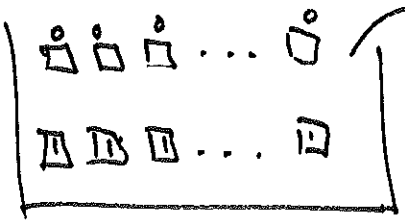
Confidence Intervals.



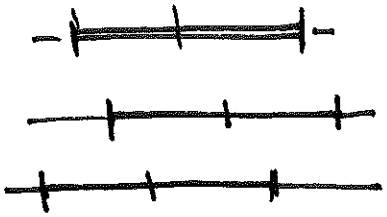
population parameter

sample. (S.R.S.).

↓

derive C.I.

- proportion of 1's in the sample
- size of sample



different samples result in (most likely) different CIs

95% CI for population % is 23% to 41%.

- The CIs for 95% of repeated samples will cover the population parameter

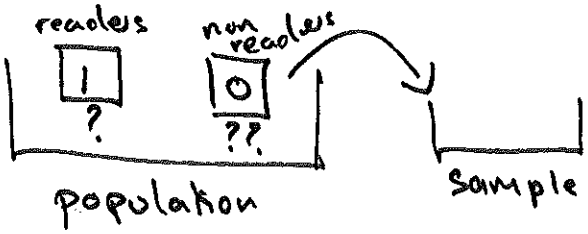
For the one sample that you have collected, you do not know if the true population parameter lies inside your confidence interval.

Example.

SRS 3500 people age > 18

Estimate % of all people over age 18 who read newspapers.

In the sample, there were 2487 newspaper readers.



size 3500.
2487 [1]

Estimate of population % = sample % = $\frac{2487}{3500} \times 100 \approx \underline{71\%}$

$$SE = \sqrt{3500} \times (1-0) \sqrt{\frac{2487}{3500} \times \frac{1013}{3500}} = 27.$$

↑ SE on # of newspaper readers in our sample.

$$SE_{\%} = \frac{27}{3500} \times 100 \approx \underline{0.8\%}.$$

95% CI for % of people aged >18 who read newspapers ≈

$$71 \pm 2 \times 0.8. \%$$

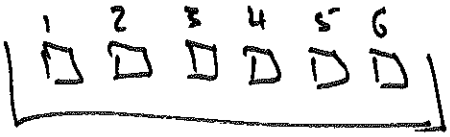
$$69.4 \rightarrow 72.6\%$$

Do we know what the true population % is?

If we drew another sample, would we get the same CI?

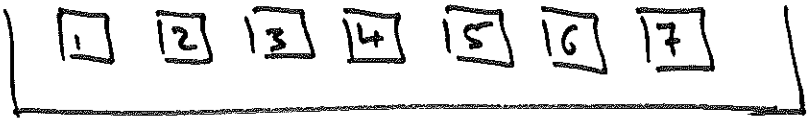
If we drew 100 samples, how many of the 100 CI's would include (cover) the ~~pop~~ true population %?

If we did this, would we know the true population %?



→ sum of draws
= total pip count on
n rolls of fair die.

Inferring <u>mean</u> for box models with arbitrary numbers on the tickets.



mean of box = 4

$SD_{box} = 2.$

look at the variability in the average of draws from this box.

Draw 25 times with replacement.

Expected value of sum of draws $= 25 \times 4 = 100$

$SE_{sum} = \sqrt{25} \times 2 = 10$

Expected value of mean of the numbers drawn

$$= \frac{EV \text{ of sum}}{\#\text{draws}} = \frac{100}{25} = 4.$$

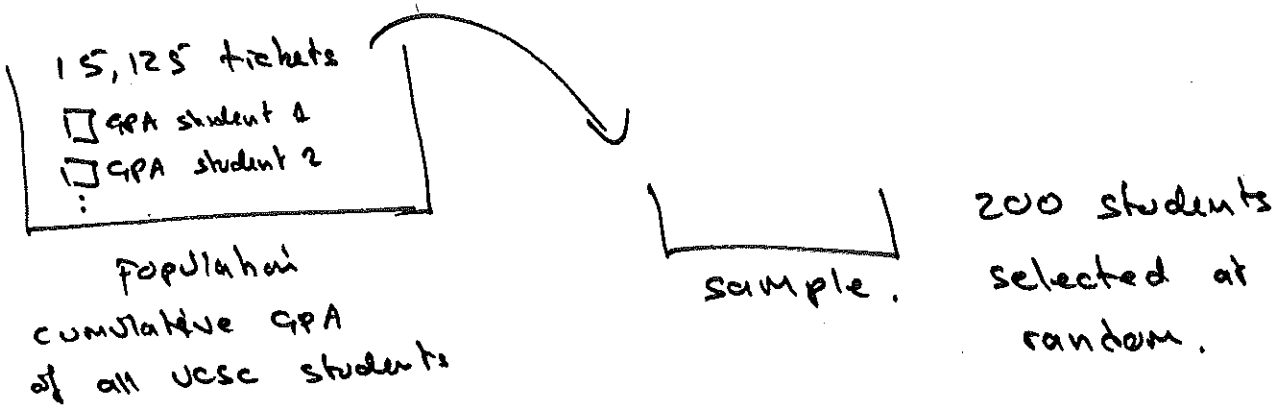
$$SE_{mean} = \frac{SE_{sum}}{\#\text{draws}} = \frac{10}{25} = 0.4$$

$$= \frac{\sqrt{\#\text{draws}} \times SD_{box}}{\#\text{draws}}$$

$$= \frac{SD_{box}}{\sqrt{\#\text{draws}}} = \frac{2}{\sqrt{25}} = \frac{10}{25} = 0.4$$

As the #draws increases, the sample averages become more tightly peaked about the expected value.

Inference for a population mean.



Q: what's the average GPA at UCSC? = average GPA of the students in the sample.

Q: what's a 95% CI for the average GPA at UCSC?

the 200 of you in class can most likely not be considered a simple random sample of all ucsc students.
(Think about grade level, major, etc).

So: I generated a population assuming GPAs uniform between 1 and 4.

then: sampled 200 students from my population.

Sample: ave GPA = 2.5

SD of sample = 0.86

SE of mean = $\frac{0.86}{\sqrt{200}}$ = 0.06

95% CI for the population mean

≡ sample mean ± 2 × SE of mean.

= 2.5 ± 0.12.

Note: SD_{box} was 0.86

CI for mean has width 0.24 ≡ 4×SE.

The CI for the mean of the box is much narrower than the spread in the data.

Hypothesis Testing.

Decision making under uncertainty.

can we attribute the effect we're seeing just to chance?

"tests of significance"

Examples.

- A vaccine is known to be only 25% effective over a period of 2 years.

  A new vaccine is tested on a sample of 2000 people. How do we test if the new vaccine is more effective?

- A machine fills bottles with 333ml of soda. Periodically, a sample of bottles is taken to determine if the average content of the bottles is too low/high.

Null Hypothesis, H_0 – everything we are observing is due to a chance variation due to the random sampling.

Alternative Hypothesis, H_1

vaccine: H_0 : the effectiveness of the new vaccine is 25%

ie new vaccine is the same as the old one.

H_1 : the effectiveness > 25%

bottle filling : H_0 : average content is 333 ml.

H_1 : average content is <u>not</u> 333 ml.

revenue neutral changes to the tax code : H_0 : tax revenues after the change will be the same as before.

H_1 : tax revenues will be different

test H_0 against H_1 using the data in the sample.

→ Build a box model for Ho.

Considering the data as draws from the box, can chance variation (due to the random sampling) explain the data?