Accuracy of Averages

Gaston Sanchez

Confidence Intervals for Averages

(FPP chapter 23)

Recap

Statistic = Parameter + Chance Error + Bias

what we want to estimate

Likely size given by SE

Hard to measure

Estimate (statistic) likely to be equal to parameter, but off by an SE or so

How accurate is an **estimated** percentage likely to be?

Recap

Statistic = Parameter + Chance Error + Bias

The amount of SE determines accuracy

Confidence Intervals

Confidence Intervals

A confidence Interval is used to give a **likely range for the parameter**

$$\text{Statistic} = \text{Parameter} + \text{Chance Error} + \text{Bias}$$

C.I. for the parameter

Confidence Intervals and Confidence Levels

C.I. at the **68%** level is:

statistic ± 1 SE

C.I. at the **95%** level is:

statistic ± 2 SE

C.I. at the **99.7%** level is:

statistic ± 3 SE

Sample Average

A random sample is taken from a box of unknown composition.

We want to estimate the average of the box.

We use the average of the draws as the statistic.

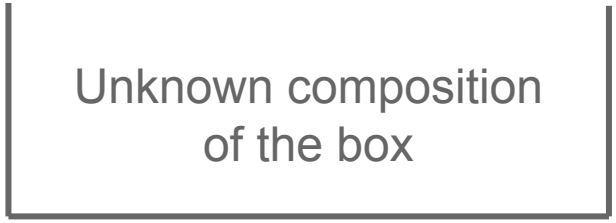
We use the SE for the sample average (i.e. bootstrap method) to compute a confidence interval for the estimated sample average.

Example: Average Income

Turning to statistical estimation and inference

Average income of 25,000 families in a town?

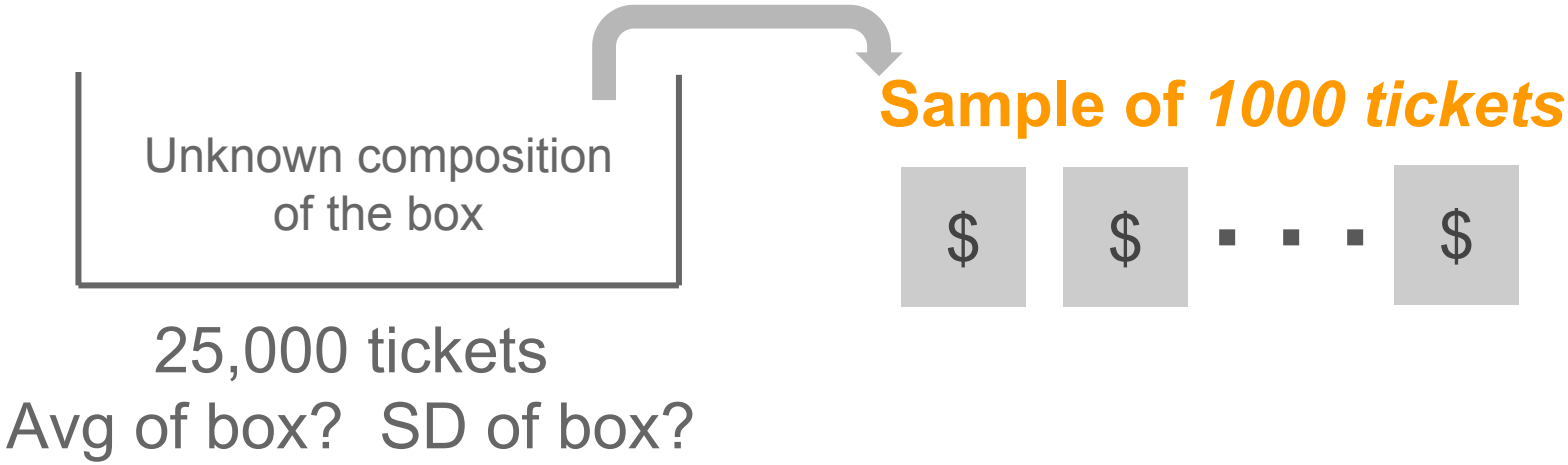
SRS of 1,000 families, with average income of $62,400, and SD of $53,000.



25,000 tickets
Avg of box? SD of box?

Sample of *1000 tickets*

Turning to statistical estimation and inference



SRS of 1,000 families with:

- Average of sample $62,400
- SD of sample $53,000

SE of Sample

SRS of 1,000 families with:

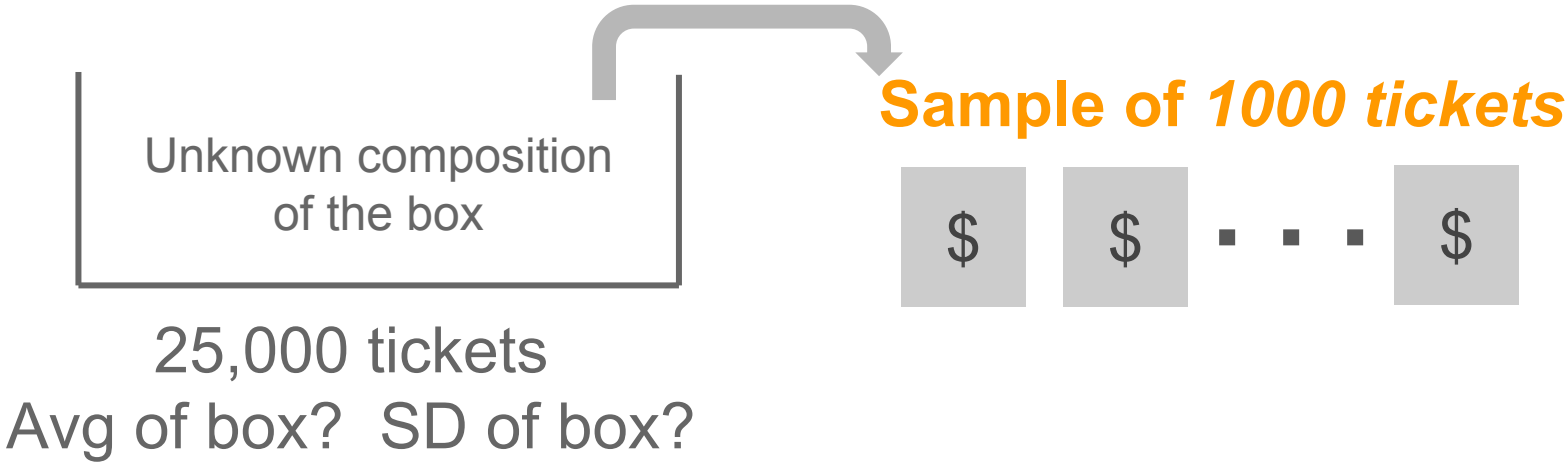
- Average of sample $62,400
- SD of sample $53,000

SD sum = 53,000

SE sum = $\sqrt{1000}$ (53,000) = 1,700,000

SE avg = 1,700,000 / 1,000 = 1,700

Turning to statistical estimation and inference



The average income of all 25,000 families can be estimated as $62,400 ± $1,700

95% Confidence Interval

From Central Limit Theorem, the sample avg follows a normal distribution

2 SEs either way from the sample average:

$62,400 ± 2($1,700) = $59,000 to $65,800

We are 95% confident that the population average is contained within $59,000 to $65,800

99.7% Confidence Interval

From Central Limit Theorem, the sample avg follows a normal distribution

3 SEs either way from the sample average:

$62,400 ± 3($1,700) = $57,300 to $67,500

We are 99.7% confident that the population average is contained within $57,300 to $67,500

Another Example

Example

SRS of 400 Cal students

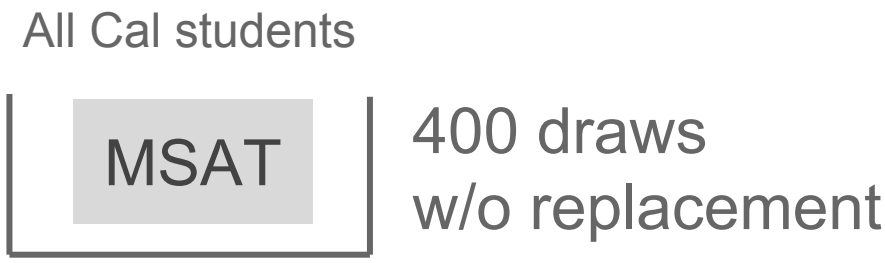
Avg MSAT in sample is 563

MSAT scores follow the normal curve

Sample SD = 90

The average MSAT score for all Cal students is around _statistic_ give or take _SE avg_

Example



SD box = ?

Bootstrap estimate of SD box

SE sum = $\sqrt{400}\,(90) = 1800$

SE avg = 1800 / 400 = 4.5

Example

95% CI for average MSAT for all Cal students is 563 ± 2(4.5)

CI: 554 to 572

99.7% CI for average MSAT for all Cal students is 563 ± 3(4.5)

CI: 549.5 to 576.5

Example

SRS of 400 Cal students

Avg MSAT in sample is 563

Sample SD = 90

The average MSAT score for all Cal students is around 563 give or take 4.5 or so.

True or False

About 68% of the students in the sample have MSAT scores between 473 and 653

(563 - 90, 563 + 90)

TRUE

We know twhat he sample Avg and SD are

True or False

About 68% of MSAT scores of all Cal students are between 558.5 and 567.5

(563 - 4.5, 563 + 4.5)

FALSE

We don’t know what the population Avg and SD are

True or False

A 68% CI for the average MSAT score in the sample is 558.5 and 567.5

(563 - 4.5, 563 + 4.5)

FALSE

Avg MSAT score in the sample is a Statistic
CI are for parameters, not statistics

True or False

For the 95% CI for the average MSAT scores (563 - 9, 563 + 9), it doesn't matter whether the MSAT scores follow the normal curve

TRUE

What matters is that the probability histogram for the sample average follows the normal curve

C.I. Rules

About Confidence Intervals

It is WRONG to say: “There’s a 95% chance that the avg MSAT of all students is between 554 and 572”

It is CORRECT to say: “We are 95% **confident** that the avg MSAT of all students is between 554 and 572”

The parameter is NOT a random number

It feels like chance, but it’s not chance

Chance variability is in the sample process, NOT in the parameter.

To make a Confidence Interval

3) CIs are for parameters

4) Chance is for sample values, sums of draws, random things

5) Average and SD are for data lists of numbers, tickets in a box

6) EV and SE are for sums of draws, sample values, sample averages, sample %'s