The Accuracy of Averages

Results parallel to those developed already for percents.

How accurate is the average of a sample?

Where are we going?

Review: EV and SE of a sum. Central limit theorem

EV and SE of sum ⇨ EV and SE of average

Inference: What does the sample average imply about the population average?

Confidence interval for population average

The Central Limit Theorem

The probability histogram of the sum of a large number of independent draws can be approximated by a normal curve.

Properties of the Sum

- It's expected value is the number of draws times the average of the box
- It's SE is the square root of the number of draws times the SD of the box
- You use these to convert to standard units in using normal curve approximations

To find the approximating normal curve for the sum of the draws you only need to know:

The box average ⇨ expected value of sum

The box SD ⇨ standard error of sum

When the number of draws is large, other facts about the box don't matter.

How Many is a Large Number?

There is no simple, universal answer. If the histogram of the values of the tickets is not too far from normal, the number can be small (say 30).

If the box contains rare tickets with very large values, the number will have to be larger.

"Usually," 100 is large enough.

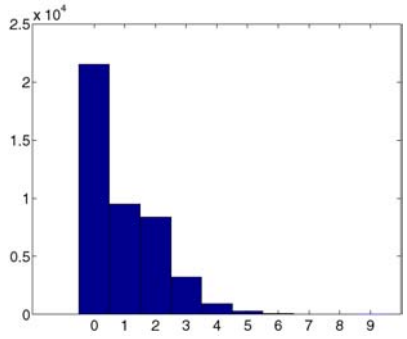
Example: Sampling a Population

The population consists of 43,886 families.

#kids	#families
0	21524
1	9490
2	8355
3	3235
4	902
5	258
6	75
7	32
8	11
9	4

Average = .95 kids/family SD = 1.16

Histogram of number of kids

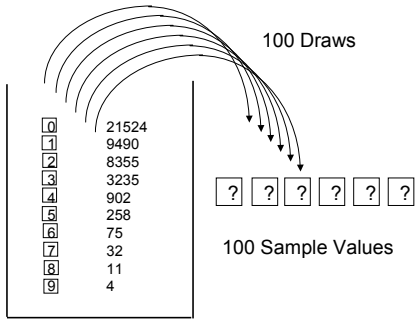


A sample of size 100

Is like drawing from a box

How many tickets in the box? *43,886*

What are the numbers on the tickets? *0,1,…,9*



Box Average = .95 SD = 1.16

100 Draws Box Average = .95 SD = 1.16

The Sum of Draws (= Sample Sum)

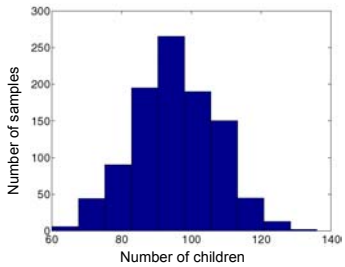
Expected value = _____

SE = _____

The sample sum will be about equal to _____ plus or minus _____ or so.

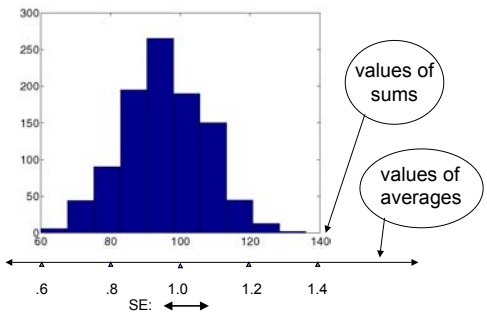
The probability histogram of the sample sum is approximately shaped like a _____.

1000 samples each of size 100. Histogram of their sums is approximately the probability histogram of the sample sum



Expected value = 95; SE = 11.6

Changing to a Histogram of Sample Averages



13

100 Draws Box Average = .95 SD = 1.16

EV of Sum = 95

SE of Sum = 11.6

EV of Average = _____?

SE of Average = _____?

14

Here is what we have found:

Expected Value of the Average = Box Average

Also Called "Population Average"

$$\text{SE of Average} = \frac{\text{SE of Sum}}{\text{sample size}}$$

15

(Notation: n = sample size)

$$\text{SE of Average} = \frac{\text{SE of Sum}}{n}$$

$$\text{SE of Average} = \frac{\sqrt{n} \times \text{Box SD}}{n}$$

$$\text{SE of Average} = \frac{\text{Box SD}}{\sqrt{n}}$$

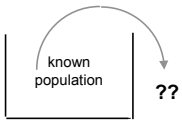
Also called "population SD"

So we know how to do this problem:

From a population with average value = .95 and SD =1.16 a sample of size 100 is taken.

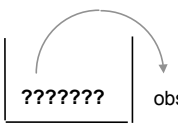
We expect the sample average to be about __.____ give or take ._____ or so.

The chance that the sample average differs from the population average by more than .23 is about _____%



?? random sample values

But what can we say about this problem?



observed sample values

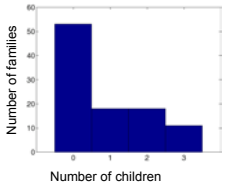
The problem of "**inference**:" What can we conclude about the population on the basis of the sample and with what certainty?

Analogous problem: Suppose you know how far from its target an arrow is likely to be. Now I tell you where an arrow has hit and I ask where the center of the target is likely to be.

19

Example: a sample of size 100 is taken. The average number of children in the sample is .87 per family.

Histogram of values in that sample

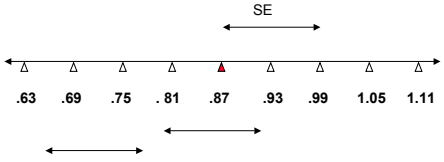


From this sample we would estimate that the average number of children per family in the population to be _____

How accurate is this estimate? What can we say about its uncertainty?

20

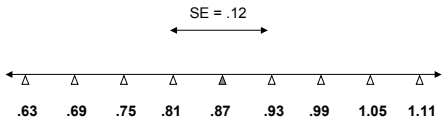
Suppose that we knew that the population SD was 1.16 so that the SE of the sample mean was .12



Is it likely that the population average is .80?

Is it likely that the population average is .65?

21



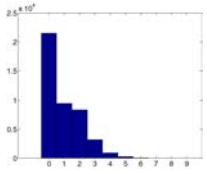
Chance that sample mean is less than 2 SE away from population mean is about _____ %.

Chance that population percent is less than 2 SE away from sample mean is about _____ %.

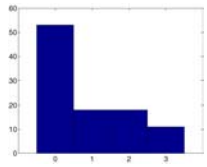
Sample mean plus or minus 2 SE is called a _____ % **confidence interval.** In this case _____ to _____

Now we have been pretending to know the population SD, which gives us the sample SE. What to do in reality, when we don't know the population SD?

Histogram of Population



Histogram of Sample



Population SD = 1.16

Sample SD = 1.07

If we use the sample SD (1.07), our estimate of the SE of the average is 1.07/10 = .11

The sample average (.87) plus or minus twice this estimate is a 95% confidence interval:

_____ to _____

Interpretation

"The chance that the interval plus or minus 2 SEs around the sample average contains the population average is 95%."

What the object subject to chance here: the sample mean or the population mean?

Does the following make any sense: "The probability that the population mean is in the interval .65 to 1.09 is 95%."

25

When can you make probability statements?

I am planning to toss a coin. Is it meaningful to say, "The probability of a head is 50%"?

I am planning a survey. Is it meaningful to say, "The chance that the population average is in a 95% confidence interval around the sample average is 95%"?

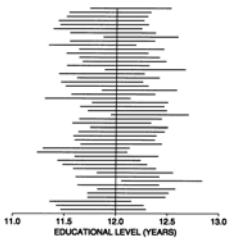
I toss the coin and show you a tail is up. Is it meaningful to say, "The probability of a head is 50%"?

I conduct my survey and I calculate a 95% confidence interval to be 2.2 to 3.8. Is it meaningful to say, "The chance that the population average is between 2.2 and 3.8 is 95%"?

26

Example

A population has an average educational level of 12 years. 95% confidence intervals from 50 simple random samples of size 400:



Why do the intervals have different centers?

Why to they have different lengths?

How many would you expect to contain 12?

How many do?

7

SDs and SEs

The book consistently makes the following distinction:

•An SD is a measure of the spread of a collection of numbers. SD of box, SD of population

•An SE is a measure of the likely size of a chance error. SE of sum, SE of average, SE of count, SE of percentage

In other books and in papers this distinction is not always maintained. You will find:

- "SD of sample mean"
- "SD of sample percentage"

The terminology of boxes and tickets is unique to this book and is used as a pedagogical device. Elsewhere, "population" means "box" and "population values" means "tickets."

Where Have We Been?

Review: SEs

Basic relationship:

SE of sum = $\sqrt{\text{\# draws}}$ x SD of box

SE of average = SE of sum / # draws

SE of count = SE of sum from 0-1 box

SE of percent = 100% x SE of count / # draws

Confidence Intervals

Our confidence intervals are of the form

estimate plus/minus a multiple of the SE of the estimate

To make a confidence interval we would like to know the SD of the box---the population SD. Since we don't know it, we use the SD of the *sample* values---either for a 0-1 in the case of percents or the numerical values in the case of the sample average.

Exercise: which of the following are subject to chance error?

- The population average
- The sample average
- The population SD
- The SE of the sample average
- The sample size
- The SD of the values in the sample

? A sample of size 1000 is taken from a population of size 100,000 adults and you are given the following information:

(1) The percent of the sample who own cars

(2) The average weekly earnings of the people in the sample.

Can you find the SE for the population parameter for (1)? For (2)?

?

100 teenagers are interviewed as they leave a shopping mall and are asked how much money they spent that day. Can a confidence interval for the average expenditure of a teenager in a mall be derived from the information that is collected?

?

A town has 50,000 households. A simple random sample of 1000 is taken and it is found that the average commute distance for the head of household is 8.7 miles and the SD is 9.0 miles.

- The average commute distance of all 50,000 heads of households is estimated to be _____ and this estimate is likely to be off by _____ or so.
- A 95% confidence interval for the average commute distance of heads of households is

_____ to _____

We can estimate that 95% of heads of households commute about this much. True or False?

Driving survey continued: There were 2500 people over the age of 16 in these households. Their average commute distance was 7.7 miles with an SD of 10.2 miles. Can we find a 95% confidence interval for the average commute distance of all the people in the town over the age of 16?

Driving survey continued: 721 of the heads of households commuted by car. Can we find a 95% confidence interval for the percent of all heads of household in the town who commute by car?

? In a simple random sample of 400 households from a large population, the average household income is $40,000 with an SD of $10,000. Circle True or False and explain in one or two sentences:

T F (a) $40,000 plus or minus $1000 is a 95% confidence interval for the average household income in the population.

T F (b) $40,000 plus or minus $1000 is a 95% confidence interval for the average household income in the sample.

T F (c) There is roughly a 95% chance that the average household income in the population is in the range $40,000 plus or minus $1000.

T F (d) Roughly 95% of the households in the population have incomes in the range $40,000 plus or minus $1000.

? Several simple random samples of equal size are taken from a population and for each sample a confidence interval for the population mean is constructed. These intervals are shown below. True or False and explain: It's quite possible that these are 95% confidence intervals.