1-Sample Inference: Confidence Intervals

Engineering Statistics
Section 7.1

Josh Engwer

TTU

30 March 2016

The Need for Sampling: Enormous Populations

Unfortunately, most populations are vastly huge:

- There are hundreds of millions of people in the US.
- There are billions of cans of soda.
- There are trillions of cells in the human body.
- There are too many birds (no one knows an accurate count!)

The enormity of most populations of interest causes various issues:

- It takes too much time & money to poll every single person in the US!
- If taste-testers tested every can of soda, there would be no soda to sell!
- If every cell was drawn from a person, the person would die!
- It's too hard for scientists to capture & tag every bird!

The fix to this intractable problem is to take a **sample** of the population:

Definition

A **sample** is a subset of a population.

As it happens, most methods of statistics involve **samples**.

1-Sample Inference

Chapter 7 starts the foray into **Statistical Inference**:

Definition

Statistical Inference (or just **inference**) is the quantitative study of samples to draw conclusions of populations.

Definition

1-Sample Inference is the quantitative study of one sample to draw conclusions of one population.

Definition

2-Sample Inference is the quantitative study of two samples to draw conclusions of two populations.

Definition

Many-Sample Inference is the quantitative study of many samples to draw conclusions of many populations.

1-Sample Inference

Chapter 7 starts the foray into **Statistical Inference**:

Definition

Statistical Inference (or just **inference**) is the quantitative study of samples to draw conclusions of populations.

Descriptive Stats	Chapter 1
Probability	Chapters 2-5
Point Estimation	Chapter 6
1-Sample Inference: Interval Estimation	**Chapter 7**
1-Sample Inference: Hypothesis Testing	Chapter 8
2-Sample Inference	Chapter 9
Many-Sample Inference	Chapters 10-11
Advanced Inference	Chapters 14-15

1-Sample Point Estimation

Point Estimation is a key building block of Statistical Inference:

Definition

Given a random sample $X_1, \ldots, X_n$ from a population with parameter θ.

Then a **point estimator** $\widehat{\theta}$ of parameter θ is a suitable statistic T of sample:

$$\widehat{\theta} = T(X_1, \ldots, X_n)$$

Recall from Chapter 5 that a statistic T is a <u>function</u> of the random sample.

Often there are several point estimators for a population parameter, but for Chapter 7, these are the only parameters & estimators considered:

POPULATION PARAMETER	POINT ESTIMATOR(S)
Mean μ	$\widehat{\mu} := \overline{X}$
Proportion p	$\widehat{p} := X/n$
Variance σ^2	$\widehat{\sigma}^2 := S^2$
Std Dev σ	$\widehat{\sigma} := S$

The Need for Interval Estimators

A point estimator $\widehat{\theta}$, as its name suggests, is simply a single number estimate for population parameter θ.

An unbiased estimator $\widehat{\theta}$ is expected to be close to the true value of θ, but there's no indication regarding how close the estimator is to the parameter!! Remember, the true value of the population parameter θ is unknown!!

Instead, how about considering an **interval estimator**?

An interval estimator fused with point estimators result in a **confidence interval**.

Confidence Intervals (Definition)

So, what is an interval estimator and confidence interval?? (Brace yourself!)

Definition

Given a population with parameter θ and random sample $\mathbf{X} := (X_1, \ldots, X_n)$.

A $100(1-\alpha)\%$ **interval estimator** $(\theta_{L,1-\alpha}(\mathbf{X}),\ \theta_{U,1-\alpha}(\mathbf{X}))$ for θ is an interval of values constructed from the random sample such that:

$$\mathbb{P}\left[\theta_{L,1-\alpha}(\mathbf{X}) < \theta < \theta_{U,1-\alpha}(\mathbf{X})\right] = 1 - \alpha$$

Moreover, suppose a sample $\mathbf{x} := (x_1, \ldots, x_n)$ is taken from the population.

The $100(1-\alpha)\%$ **confidence interval** $(\theta_{L,1-\alpha}(\mathbf{x}),\ \theta_{U,1-\alpha}(\mathbf{x}))$ for θ is the $100(1-\alpha)\%$ interval estimator but replacing each parameter involved in $\theta_{L,1-\alpha}(\mathbf{X})$ & $\theta_{U,1-\alpha}(\mathbf{X})$ with its point estimator.

The percent $100(1-\alpha)\%$ is called the **confidence level**.

If you don't understand these definitions, don't panic!
When working HW & exam problems, these defn's will never be invoked.

Confidence Intervals (Wrong Interpretation)

Suppose the 95% CI for the average lifetime μ (in years) for all light bulbs is:

$$(9.42, 15.77)$$

How does one interpret this confidence interval??

WRONG INTERPRETATIONS:

"The probability that μ is between 9.42 years & 15.77 years is 0.95" OR
"There is a 95% **chance** that μ is between 9.42 years & 15.77 years."

WHY ARE THESE INTERPRETATIONS WRONG????

Because the population parameter μ is **not random**!!!
i.e. μ is not changing, μ is some fixed value – we just don't know that value!

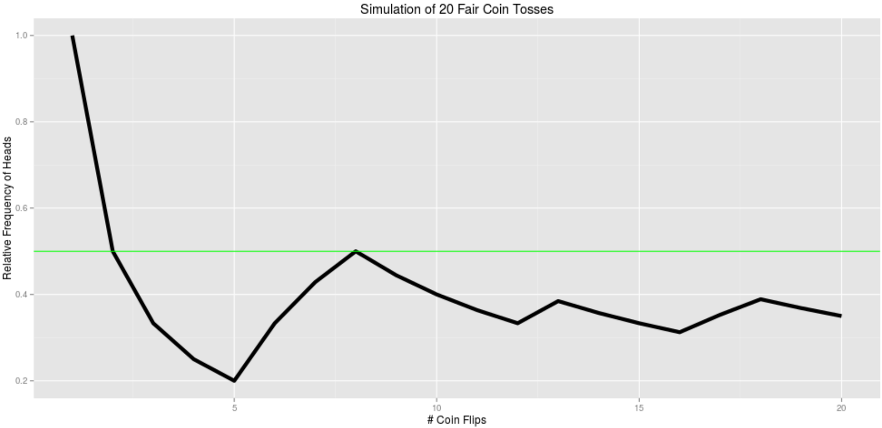
Now, the sample mean $\overline{X}$ of a random sample from this population is random.
Then, it can be shown that the CI is equivalent to the following probability:

$$\mathbb{P}(\overline{X} - 3.325 < \mu < \overline{X} + 3.325) = 0.95$$

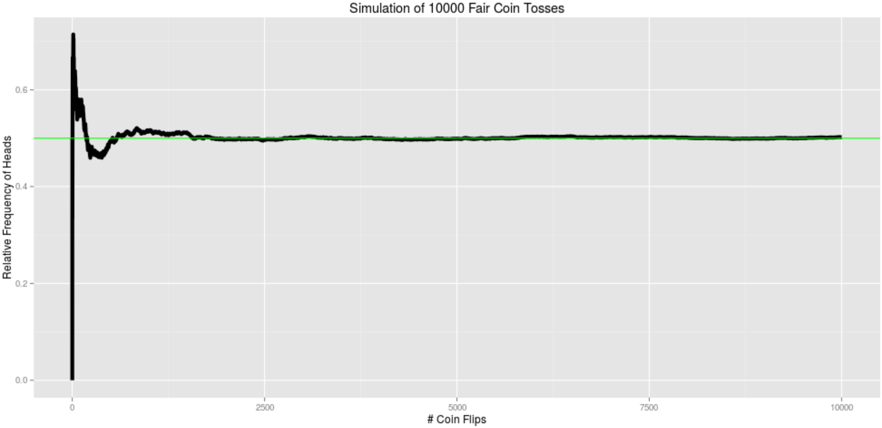
So, in effect, what is actually random is the interval itself!!!!

Deep Interpretation of Probability (Reminder)

The axioms & properties do not give a complete interpretation of probability!!
The most intuitive interpretation is to treat probability as a relative frequency:



The axioms & properties do not give a complete interpretation of probability!!
The most intuitive interpretation is to treat probability as a relative frequency:



i.e. After many many fair coin flips, about 1/2 of them will be Heads.

Confidence Intervals (Right Interpretation)

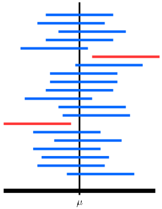
Suppose the 95% CI for the average lifetime μ (in years) for all light bulbs is:

$$(9.42, 15.77)$$

How does one interpret this confidence interval??

RIGHT INTERPRETATION:

"There is 95% **confidence** that μ is between 9.42 years and 15.77 years."
"The probability that the next sample's computed CI will contain μ is 0.95"
"After taking many many samples from the population,
about 95% of the CI's resulting from these samples will contain μ."



Twenty CI's from twenty samples from the population.
Two of the CI's (in **red**) do not contain parameter μ.

How to Improve the Precision of a Confidence Interval

Increasing the confidence level increases reliability but decreases precision:

90% CI for θ	$(3.81, 6.03)$
95% CI for θ	$(1.93, 8.11)$
99% CI for θ	$(0.22, 9.67)$
100% CI for θ	$(-\infty, \infty)$

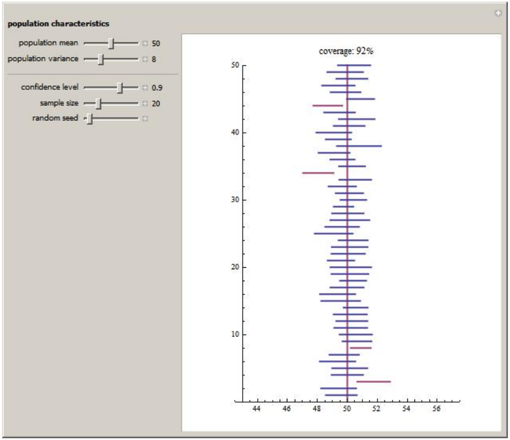
The only way to increase precision without decreasing reliability is to

INCREASE THE SAMPLE SIZE!!!!

This makes sense because if the entire population could be sampled, then the CI would be not an interval, but the exact value of the population parameter.

Confidence Intervals (Demo)

(DEMO) VISUAL INTERPRETATION OF CONFIDENCE INTERVALS (Click below):



Confidence Intervals for a Mean from the Wolfram Demonstrations Project by Chris Boucher

Construction of a Confidence Interval (Procedure)

Proposition

GIVEN: *Random sample* $\mathbf{X} := (X_1, \ldots, X_n)$ *of a population with parameter* θ.
TASK: *Construct the* $100(1-\alpha)\%$ *Confidence Interval for* θ. $(0 < \alpha < 1)$

*(1) Produce a **pivot*** $Q(\mathbf{X}; \theta)$ *which is a statistic such that:*

- Q *is a function of both random sample* $X_1, \ldots, X_n$ *and parameter* θ.
- *The pdf of* Q, $f_Q(x)$, *does not depend on* θ *or any other parameters.*

(2) Find real numbers $a < b$ *such that the following probability holds:*

$$\mathbb{P}[a < Q(\mathbf{X}; \theta) < b] = 1 - \alpha$$

(3) Manipulate the above inequalities to isolate parameter θ:

$$\mathbb{P}\left[\theta_{L,1-\alpha}(\mathbf{X}) < \theta < \theta_{U,1-\alpha}(\mathbf{X})\right] = 1 - \alpha$$

(4) The $100(1-\alpha)\%$ *interval estimator for* θ *is:* $\theta \in (\theta_{L,1-\alpha}(\mathbf{X}),\ \theta_{U,1-\alpha}(\mathbf{X}))$
(5) Obtain a sample $\mathbf{x} := (x_1, \ldots, x_n)$ *from the population*
(6) Compute point estimator(s) for each parameter in $\theta_{L,1-\alpha}(\mathbf{X})$ *&* $\theta_{U,1-\alpha}(\mathbf{X})$
(7) Replace the each parameter with its point estimate, resulting in the CI

$$(\theta_{L,1-\alpha}(\mathbf{x}),\ \theta_{U,1-\alpha}(\mathbf{x}))$$

Textbook Logistics for Section 7.1

- Difference(s) in Notation:

CONCEPT	TEXTBOOK NOTATION	SLIDES/OUTLINE NOTATION
Probability of Event	$P(A)$	$\mathbb{P}(A)$
Pivot	$h(X_1, \ldots, X_n; \theta)$	$Q(\mathbf{X}; \theta)$
Upper Confidence Limit	$u(x_1, x_2, \ldots, x_n)$	$\theta_{U,1-\alpha}(\mathbf{x})$
Lower Confidence Limit	$l(x_1, x_2, \ldots, x_n)$	$\theta_{L,1-\alpha}(\mathbf{x})$

- Ignore "Bootstrap Confidence Intervals" (pg 284)
  - Bootstrap Sampling is useful when a point estimator's expression is either impossible to find or far too complicated to use.
  - Bootstrap Sampling is a numerical method in Statistics.
  - A **bootstrap CI** is centered at an unbiased estimator $\widehat{\theta}$ for parameter θ, but the percentiles are found using bootstrap sampling.
  - Bootstrap Sampling & Bootstrap CI's are beyond the scope of this course.

Fin.