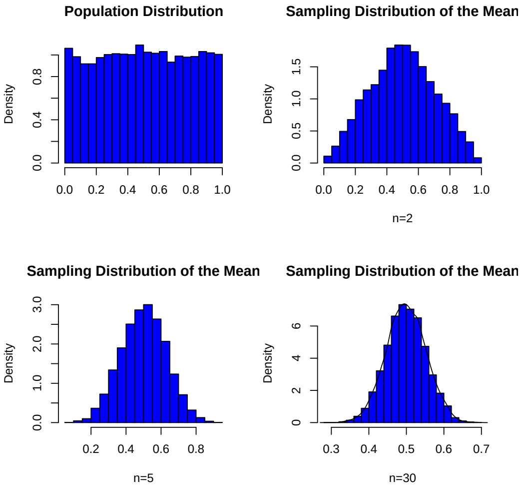
Bernoulli (Coin Flip) Distribution



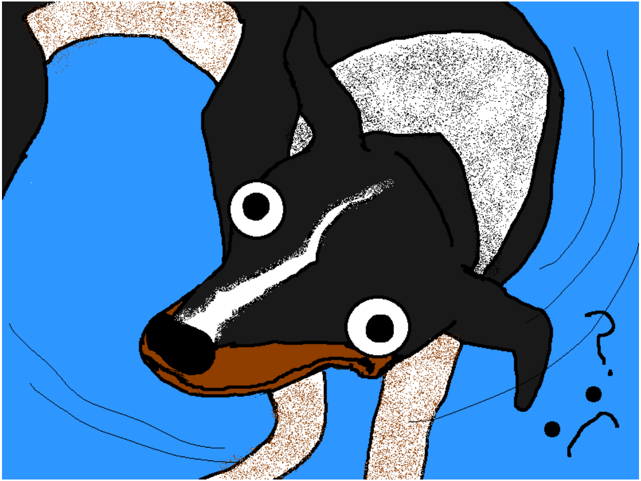
Poisson (Count) Distribution



Uniform Distribution

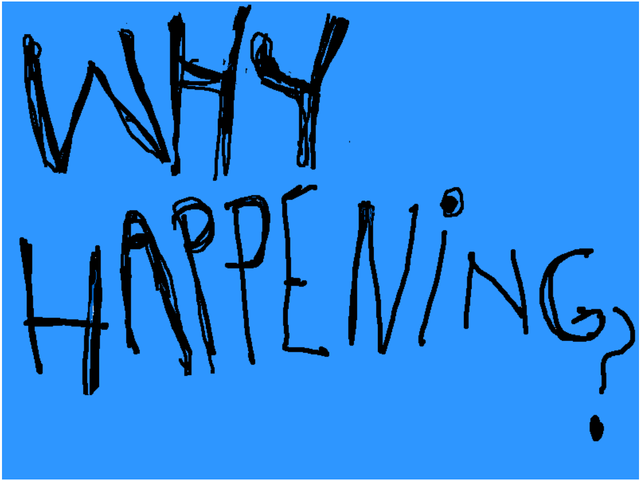


Why would this be true?



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

Why would this be true?



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

The Central Limit Theorem

The Central Limit Theorem

In the previous slides, as n increases, the sampling distribution of $\overline{X}_n$ appeared to become more bell-shaped.

The Central Limit Theorem

In the previous slides, as n increases, the sampling distribution of $\overline{X}_n$ appeared to become more bell-shaped. This is the basic implication of the Central Limit Theorem:

The Central Limit Theorem

In the previous slides, as n increases, the sampling distribution of $\overline{X}_n$ appeared to become more bell-shaped. This is the basic implication of the Central Limit Theorem:

If $X_1, \ldots, X_n \sim_{i.i.d.} ?(\mu, \sigma^2)$ and n is large, then

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

The Central Limit Theorem: What are we glossing over?

To understand the Central Limit Theorem mathematically we need a few basic definitions in place first.

The Central Limit Theorem: What are we glossing over?

To understand the Central Limit Theorem mathematically we need a few basic definitions in place first.

Definition (Convergence in Probability)

A sequence $X_1, ..., X_n$ of random variables converges in probability towards a real number a if, for all accuracy levels $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - a| \geq \varepsilon\right) = 0$$

We write this as

$$X_n \overset{p}{\to} a \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} X_n = a.$$

The Central Limit Theorem: What are we glossing over?

The Central Limit Theorem: What are we glossing over?

Definition (Law of Large Numbers)

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables, each with finite mean μ. Then for all $\varepsilon > 0$,

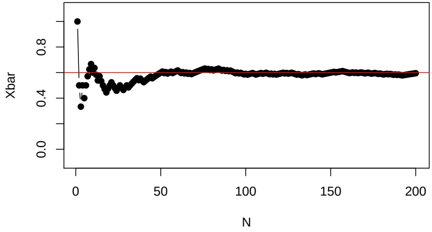
$$\overline{X}_n \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty$$

or equivalently,

$$\lim_{n \to \infty} \Pr\left(|\overline{X}_n - \mu| \geq \varepsilon \right) = 0$$

where $\overline{X}_n$ is the sample mean.

Example: Mean of N independent tosses of a coin:



The Central Limit Theorem: What are we glossing over?

The Central Limit Theorem: What are we glossing over?

Definition (Convergence in Distribution)

Consider a sequence of random variables $X_1, ..., X_n$, each with CDFs $F_1, ..., F_n$. The sequence is said to converge in distribution to a limiting random variable X with CDF F if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every point x at which F is continuous. We write this as

$$X_N \stackrel{d}{\to} X.$$

The Central Limit Theorem: What are we glossing over?

Definition (Convergence in Distribution)

Consider a sequence of random variables $X_1, ..., X_n$, each with CDFs $F_1, ..., F_n$. The sequence is said to converge in distribution to a limiting random variable X with CDF F if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every point x at which F is continuous. We write this as

$$X_N \stackrel{d}{\to} X.$$

- As n grows, the distribution of X_n converges to the distribution of X.

The Central Limit Theorem: What are we glossing over?

Definition (Convergence in Distribution)

Consider a sequence of random variables $X_1, ..., X_n$, each with CDFs $F_1, ..., F_n$. The sequence is said to converge in distribution to a limiting random variable X with CDF F if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every point x at which F is continuous. We write this as

$$X_N \xrightarrow{d} X.$$

- As n grows, the distribution of X_n converges to the distribution of X.
- Convergence in probability is a special case of convergence in distribution in which the distribution converges to a degenerate distribution (i.e. a probability distribution which only takes a single value).

The Central Limit Theorem: What are we glossing over?

The Central Limit Theorem: What are we glossing over?

Definition (Lindeberg-Lévy Central Limit Theorem)

Let $X_1, ..., X_n$ a sequence of i.i.d. random variables each with mean μ and variance $\sigma^2 < \infty$. Then, for *any* population distribution of X,

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

- As n grows, the $\sqrt{n}$-scaled sample mean converges to a normal random variable.

The Central Limit Theorem: What are we glossing over?

Definition (Lindeberg-Lévy Central Limit Theorem)

Let $X_1, ..., X_n$ a sequence of i.i.d. random variables each with mean μ and variance $\sigma^2 < \infty$. Then, for *any* population distribution of X,

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

- As n grows, the $\sqrt{n}$-scaled sample mean converges to a normal random variable.
- CLT also implies that the standardized sample mean converges to a standard normal random variable:

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

The Central Limit Theorem: What are we glossing over?

Definition (Lindeberg-Lévy Central Limit Theorem)

Let $X_1, \ldots, X_n$ a sequence of i.i.d. random variables each with mean μ and variance $\sigma^2 < \infty$. Then, for *any* population distribution of X,

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

- As n grows, the $\sqrt{n}$-scaled sample mean converges to a normal random variable.
- CLT also implies that the standardized sample mean converges to a standard normal random variable:

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

- Note that CLT holds for a random sample from *any* population distribution (with finite mean and variance) — what a convenient result!

Question:

Question:

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

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.

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.

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.

Point Estimation

Point Estimation

Suppose we are primarily interested in specific characteristics of the population distribution.

Point Estimation

Suppose we are primarily interested in specific characteristics of the population distribution.

For example, suppose we are primarily interested in $E[X]$.

Point Estimation

Suppose we are primarily interested in specific characteristics of the population distribution.

For example, suppose we are primarily interested in $E[X]$.

We refer to characteristics of the population distribution (e.g., $E[X]$) as parameters. These are often denoted with a greek letter (e.g. μ).

Point Estimation

Suppose we are primarily interested in specific characteristics of the population distribution.

For example, suppose we are primarily interested in $E[X]$.

We refer to characteristics of the population distribution (e.g., $E[X]$) as parameters. These are often denoted with a greek letter (e.g. μ).

We use a statistic (e.g., $\overline{X}$) to estimate a parameter, and we will denote this with a hat (e.g. $\hat{\mu}$). A statistic is a function of the sample.

Why Point Estimation?

Why Point Estimation?

- Estimating one number is typically easier than estimating many (or an infinite number of) numbers.

Why Point Estimation?

- Estimating one number is typically easier than estimating many (or an infinite number of) numbers.
- The question of interest may be answerable with single characteristic of the distribution (e.g., if $E[Y] - E[X]$ identifies the proportion angered by the sensitive item, then it may be sufficient to estimate $E[Y]$ and $E[X]$)

Estimators for μ

Estimators for μ

Some possible estimators $\widehat{\mu}$ for the balance point μ:

Estimators for μ

Some possible estimators $\widehat{\mu}$ for the balance point μ:

A) $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$, the sample average

Estimators for μ

Some possible estimators $\widehat{\mu}$ for the balance point μ:

A) $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$, the sample average

B) $\tilde{X}_n = \text{median}(X_1, \ldots, X_n)$, the sample median

Estimators for μ

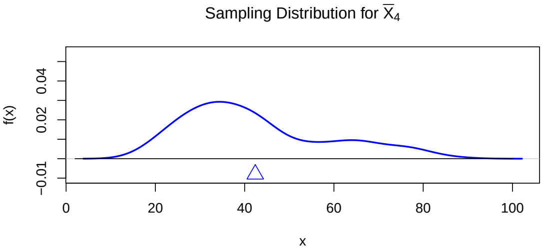
Some possible estimators $\widehat{\mu}$ for the balance point μ:

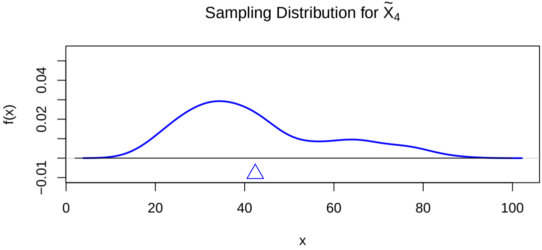
A) $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$, the sample average

B) $\tilde{X}_n = \text{median}(X_1, \ldots, X_n)$, the sample median

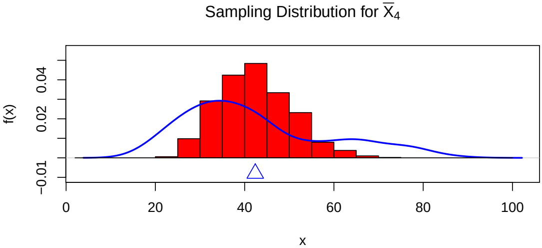
Clearly, one of these estimators is better than the other, but how can we define "better"?

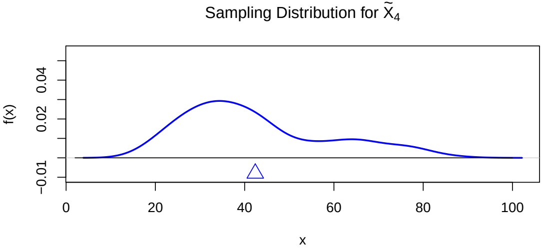
Age population distribution in blue



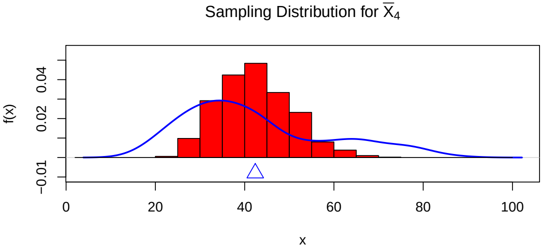


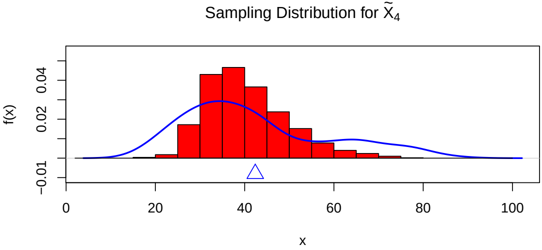
Age population distribution in blue, sampling distributions in red





Age population distribution in blue, sampling distributions in red





Methods of Finding Estimators

Methods of Finding Estimators

We will primarily discuss the Method of Least Squares for finding estimators in this course. However, many of the estimators we discuss can also be derived by Method of Moments or Method of Maximum Likelihood (covered in Soc504).

Methods of Finding Estimators

We will primarily discuss the Method of Least Squares for finding estimators in this course. However, many of the estimators we discuss can also be derived by Method of Moments or Method of Maximum Likelihood (covered in Soc504).

When estimating simple features of a distribution we can use the plug-in principle, the idea that you write down the feature of the distribution you are interested in and estimate with the sample analog. Formally this is using the Empirical CDF to estimate features of the population.

Desirable Properties of Estimators

Desirable Properties of Estimators

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

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.

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.

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.

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

Properties of Estimators

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

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.

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.

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):

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? $E[\hat{\mu}] = \mu$

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? $V[\hat{\mu}_1] < V[\hat{\mu}_2]$

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):

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?

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)

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.

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.

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

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,

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

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 iff:

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

Example: Estimators for Population Mean

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

Bias of Example Estimators

Which of these estimators are unbiased?

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

Bias of Example Estimators

Which of these estimators are unbiased?

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

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. $E[42 - \mu] =$
4. $E[\overline{Y}_n - \mu] =$

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. $E[\overline{Y}_n - \mu] =$

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. $E[\overline{Y}_n - \mu] = \frac{1}{n}\sum_1^n E[Y_i] - \mu$

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. $E[\overline{Y}_n - \mu] = \frac{1}{n}\sum_1^n E[Y_i] - \mu = \mu - \mu = 0$

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.

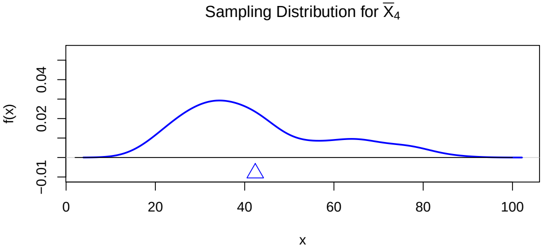
Bias of Example Estimators

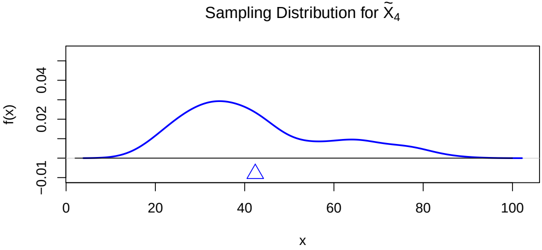
Which of these estimators are unbiased?

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. $E[\overline{Y}_n - \mu] = \frac{1}{n}\sum_1^n E[Y_i] - \mu = \mu - \mu = 0$

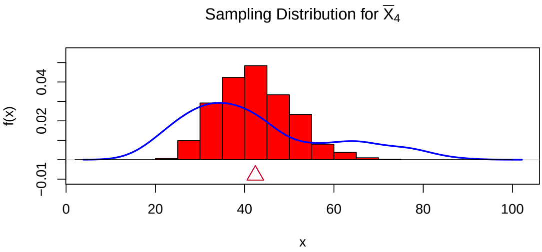
- Estimators 1,2, and 4 are unbiased because they get the right answer on average.
- Estimator 3 is biased.

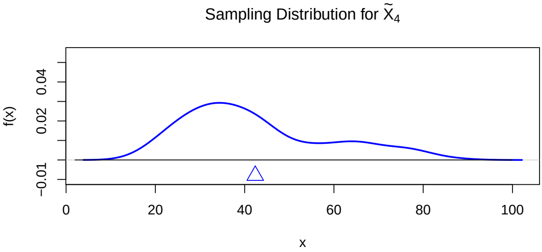
Age population distribution in blue



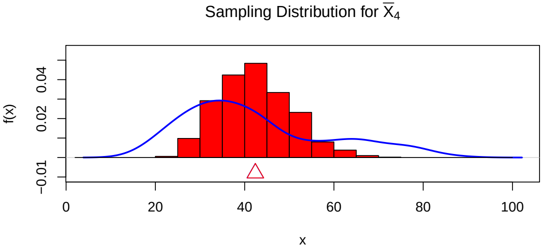


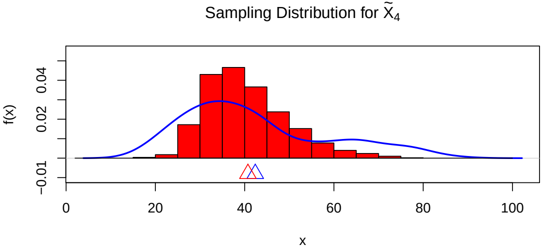
Age population distribution in blue, sampling distributions in red





Age population distribution in blue, sampling distributions in red





2: Efficiency (doesn't change much sample to sample)

- How should we choose among unbiased estimators?

2: Efficiency (doesn't change much sample to sample)

- How should we choose among unbiased estimators?
- All else equal, we prefer estimators that have a sampling distribution with smaller variance.

2: Efficiency (doesn't change much sample to sample)

- How should we choose among unbiased estimators?
- 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 two 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

2: Efficiency (doesn't change much sample to sample)

- How should we choose among unbiased estimators?
- 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 two 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

2: Efficiency (doesn't change much sample to sample)

- How should we choose among unbiased estimators?
- 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 two 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

Variance of Example Estimators

What is the variance of our estimators?

1. $V[Y_1] =$
2. $V[\frac{1}{2}(Y_1 + Y_n)] =$
3. $V[42] =$
4. $V[\overline{Y}_n] =$

Variance of Example Estimators

What is the variance of our estimators?

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

Variance of Example Estimators

What is the variance of our estimators?

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

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. $V[42] =$
4. $V[\overline{Y}_n] =$

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. $V[\overline{Y}_n] =$

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. $V[\overline{Y}_n] = \frac{1}{n^2}\sum_1^n V[Y_i]$

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. $V[\overline{Y}_n] = \frac{1}{n^2}\sum_1^n V[Y_i] = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2$

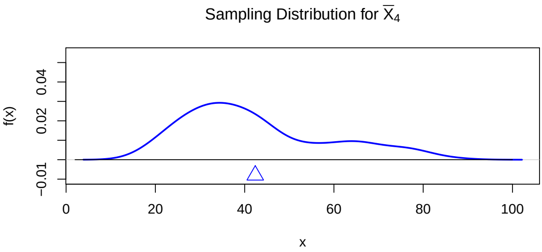
Variance of Example Estimators

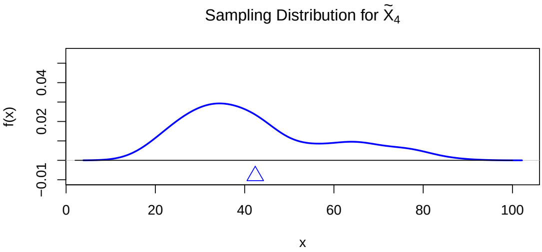
What is the variance of our estimators?

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. $V[\overline{Y}_n] = \frac{1}{n^2}\sum_1^n V[Y_i] = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2$

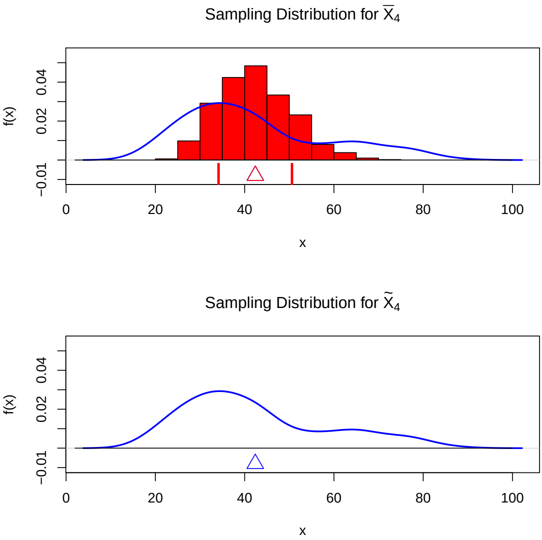
Among the unbiased estimators, the sample average has the smallest variance. This means that Estimator 4 (the sample average) is likely to be closer to the true value μ, than Estimators 1 and 2.

Age population distribution in blue

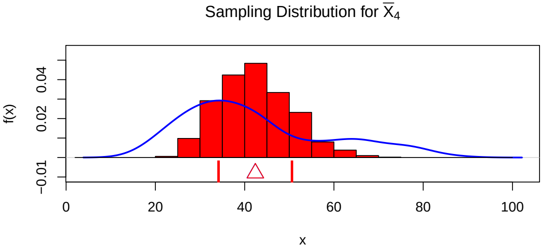


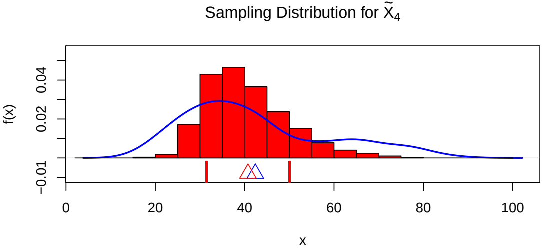


Age population distribution in blue, sampling distributions in red



Age population distribution in blue, sampling distributions in red





3-4: Asymptotic Evaluations (what happens as sample size increases)

3-4: Asymptotic Evaluations (what happens as sample size increases)

- Unbiasedness and efficiency are finite-sample properties of estimators, which hold regardless of sample size

3-4: Asymptotic Evaluations (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

3-4: Asymptotic Evaluations (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:

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

3-4: Asymptotic Evaluations (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:

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

3-4: Asymptotic Evaluations (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.

Stochastic Convergence

- When a sequence of random variables stabilizes to a certain probabilistic behavior as $n \to \infty$, the sequence is said to show stochastic convergence.

Stochastic Convergence

- When a sequence of random variables stabilizes to a certain probabilistic behavior as $n \to \infty$, the sequence is said to show stochastic convergence.
- Two types of stochastic convergence are of particular importance:
  1. Convergence in probability: values in the sequence eventually take a constant value
     (i.e. the limiting distribution is a point mass)
  2. 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)

Convergence in Probability

Definition (Convergence in Probability)

A sequence $X_1, ..., X_n$ of random variables converges in probability towards a real number a if, for all accuracy levels $\varepsilon > 0$,

$$\lim_{n \to \infty} \Pr\left(|X_n - a| \geq \varepsilon\right) = 0$$

We write this as

$$X_n \xrightarrow{p} a \quad \text{or} \quad \operatorname*{plim}_{n \to \infty} X_n = a.$$

- As n increases, almost all of the PDF/PMF of X_n will be concentrated in the ε-interval around a, $[a - \varepsilon, a + \varepsilon]$

Convergence in Probability

Definition (Convergence in Probability)

A sequence $X_1, ..., X_n$ of random variables converges in probability towards a real number a if, for all accuracy levels $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - a| \geq \varepsilon\right) = 0$$

We write this as

$$X_n \xrightarrow{p} a \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} X_n = a.$$

- As n increases, almost all of the PDF/PMF of X_n will be concentrated in the ε-interval around a, $[a - \varepsilon, a + \varepsilon]$
- A sufficient (but not necessary) condition for convergence in probability:

$$E[X_n] \to a \quad \text{and} \quad V[X_n] \to 0 \quad \text{as} \quad n \to \infty$$

Convergence in Probability

Definition (Convergence in Probability)

A sequence $X_1, ..., X_n$ of random variables converges in probability towards a real number a if, for all accuracy levels $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|X_n - a| \geq \varepsilon\right) = 0$$

We write this as

$$X_n \xrightarrow{p} a \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} X_n = a.$$

- As n increases, almost all of the PDF/PMF of X_n will be concentrated in the ε-interval around a, $[a - \varepsilon, a + \varepsilon]$
- A sufficient (but not necessary) condition for convergence in probability:

$$E[X_n] \to a \quad \text{and} \quad V[X_n] \to 0 \quad \text{as} \quad n \to \infty$$

- For example, the sample mean $\bar{X}_n$ converges to the population mean μ in probability because

$$E[\bar{X}_n] = \mu \quad \text{and} \quad V[\bar{X}_n] = \sigma^2/n \to 0 \quad \text{as} \quad n \to \infty$$

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

Definition

An estimator θ_n is consistent if the sequence $\theta_1, ..., \theta_n$ converges in probability to the true parameter value θ as sample size n grows to infinity:

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

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

Definition

An estimator θ_n is consistent if the sequence $\theta_1, ..., \theta_n$ converges in probability to the true parameter value θ as sample size n grows to infinity:

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

- Often seen as a minimal requirement for estimators
- A consistent estimator may still perform badly in small samples

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

Definition

An estimator θ_n is consistent if the sequence $\theta_1, ..., \theta_n$ converges in probability to the true parameter value θ as sample size n grows to infinity:

$$\theta_n \xrightarrow{p} \theta \quad \text{or} \quad \text{plim}_{n\to\infty} \theta_n = \theta$$

- 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 $E[\theta_n] \to \theta$ and $V[\theta_n] \to 0$
  2. Simulation: Increase n and see how the sampling distribution changes

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

Definition

An estimator θ_n is consistent if the sequence $\theta_1, ..., \theta_n$ converges in probability to the true parameter value θ as sample size n grows to infinity:

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

- 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 $E[\theta_n] \to \theta$ and $V[\theta_n] \to 0$
  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?

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 $V[\widehat{\mu}_1] = \sigma^2$

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 $V[\widehat{\mu}_2] = 0$

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 $V[\widehat{\mu}_3] = \frac{1}{n}\sigma^2$

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.

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

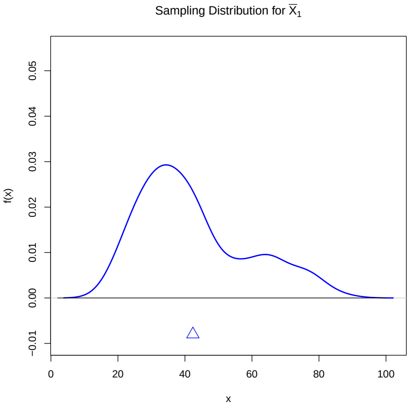
Consistency

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.

$$n =$$



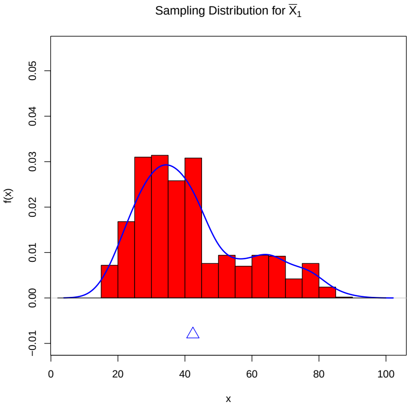
Consistency

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.

$$n = 1$$



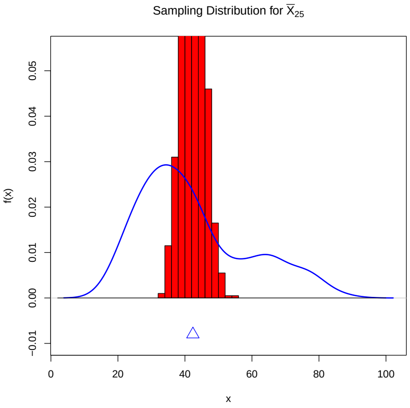
Consistency

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.

$$n = 25$$



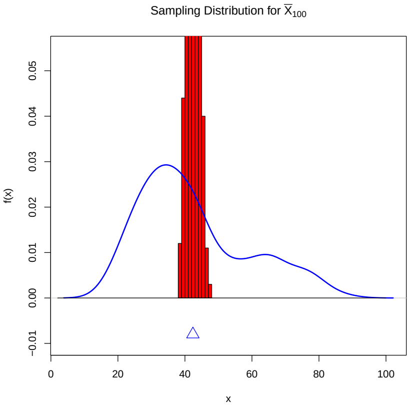
Consistency

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.

$$n = 100$$



Inconsistency

Inconsistency

An estimator can be inconsistent in several ways:

Inconsistency

An estimator can be inconsistent in several ways:

- The sampling distribution collapses around the wrong value

Inconsistency

An estimator can be inconsistent in several ways:

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

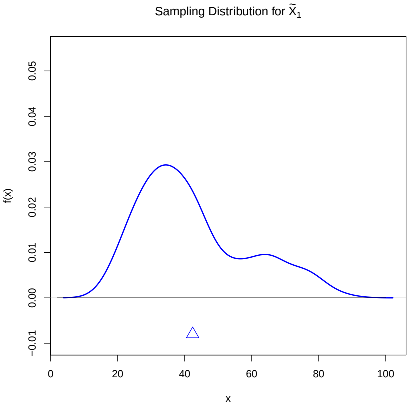
Inconsistency

Consider the median estimator: $\tilde{X}_n =$ median$(Y_1, ..., Y_n)$

Inconsistency

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

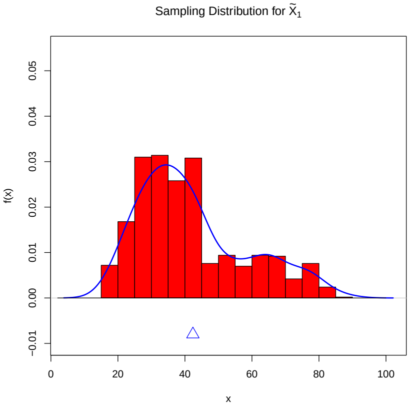
$n =$



Inconsistency

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

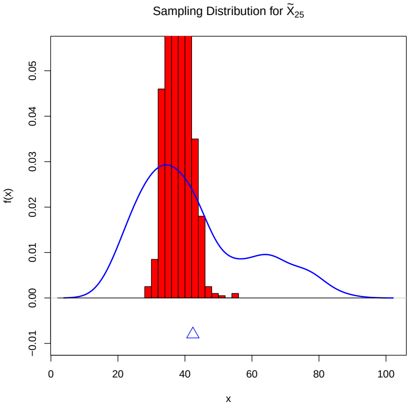
$n = 1$



Inconsistency

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

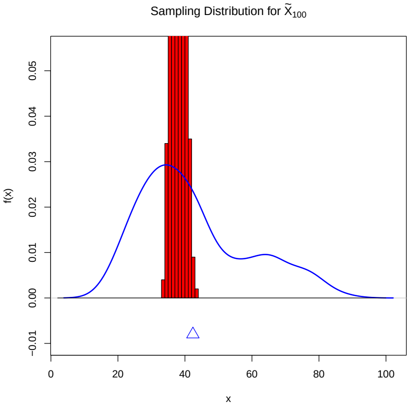
$n = 25$



Inconsistency

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

$n = 100$



4: Asymptotic Distribution (known sampling distribution for large sample size)

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.

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.

The sampling distributions of many estimators converge towards a normal distribution.

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.

The sampling distributions of many estimators converge towards a normal distribution.

For example, we've seen that the sampling distribution of the sample mean converges to the normal distribution.

Mean Squared Error

How can we choose between an unbiased estimator and a biased, but more efficient estimator?

Mean Squared Error

How can we choose between an unbiased estimator and a biased, but more efficient 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:

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

Review and Example

Gerber, Green, and Larimer (*American Political Science Review*, 2008)

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY—VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Basic Analysis

Basic Analysis

```
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
```

Basic Analysis

```
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
```

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

Basic Analysis

```
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?

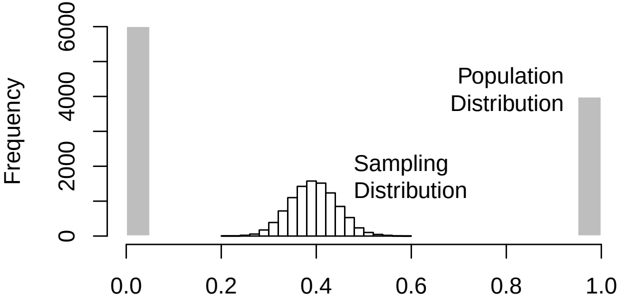
Population vs. Sampling Distribution

Population vs. Sampling Distribution

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

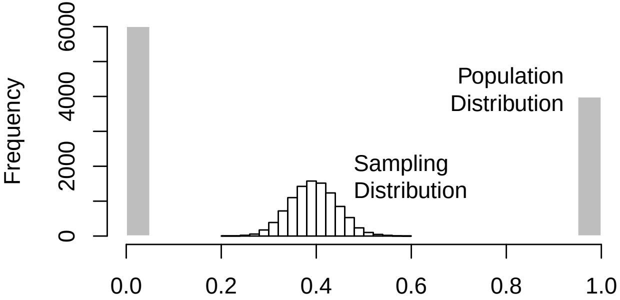
Population vs. Sampling Distribution

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



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

Asymptotic Normality

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

Asymptotic Normality

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

- The estimator is difference in means

Asymptotic Normality

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

- The estimator is difference in means
- The estimate is 0.063

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 $\hat{\text{SE}}(\hat{\theta}) = 0.02$.

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 $(\mu_y - \mu_x = 0)$?

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 $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$

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 $\hat{\theta} \sim \mathcal{N}(0, \text{SE}(\hat{\theta}))$

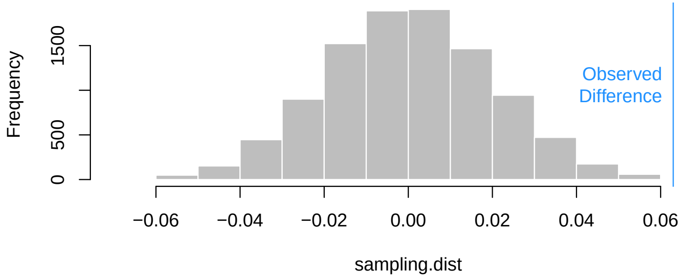
Asymptotic Normality

Asymptotic Normality

We can plot this to get a feel for it.

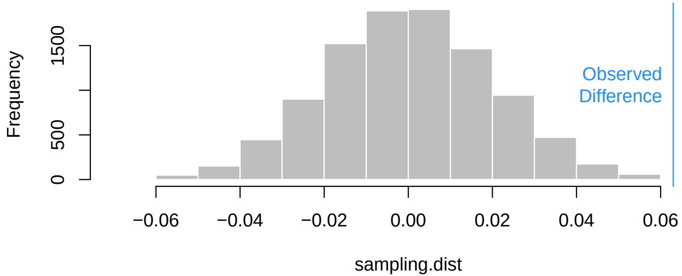
Asymptotic Normality

We can plot this to get a feel for it.



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?

Summary of Today

Summary of Today

- Sampling distributions provide away for studying the properties of sample statistics.

Summary of Today

- Sampling distributions provide away for studying the properties of sample statistics.
- We must usually make assumptions and/or appeal to a large *n* in order to derive a sampling distribution.

Summary of Today

- Sampling distributions provide away for studying the properties of sample statistics.
- We must usually make assumptions and/or appeal to a large *n* in order to derive a sampling distribution.
- Choosing a point estimator may require tradeoffs between desirable properties.

Summary of Today

- Sampling distributions provide away for studying the properties of sample statistics.
- We must usually make assumptions and/or appeal to a large n in order to derive a sampling distribution.
- Choosing a point estimator may require tradeoffs between desirable properties.

Next Class:

Summary of Today

- Sampling distributions provide away for studying the properties of sample statistics.
- We must usually make assumptions and/or appeal to a large n in order to derive a sampling distribution.
- Choosing a point estimator may require tradeoffs between desirable properties.

Next Class: interval estimation

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

Fun with Hidden Populations

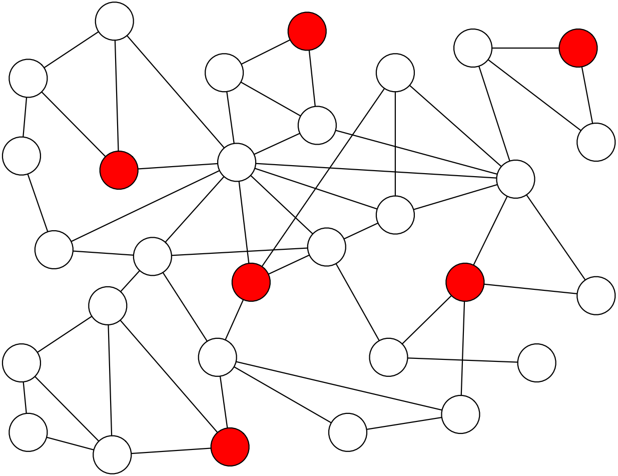
Dennis M. Feehan and Matthew J. Salganik "Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations"

Slides graciously provided by Matt Salganik.

Scale-up Estimator

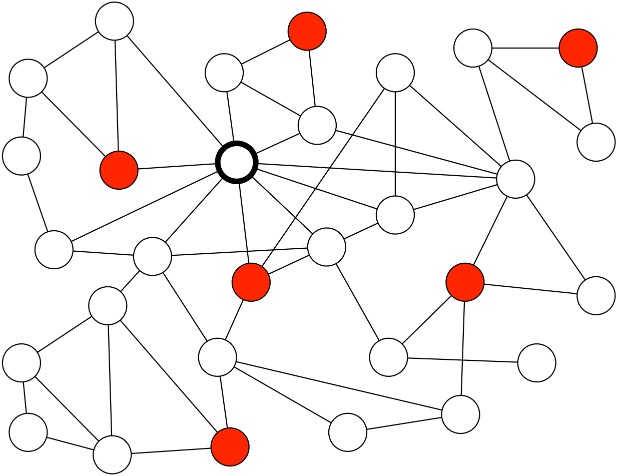
Basic insight from Bernard et al. (1989)

Network scale-up method



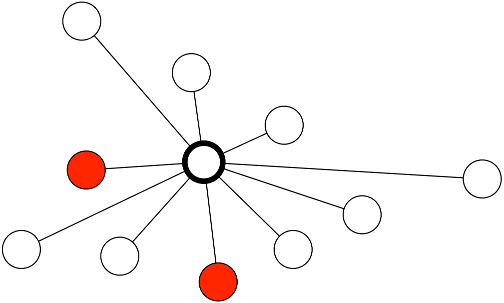
$$\hat{N}_T = \frac{\sum_i y_{i,T}}{\sum_i \hat{d}_i} \times N$$

Network scale-up method



$$\hat{N}_T = \frac{\sum_i y_{i,T}}{\sum_i \hat{d}_i} \times N$$

Network scale-up method



$$\hat{N}_T = \frac{2}{10} \times 30 = 6$$

If $\underbrace{y_{i,k} \sim Bin(d_i, N_k/N)}_{\text{basic scale-up model}}$, then maximum likelihood estimator is

$$\hat{N}_T = \frac{\sum_i y_{i,T}}{\sum_i \hat{d}_i} \times N$$

- $\hat{N}_T$: number of people in the target population
- $y_{i,T}$: number of people in target population known by person i
- $\hat{d}_i$: estimated number of people known by person i
- N: number of people in the population

See Killworth et al., (1998)

Target population	Location	Citation
Mortality in earthquake	Mexico City, Mexico	Bernard et al. (1989)
Rape victims	Mexico City, Mexico	Bernard et al. (1991)
HIV prevalence, rape, & homelessness	U.S.	Killworth et al. (1998)
Heroin use	14 U.S. cities	Kadushin et al. (2006)
Choking incidents in children	Italy	Snidero et al. (2007, 2009)
Groups most at-risk for HIV/AIDS	Ukraine	Paniotto et al. (2009)
Heavy drug users	Curitiba, Brazil	Salganik et al. (2011)
Men who have sex with men	Japan	Ezoe et al. (2012)
Groups most at risk for HIV/AIDS	Almaty, Kazakhstan	Scutelniciuc (2012a)
Groups most at risk for HIV/AIDS	Moldova	Scutelniciuc (2012b)
Groups most at risk for HIV/AIDS	Thailand	Aramrattan (2012)
Groups most at risk for HIV/AIDS	Chongqing, China	Guo (2012)
Groups most at risk for HIV/AIDS	Rwanda	Rwanda Biomedical Center (2012)

Does it Work? Under What Conditions?

Does it Work? Under What Conditions?

- Feehan and Salganik study the properties of the estimator

Does it Work? Under What Conditions?

- Feehan and Salganik study the properties of the estimator
- They show that for the estimator to be unbiased and consistent requires a particular assumption that average personal network size is the same in the hidden population as the remainder.

Does it Work? Under What Conditions?

- Feehan and Salganik study the properties of the estimator
- They show that for the estimator to be unbiased and consistent requires a particular assumption that average personal network size is the same in the hidden population as the remainder.
- This was unknown up to this point!

Does it Work? Under What Conditions?

- Feehan and Salganik study the properties of the estimator
- They show that for the estimator to be unbiased and consistent requires a particular assumption that average personal network size is the same in the hidden population as the remainder.
- This was unknown up to this point!
- Analyzing the estimator let them see that the problem can be addressed by collecting a new kind of data on the visibility of hidden population (which can easily be collected with respondent driven sampling)

Meta points

Meta points

- Studying estimators can not only expose problems but suggest solutions

Meta points

- Studying estimators can not only expose problems but suggest solutions
- Another example of creative and interesting ideas coming from the applied people

Meta points

- Studying estimators can not only expose problems but suggest solutions
- Another example of creative and interesting ideas coming from the applied people
- Formalizing methods is important because it is what allows them to be studied- it was a long time before anyone discovered the bias/consistency concerns!

References

- Kuklinski et al. 1997 “Racial prejudice and attitudes toward affirmative action” *American Journal of Political Science*
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102.01 (2008): 33-48.

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

- Last Week

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

- Last Week
  - random variables

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

- Last Week
  - random variables
  - joint distributions

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

- Last Week
  - random variables
  - joint distributions
- This Week

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - ★ sampling and sampling distributions
    - ★ point estimates
    - ★ properties (bias, variance, consistency)

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - ★ sampling and sampling distributions
    - ★ point estimates
    - ★ properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - ★ confidence intervals

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?

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

- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run

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

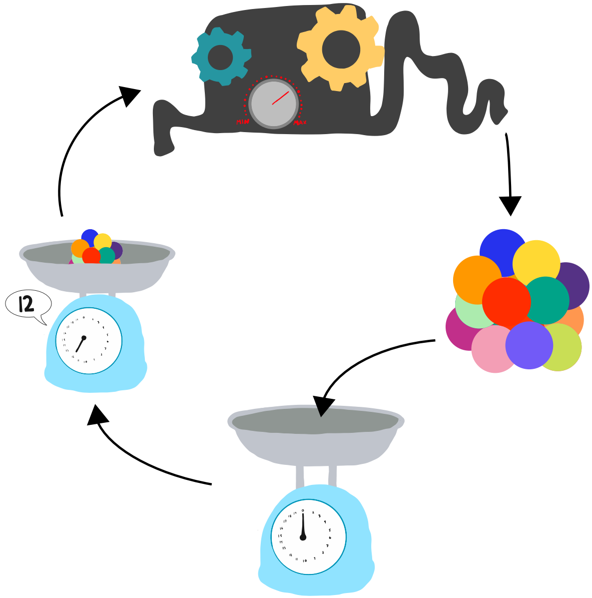
- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

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

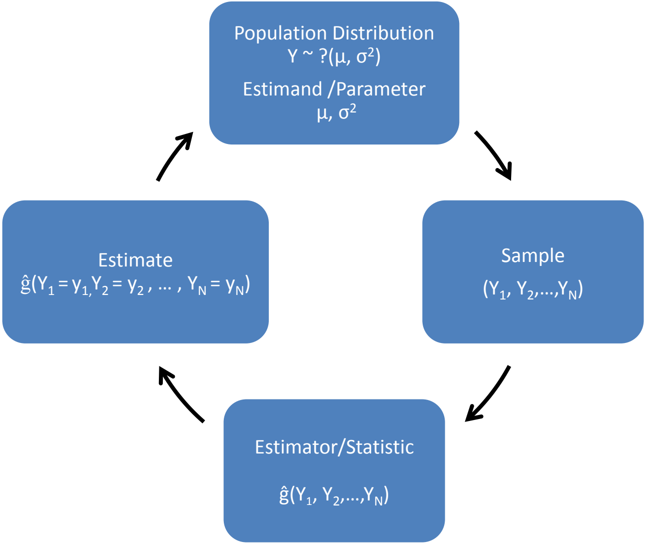
- Last Week
  - random variables
  - joint distributions
- This Week
  - Monday: Point Estimation
    - sampling and sampling distributions
    - point estimates
    - properties (bias, variance, consistency)
  - Wednesday: Interval Estimation
    - confidence intervals
    - comparing two groups
- Next Week
  - hypothesis testing
  - what is regression?
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?

Last Time



Last Time



Population Distribution
Y ~ ?(μ, σ²)
Estimand /Parameter
μ, σ²
Sample
(Y₁, Y₂,…,Y_N)
Estimator/Statistic
ĝ(Y₁, Y₂,…,Y_N)
Estimate
ĝ(Y₁ = y₁, Y₂ = y₂ , … , Y_N = y_N)

What is Interval Estimation?

- A point estimator $\hat{\theta}$ estimates a scalar population parameter θ with a single number.

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.

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.

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.

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.

Normal Population with Known σ^2

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

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

$$\overline{X}_n \sim$$

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

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

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

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

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

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

Therefore, the standardized sample average is distributed as follows:

$$\frac{\overline{X}_n - \mu}{1/\sqrt{n}} \; \sim$$

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

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

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

Therefore, the standardized sample average is distributed as follows:

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

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

From previous lecture, we know that the sampling distribution of the sample average 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

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

Normal Population with Known σ^2

Suppose we have an i.i.d. random sample of size n, $X_1, ..., X_n$, from $X \sim N(\mu, 1)$.

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

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

Therefore, the standardized sample average is distributed as follows:

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

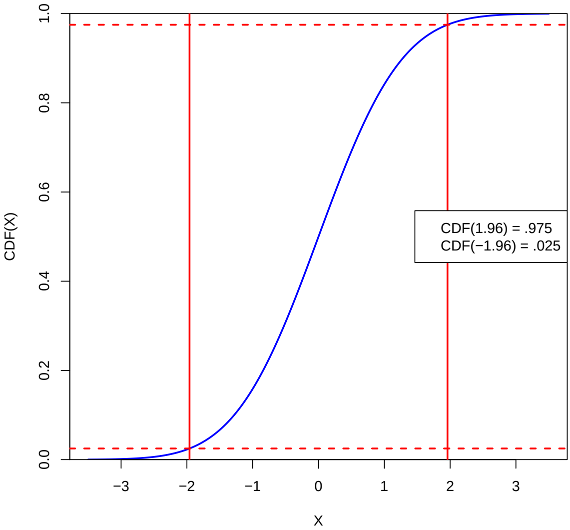
This implies

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

Why?

CDF of the Standard Normal Distribution

CDF of the Standard Normal Distribution



Constructing a Confidence Interval with Known σ^2

So we know that:

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

Constructing a Confidence Interval with Known σ^2

So we know that:

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

Rearranging yields:

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

Constructing a Confidence Interval with Known σ^2

So we know that:

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

Rearranging yields:

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

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.

Constructing a Confidence Interval with Known σ^2

So we know that:

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

Rearranging yields:

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

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

$$\overline{Y} \sim_{approx} ?(?, ?)$$

Kuklinski Example

$$\overline{Y} \sim_{approx} ?(\mu, ?)$$

Kuklinski Example

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

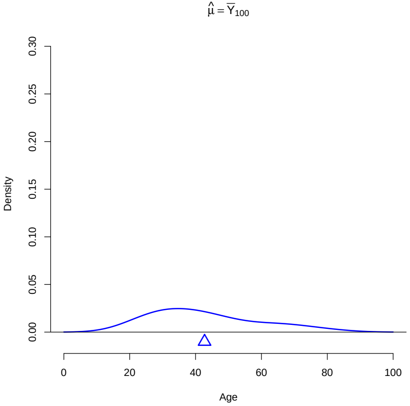
Kuklinski Example

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

Kuklinski Example

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

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



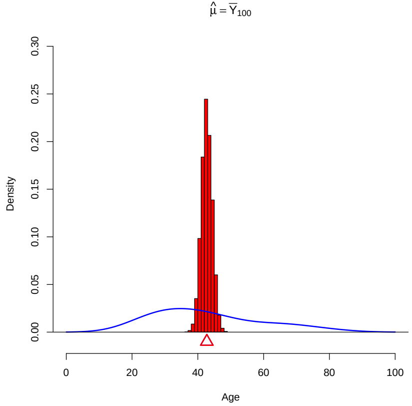
Kuklinski Example

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

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:

$$\overline{Y}_{100} \sim_{approx} ?(?,?)$$



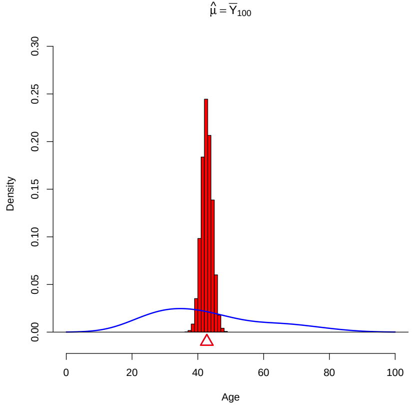
Kuklinski Example

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

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:

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



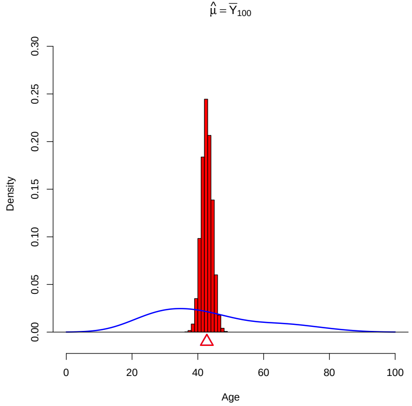
Kuklinski Example

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

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:

$$\overline{Y}_{100} \sim_{approx} ?(42.7, 257.9/100)$$



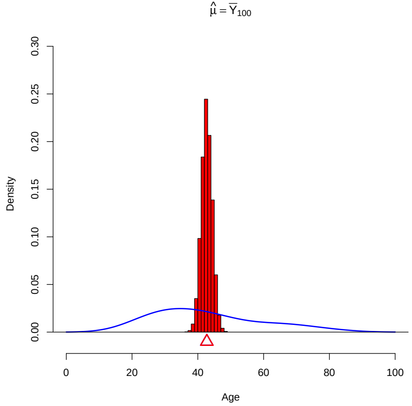
Kuklinski Example

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

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 $\overline{Y}$

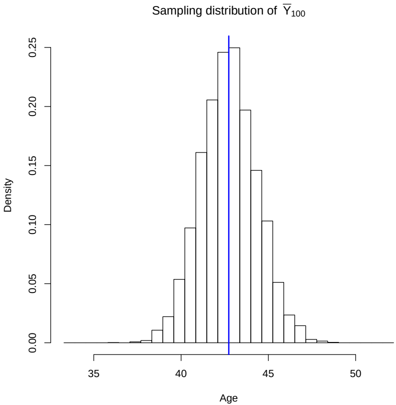
The standard error of the sample mean is the standard deviation of the sampling distribution for $\overline{Y}$:

$$SE(\overline{Y}) =$$

The standard error of $\overline{Y}$

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

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

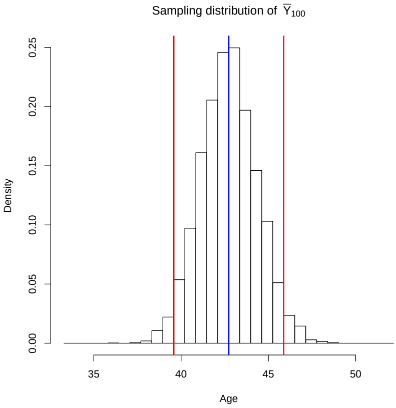


The standard error of $\overline{Y}$

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?

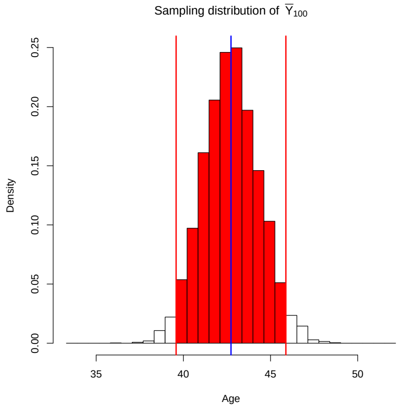


The standard error of $\overline{Y}$

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?



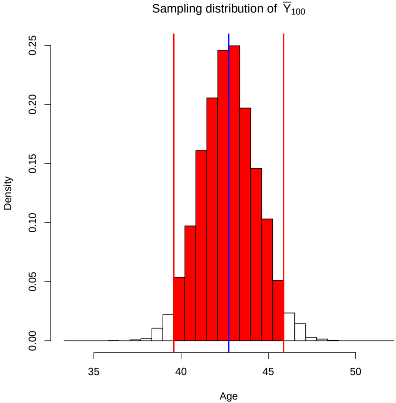
The standard error of $\overline{Y}$

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?

But the 1,161 is actually the sample (not the population).



Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, where σ^2 is unknown.

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, where σ^2 is unknown. Then, as before,

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

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, 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:

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

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, 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

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, 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.

Normal Population with Unknown σ^2

In practice, it is rarely the case that we somehow know the true value of σ^2.

Suppose now that we have an i.i.d. random sample of size n $X_1, ..., X_n$ from $X \sim N(\mu, \sigma^2)$, 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:

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?

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.

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

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

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

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

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

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

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:

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

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

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

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

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

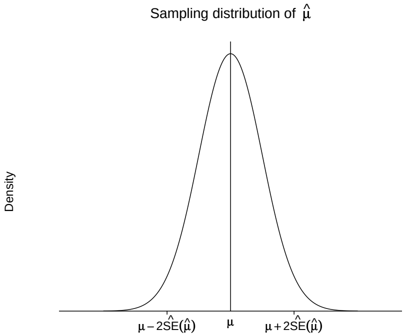
We will plug in S for σ and our estimated standard error will be

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

95% Confidence Intervals

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

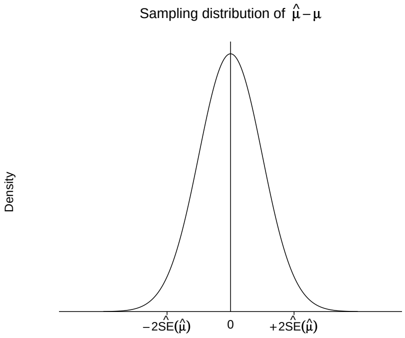
$$\widehat{\mu} \quad \sim \quad N(\mu, (\widehat{SE}[\hat{\mu}])^2)$$



95% Confidence Intervals

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

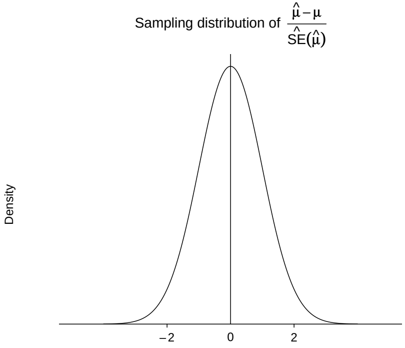
$$\begin{aligned}\widehat{\mu} &\sim N(\mu, (\widehat{SE}[\hat{\mu}])^2)\\ \widehat{\mu} - \mu &\sim N(0, (\widehat{SE}[\hat{\mu}])^2)\end{aligned}$$



95% Confidence Intervals

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

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



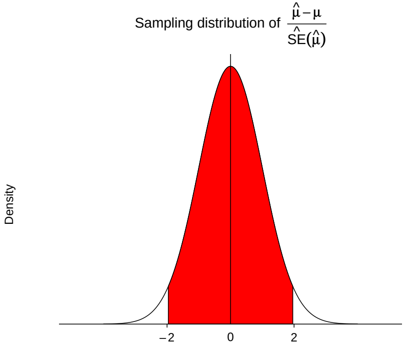
95% Confidence Intervals

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:

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

95% Confidence Intervals

We can work backwards from this:

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

95% Confidence Intervals

We can work backwards from this:

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

95% Confidence Intervals

We can work backwards from this:

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

The random quantities in this statement are $\widehat{\mu}$ and $\widehat{SE}[\hat{\mu}]$.

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!

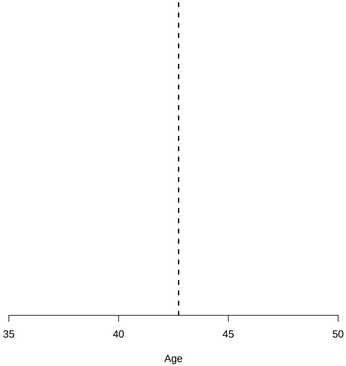
What does this mean?

We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

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

3) Construct the 95% CI:



What does this mean?

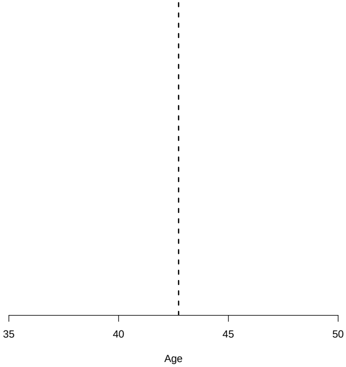
We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

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

$$\hat{\mu} = 42.32 \quad \widehat{SE}[\hat{\mu}] = 1.498$$

3) Construct the 95% CI:



What does this mean?

We can simulate this process using the Kuklinski data:

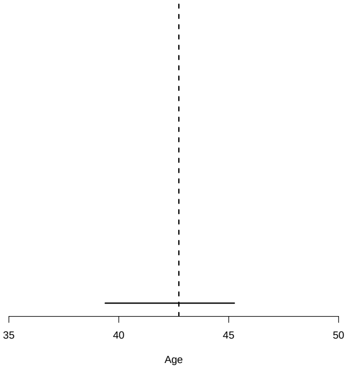
1) Draw a sample of size 100:

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

$$\hat{\mu} = 42.32 \quad \widehat{SE}[\hat{\mu}] = 1.498$$

3) Construct the 95% CI:

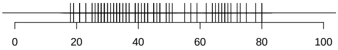
$$(39.4, 45.3)$$



What does this mean?

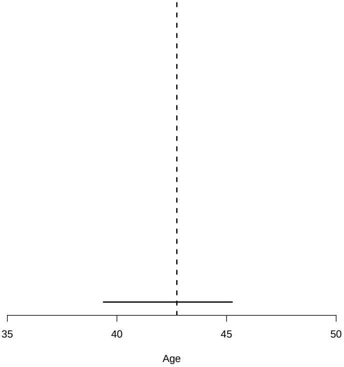
We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:



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

3) Construct the 95% CI:



What does this mean?

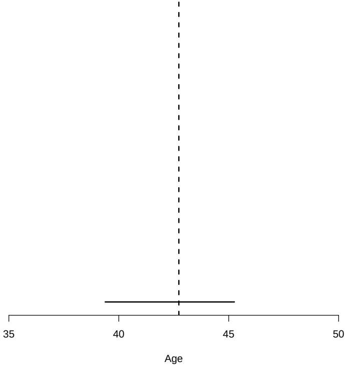
We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

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

$$\hat{\mu} = 41.93 \quad \widehat{SE}[\hat{\mu}] = 1.604$$

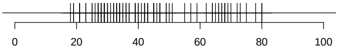
3) Construct the 95% CI:



What does this mean?

We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

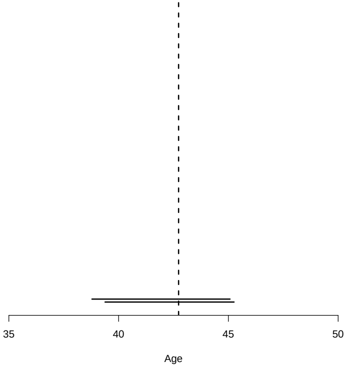


2) Calculate $\hat{\mu}$ and $\widehat{SE}[\hat{\mu}]$:

$$\hat{\mu} = 41.93 \quad \widehat{SE}[\hat{\mu}] = 1.604$$

3) Construct the 95% CI:

$$(38.8, 45.1)$$



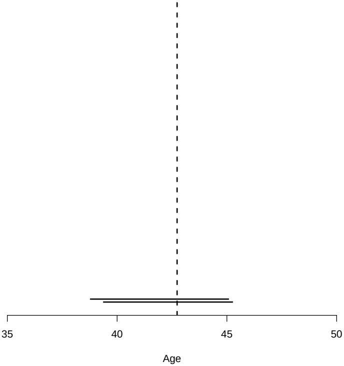
What does this mean?

We can simulate this process using the Kuklinski data:

1) Draw a sample of size 100:

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

3) Construct the 95% CI:



What does this mean?

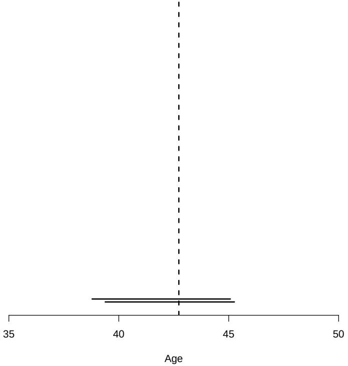
We can simulate this process using the Kuklinski data:

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:



What does this mean?

We can simulate this process using the Kuklinski data:

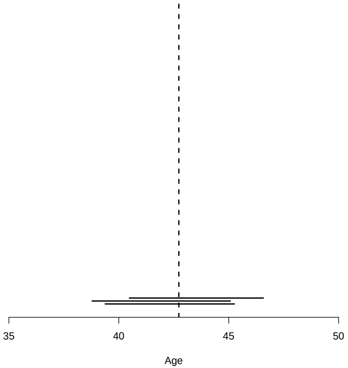
1) Draw a sample of size 100:

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

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

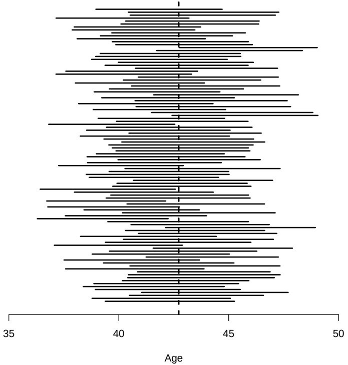
3) Construct the 95% CI:

$$(40.5, 46.6)$$



What does this mean?

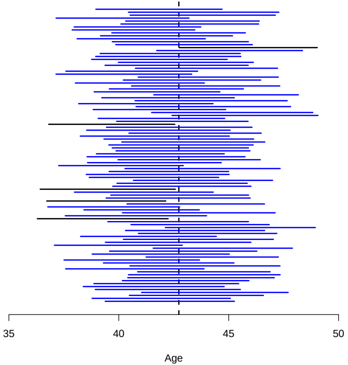
By repeating this process, we generate the sampling distribution of the 95% CIs.



What does this mean?

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

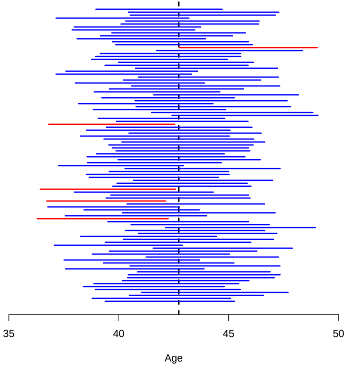
Most of the CIs cover the true μ;



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.

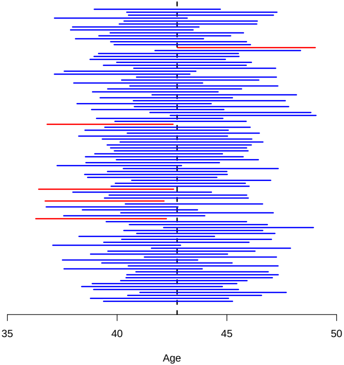


What does this mean?

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

Interpreting a Confidence Interval

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

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 $[2.5,\ 4]$.

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?

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!

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:

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 estimate from a particular sample either contains μ or not

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 estimate from a particular sample either contains μ or not
  - However, if we were to repeatedly calculate the interval estimator over many random samples from the same population, 95% of the time the constructed confidence intervals would cover μ

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 estimate from a particular sample either contains μ or not
  - However, if we were to repeatedly calculate the interval estimator 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.

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:

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

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

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

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

- Best: for a fixed confidence level/coverage probability, find the smallest interval

Is 95% all there is?

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?

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?

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

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?

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.

Normal PDF

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

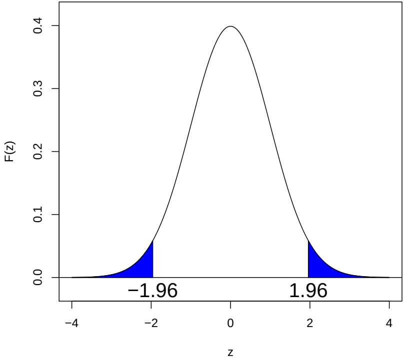
When $(1-\alpha) = 0.95$, we want to pick z so that 2.5% of the probability is in each tail.

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?

Normal PDF

What if we want a 50% confidence interval?

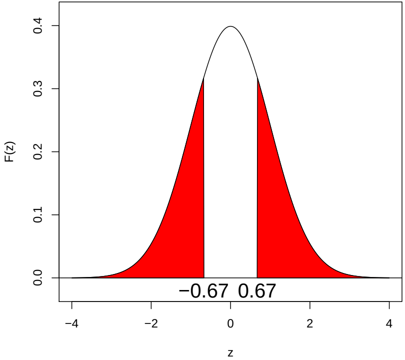
When $(1 - \alpha) = 0.50$, we want to pick z so that 25% of the probability is in each tail.

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

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

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

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

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

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

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

$(1-\alpha)\%$ 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.

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

The problem with small samples

The problem with small samples

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

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.

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).

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).

Canonical Small Sample Example

What happens if we use the large-sample formula?

Canonical Small Sample Example

What happens if we use the large-sample formula?

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

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

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

Canonical Small Sample Example

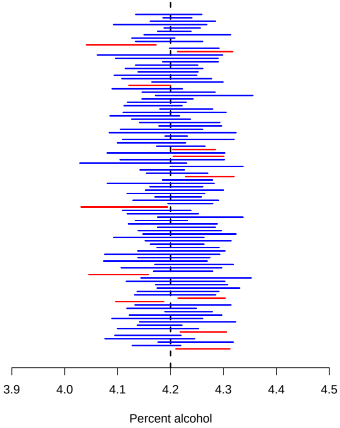
What happens if we use the large-sample formula?

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

The t distribution

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

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

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

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

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

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

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.

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.

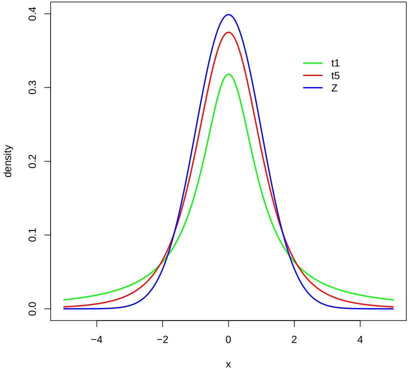
As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

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.

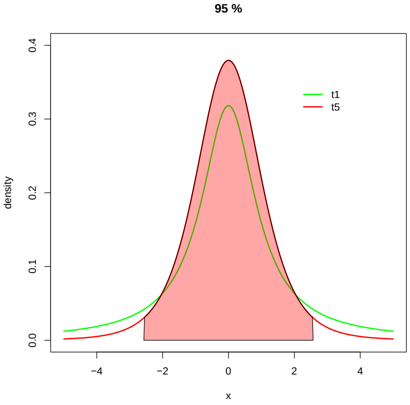


The t distribution

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.

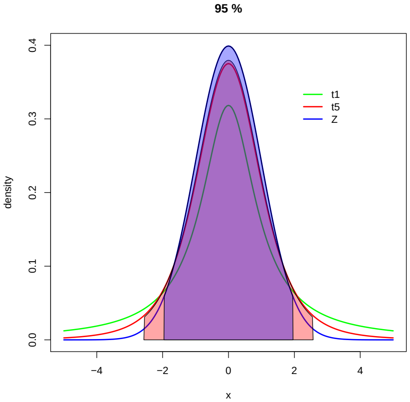


The t distribution

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:

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

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

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

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

$(1-\alpha)\%$ 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.

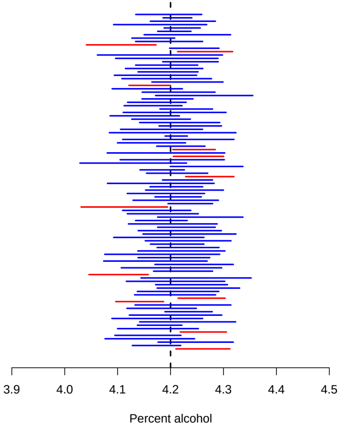
$$\hat{\mu} \pm t_{\alpha/2}\widehat{SE}[\hat{\mu}]$$

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.



Small Sample Example

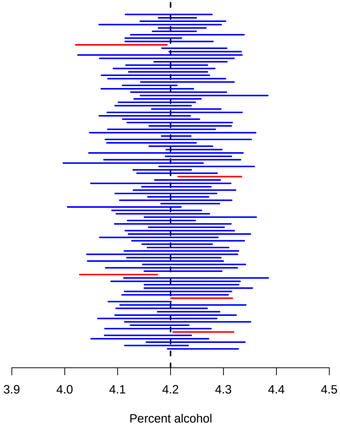
When we generated 95% CIs with the large sample formula

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

only 88 out of 100 intervals covered the true value.

When we use the correct small-sample formula

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



Small Sample Example

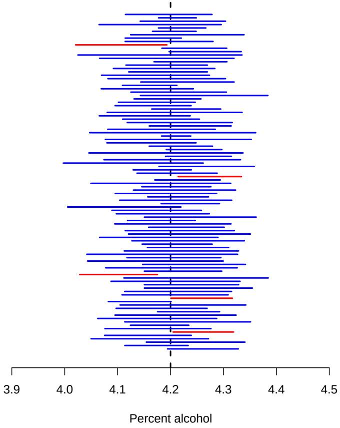
When we generated 95% CIs with the large sample formula

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

only 88 out of 100 intervals covered the true value.

When we use the correct small-sample formula

$$\hat{\mu} \pm 2.57\widehat{SE}[\hat{\mu}]$$



Small Sample Example

When we generated 95% CIs with the large sample formula

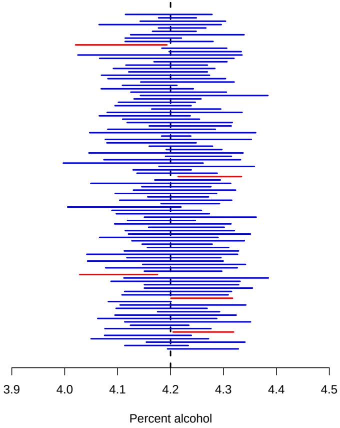
$$\hat{\mu} \pm 1.96\widehat{SE}[\hat{\mu}]$$

only 88 out of 100 intervals covered the true value.

When we use the correct small-sample formula

$$\hat{\mu} \pm 2.57\widehat{SE}[\hat{\mu}]$$

95 of the 100 CIs in this sample cover the truth.



Another Rationale for the t-Distribution

Does $\overline{X}_n \sim N(\mu, S_n^2/n)$, which would imply $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim N(0,1)$?

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).

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.

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,

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

Appendix

Kuklinski Example Returns

Kuklinski Example Returns

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

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?

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

Comparing Two Groups

We will often assume the following when comparing two groups,

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.

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 $\overline{X}_1 - \overline{X}_2$?

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

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

$$E[\overline{X}_1 - \overline{X}_2]$$

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

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

$$E[\overline{X}_1 - \overline{X}_2] = E[\overline{X}_1] - E[\overline{X}_2]$$

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

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

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

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

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

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

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 $\overline{X}_1 - \overline{X}_2$?

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

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

$$Var[\overline{X}_1 - \overline{X}_2]$$

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

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

$$Var[\overline{X}_1 - \overline{X}_2] = Var[\overline{X}_1] + Var[\overline{X}_2]$$

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

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

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

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

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

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

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 $\overline{X}_1 - \overline{X}_2$?

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 $\sim N(\mu_1, \frac{\sigma_1^2}{n_1})$.

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 $\sim N(\mu_2, \frac{\sigma_2^2}{n_2})$.

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: $E[Y_i] = \pi$.

Interval estimation of the population proportion

- Let's say that we have a sample of iid Bernoulli random variables, $Y_1, \dots, Y_n$, 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: $\text{Var}(Y_i) = \pi(1 - \pi)$.

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 $\frac{\pi(1-\pi)}{n}$.

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: $\frac{\hat{\pi}(1-\hat{\pi})}{n}$.

Interval estimation of the population proportion

- Let's say that we have a sample of iid Bernoulli random variables, $Y_1, \dots, Y_n$, 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

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:

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.

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 $\hat{\pi}_C \sim N(\pi_C, (\pi_C(1 - \pi_C))/n_C)$

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:

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

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:

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

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:

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

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

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:

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

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 $[0.056267, 0.069733]$.

Summary of Interval Estimation

Summary of Interval Estimation

- Interval estimates provide a means of assessing uncertainty.

Summary of Interval Estimation

- Interval estimates provide a means of assessing uncertainty.
- Interval estimators have sampling distributions.

Summary of Interval Estimation

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

Next Week

Next Week

- Hypothesis testing

Next Week

- Hypothesis testing
- What is regression?

Next Week

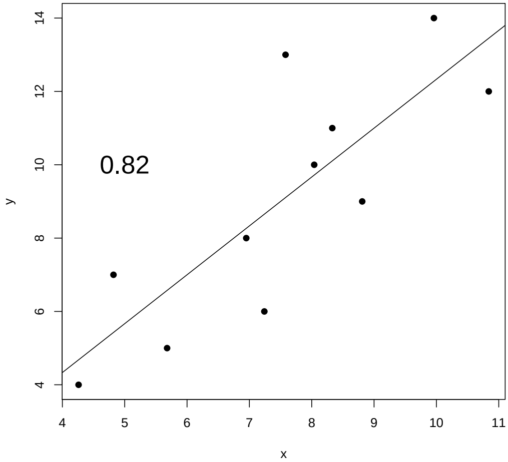
- Hypothesis testing
- What is regression?
- Reading

Next Week

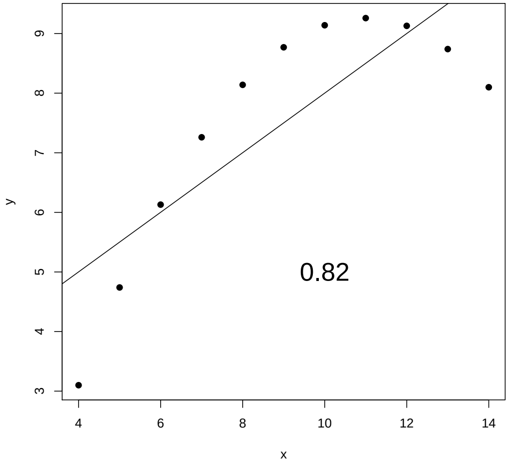
- Hypothesis testing
- What is regression?
- Reading
  - Aronow and Miller 3.2.2
  - Fox Chapter 2: What is Regression Analysis?
  - Fox Chapter 5.1 Simple Regression
  - Aronow and Miller 4.1.1 (bivariate regression)
  - "Momentous Sprint at the 2156 Olympics" by Andrew J Tatem et al. *Nature* 2004

Fun with Anscombe's Quartet

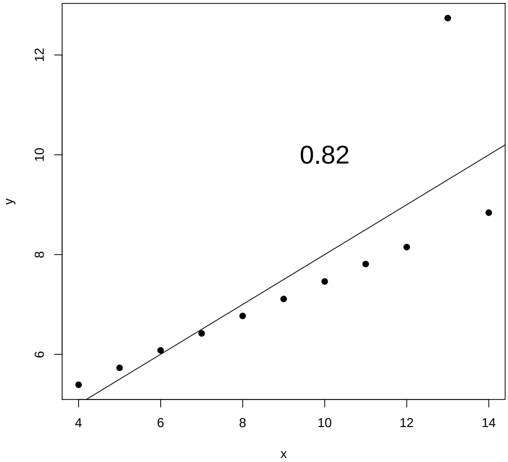
Fun with Anscombe's Quartet



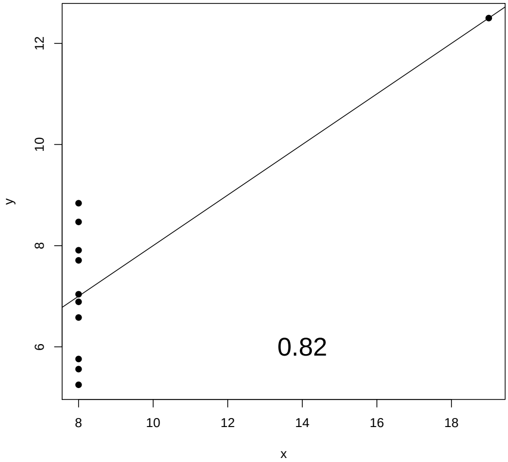
Fun with Anscombe's Quartet



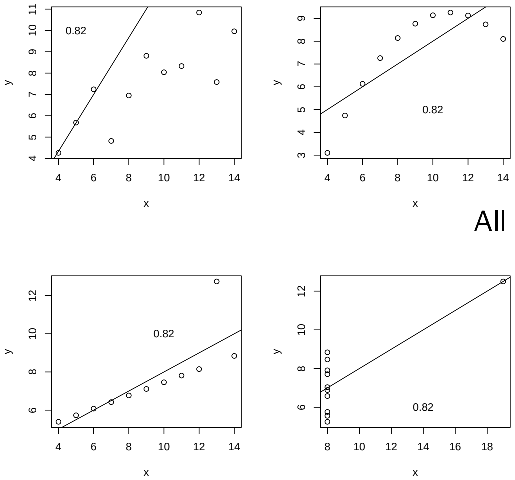
Fun with Anscombe's Quartet



Fun with Anscombe's Quartet



Fun with Anscombe's Quartet



All yield same regression model!

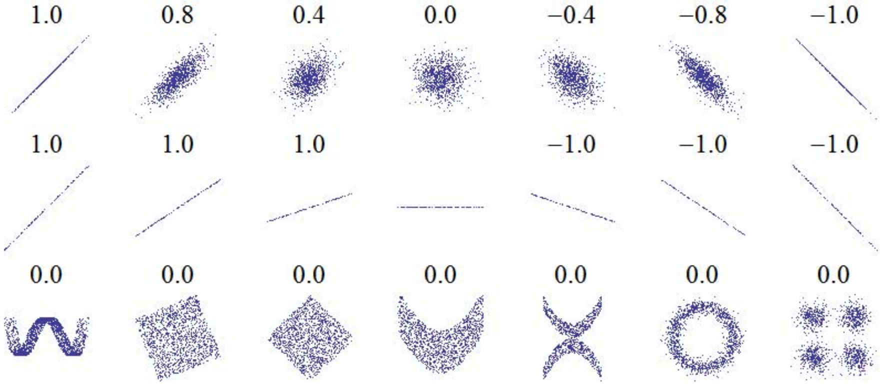
Fun Beyond Correlation

Fun Beyond Correlation

Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. (2011). "Detecting Novel Associations in Large Data Sets". *Science* 334 (6062): 1518-1524.

Fun Beyond Correlation

Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. (2011). "Detecting Novel Associations in Large Data Sets". *Science* 334 (6062): 1518-1524.

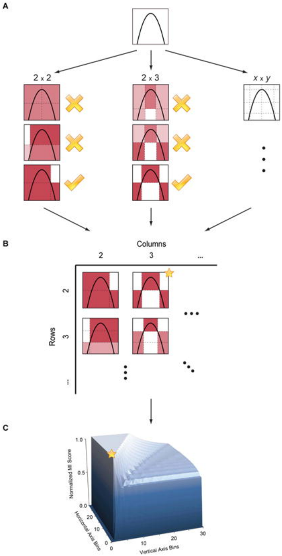


Fun with Correlation Part 2

Enter the Maximal Information Coefficient

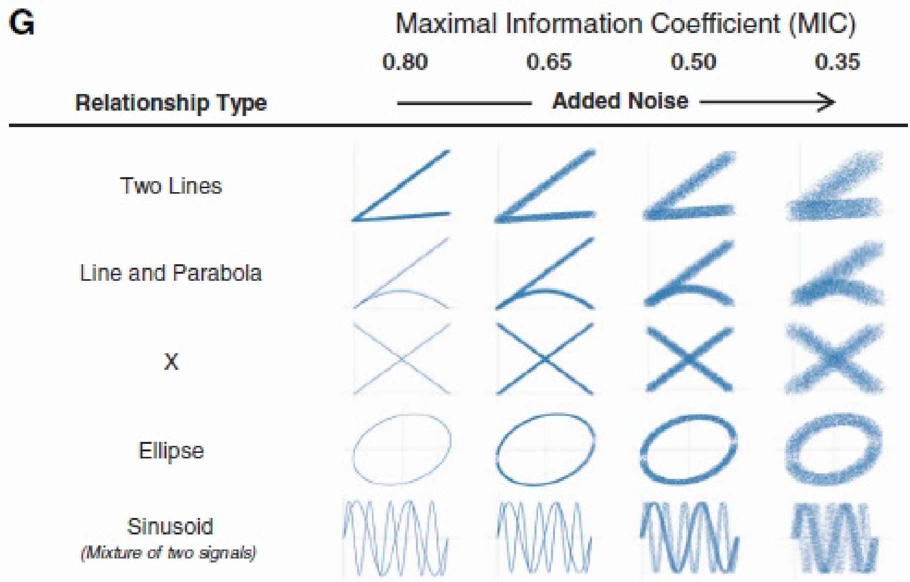
Fun with Correlation Part 2

Enter the Maximal Information Coefficient



Fun with Correlation Part 2

Enter the Maximal Information Coefficient



Fun with Correlation Part 2

Enter the Maximal Information Coefficient

MATHEMATICS

A Correlation for the 21st Century

Terry Speed

Bioinformatics Division, Walter and Eliza Hall Institute of M
Department of Statistics, University of California, Berkeley
E-mail: terry{at}stat.berkeley.edu

Fun with Correlation Part 2

Concerns with MIC

Fun with Correlation Part 2

Concerns with MIC

- low power

Fun with Correlation Part 2

Concerns with MIC

- low power
- originality?

Fun with Correlation Part 2

Concerns with MIC

- low power
- originality?
- heuristic binning mechanism

Fun with Correlation Part 2

Concerns with MIC

- low power
- originality?
- heuristic binning mechanism
- issues with equitability criterion

Fun with Correlation Part 2

Concerns with MIC

- low power
- originality?
- heuristic binning mechanism
- issues with equitability criterion

This is still an open issue!

A Sketch of why the Student t-distribution

Back

A Sketch of why the Student t-distribution

Back

- We have need the distribution of $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$

A Sketch of why the Student t-distribution

Back

- We have need the distribution of $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$
- $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is now a random variable

A Sketch of why the Student t-distribution

Back

- We have need the distribution of $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$
- $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is now a random variable
- Because we already derived the sampling distribution for $\frac{\overline{X}_n - \mu}{\sigma^2/\sqrt{n}}$, we want to derive the sampling distribution for $\frac{S_n}{\sigma^2}$ because the σ^2 term will cancel.

A Sketch of why the Student t-distribution

Back

- We have need the distribution of $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$
- $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is now a random variable
- Because we already derived the sampling distribution for $\frac{\overline{X}_n - \mu}{\sigma^2/\sqrt{n}}$, we want to derive the sampling distribution for $\frac{S_n}{\sigma^2}$ because the σ^2 term will cancel.
- Some math will show our distribution is going to be of the form $\sum Z^2$ where $Z \sim N(0,1)$.

A Sketch of why the Student t-distribution

Back

- We have need the distribution of $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$
- $S_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ is now a random variable
- Because we already derived the sampling distribution for $\frac{\overline{X}_n - \mu}{\sigma^2/\sqrt{n}}$, we want to derive the sampling distribution for $\frac{S_n}{\sigma^2}$ because the σ^2 term will cancel.
- Some math will show our distribution is going to be of the form $\sum Z^2$ where $Z \sim N(0, 1)$.
- Let's figure out what distribution that will be

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$F_X(x) \quad = \quad P(X \leq x)$$

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(Z^2 \leq x) \end{aligned}$$

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$
\begin{aligned}
F_X(x) &= P(X \leq x) \\
&= P(Z^2 \leq x) \\
&= P(-\sqrt{x} \leq Z \leq x)
\end{aligned}
$$

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$
\begin{aligned}
F_X(x) &= P(X \leq x) \\
&= P(Z^2 \leq x) \\
&= P(-\sqrt{x} \leq Z \leq x) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{z^2}{2}} dt
\end{aligned}
$$

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$
\begin{aligned}
F_X(x) &= P(X \leq x) \\
&= P(Z^2 \leq x) \\
&= P(-\sqrt{x} \leq Z \leq x) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{z^2}{2}} dt \\
&= F_Z(\sqrt{x}) - F_Z(-\sqrt{x})
\end{aligned}
$$

The pdf then is

χ^2 Distribution

Suppose $Z \sim \text{Normal}(0, 1)$.
Consider $X = Z^2$

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(Z^2 \leq x) \\ &= P(-\sqrt{x} \leq Z \leq x) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{z^2}{2}} dt \\ &= F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) \end{aligned}$$

The pdf then is

$$\frac{\partial F_X(x)}{\partial x} = f_Z(\sqrt{x}) \frac{1}{2\sqrt{x}} + f_Z(-\sqrt{x}) \frac{1}{2\sqrt{x}}$$

Definition

Suppose X is a continuous random variable with $X \geq 0$, with pdf

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

Then we will say X is a χ^2 distribution with n degrees of freedom. Equivalently,

$$X \sim \chi^2(n)$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$E[X] = E\left[\sum_{i=1}^{N} Z_i^2\right]$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$
\begin{aligned}
E[X] &= E\left[\sum_{i=1}^{N} Z_i^2\right] \\
&= \sum_{i=1}^{N} E[Z_i^2]
\end{aligned}
$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$
\begin{aligned}
E[X] &= E\left[\sum_{i=1}^{N} Z_i^2\right] \\
&= \sum_{i=1}^{N} E[Z_i^2] \\
\text{var}(Z_i) &= E[Z_i^2] - E[Z_i]^2
\end{aligned}
$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^{N} Z_i^2\right] \\ &= \sum_{i=1}^{N} E[Z_i^2] \\ \text{var}(Z_i) &= E[Z_i^2] - E[Z_i]^2 \\ 1 &= E[Z_i^2] - 0 \end{aligned}$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$
\begin{aligned}
E[X] &= E\left[\sum_{i=1}^{N} Z_i^2\right] \\
&= \sum_{i=1}^{N} E[Z_i^2] \\
\text{var}(Z_i) &= E[Z_i^2] - E[Z_i]^2 \\
1 &= E[Z_i^2] - 0 \\
E[X] &= N
\end{aligned}
$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$\text{var}(X) = \sum_{i=1}^{N} \text{var}(Z_i^2)$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$\begin{aligned} \text{var}(X) &= \sum_{i=1}^{N} \text{var}(Z_i^2) \\ &= \sum_{i=1}^{N} \left(E[Z_i^4] - E[Z_i]^2 \right) \end{aligned}$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$\begin{aligned} \text{var}(X) &= \sum_{i=1}^{N} \text{var}(Z_i^2) \\ &= \sum_{i=1}^{N} \left(E[Z_i^4] - E[Z_i]^2\right) \\ &= \sum_{i=1}^{N} (3-1) = 2N \end{aligned}$$

χ^2 Properties

Suppose $X \sim \chi^2(n)$

$$\begin{aligned}\text{var}(X) &= \sum_{i=1}^{N} \text{var}(Z_i^2) \\ &= \sum_{i=1}^{N} \left(E[Z_i^4] - E[Z_i]^2\right) \\ &= \sum_{i=1}^{N} (3-1) = 2N\end{aligned}$$

Student's t-Distribution

Definition

Suppose $Z \sim \text{Normal}(0, 1)$ and $U \sim \chi^2(n)$. Define the random variable Y as,

$$Y = \frac{Z}{\sqrt{\frac{U}{n}}}$$

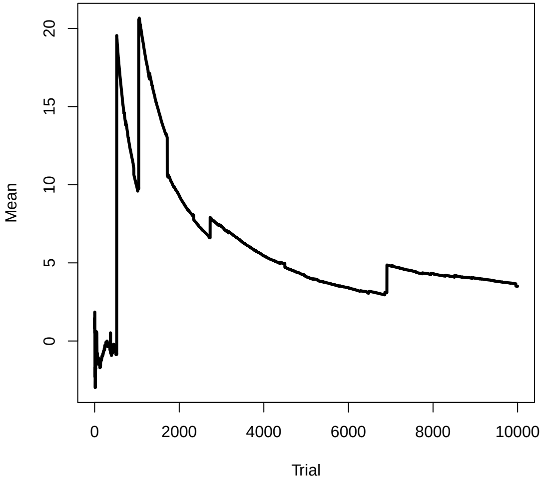
If Z and U are independent then $Y \sim t(n)$, with pdf

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

We will use the t-distribution extensively for test-statistics

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution



Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim \text{Cauchy}(1)$, then:

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim \text{Cauchy}(1)$, then:
$E[X] = \text{undefined}$

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim \text{Cauchy}(1)$, then:
$E[X] = \text{undefined}$
$\text{var}(X) = \text{undefined}$

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim$ Cauchy(1), then:
$E[X] =$ undefined
$\text{var}(X) =$ undefined
If $X \sim t(2)$

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim$ Cauchy(1), then:
$E[X] =$ undefined
$\text{var}(X) =$ undefined
If $X \sim t(2)$
$\text{E[X]} = 0$

Student's t-Distribution, Properties

Suppose $n = 1$, Cauchy distribution
If $X \sim$ Cauchy(1), then:
$E[X] =$ undefined
var$(X) =$ undefined
If $X \sim t(2)$
E[X] $= 0$
var$(X) =$ undefined

Student's t-Distribution, Properties

Suppose $n > 2$, then
$var(X) = \frac{n}{n-2}$
As $n \to \infty$ var$(X) \to 1$.