95% CI for population %

\[ = \text{sample} \pm 2 \text{SE} \%
\]

If generate repeated samples, and derive a CI from each sample, then about in about 95% of the samples, the population % will lie in the confidence interval.

(The CI "covers" the population parameter.)
$$SD_{\text{box}} = \left( \frac{\text{large number of tickets}}{\text{small number of tickets}} \right) \sqrt{\frac{\frac{500}{\text{fraction of tickets with large number}} \times \frac{800}{\text{fraction of tickets with small number}}}{1300}}$$

\[
\begin{array}{c|c}
500 & 800 \\
1 & 10 \\
\end{array}
\]

$$SD_{\text{box}} = (1 - 0) \sqrt{\frac{\frac{500}{1300} \times \frac{800}{1300}}{}}$$

\[
\begin{array}{c|c}
27 & 4 \\
500 & 800 \\
\end{array}
\]

$$SD_{\text{box}} = (27 - 4) \sqrt{\frac{\frac{500}{1300} \times \frac{800}{1300}}{}}$$
Example.

Simple Random Sample of 3500 people age 18+
to estimate % who read newspapers.

In the sample there were 2487 newspaper readers.

Estimate of population % = \frac{2487}{3500} \times 100 \approx 71%.

\text{Estimate } SE = \sqrt{\frac{3500}{n}} \times \sqrt{0.71 \times 0.29 \times (1-0)}.

\approx 27

\text{Estimated } SE \text{ on } % \text{ of sample size.}

\frac{27 \times 100}{3500} \approx 0.8%.

95\% \text{ CI = estimated } \pm 2 \times \text{Estimated } SE

= \approx 71 \pm 1.6\%.

- Do we know what the true population % is?
- If we generated another sample, would we get the same CI?
- If we generated 100 samples, the 95\% CIs would cover the true value how often?
Accuracy of Averages.

- What can we say about a population mean from a sample from the population?

The tickets in the box (data on the population members) can now be arbitrary.

\[ \begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\end{array} \]

How much variability do we expect in the average of draws from this box?

Draw 25 times with replacement:

Expected value of sum = 25 \times 4 = 100

\[ \text{SD}_{\text{box}} = \sqrt{\frac{(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2}{7}}. \]

\[ = \sqrt{4 + 4 + 1 + 0 + 1 + 4 + 9}. \]

\[ = 2. \]
\[ \text{SE of sum} = \sqrt{25} \times 2 = 10. \]

\[ \text{Expected value of \underline{average}} = \frac{100}{25} = 4. \] (\text{\underline{average of contents of box}})

\[ \text{SE of average} = \frac{10}{25} = 0.4 \]

\[ = \frac{\sqrt{\# \text{draws} \times \text{SD of box}}}{\# \text{draws}} \]

\[ = \frac{\text{SD of box}}{\sqrt{\# \text{draws}}} \]

As the number of draws increases, the \underline{sample means} become more tightly peaked around the \underline{expected value}. 
Expected value of mean = 104

SD_{box} = 2

Sample of size 25

SE of average = \frac{2}{\sqrt{25}} = 0.4
How to do a survey when people can reasonably be expected to lie about the answers?

Eg Did you smoke marijuana in the last week?

Before you answer, toss a coin.

If H -> say "yes"

If T -> answer honestly

For any individual a yes answer doesn't tell me whether they smoked or not.

However from the aggregate responses, I can estimate the % of you who did.

See the practice midterm question.
Inference for population mean.

Cumulative GPA:

Population:

Population mean

= average GPA of all UCSC students.

Try to find a 95% CI for the mean GPA.

I generated a population by assuming that GPA's are uniformly distributed between 0 and 4.

I could have asked all of you for your GPA's, but:

- some of you would lie
- it would not be a simple random sample.

Sample of size 200.

Mean GPA = 2.5

SD of sample = 0.86 (calculated by finding SD of the list of 200 numbers)

I use this in place of population SD.
SE of mean GPA = \( \frac{0.86}{\sqrt{200}} = 0.06 \).

95\% CI for mean GPA of all students in

\[ 2.5 \pm 2 \times 0.06 = 0.06. \]

\[ 2.05 \pm 0.12. \]

SD box was 0.86.

CI for mean has width 0.24 — this is 4 x SE.

The CI for the mean of the draws in much narrower than the spread in the data.
Can my boyfriend come along?

I'm not your boyfriend, you totally are.

But you spend twice as much time with me as with anyone else. I'm a clear outlier.

Your math is face it - I'm statistically significant other.
Box plot:
- Smallest value
- Lower quartile
- Median
- Upper quartile
- Largest non-outlier
- Outliers
Hypothesis Testing

making decisions when uncertainty and variability are present.

- is the effect we're seeing due to chance?

or is the effect we're seeing too large to be accounted for just by chance variation?

"tests of significance"

Example.

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

A new vaccine is being tested on a random sample of 2000 people.

How do we test if the new vaccine is more effective?
A machine fills bottles with 333 ml of liquid. Periodically a sample of bottles is taken to determine if the average amount is too low/high.

Null Hypothesis + Alternative Hypothesis.

Null Hypothesis: $H_0$ - nothing has changed, all variability is due to chance.

Alternative Hypothesis: $H_1$. 