The size of chance errors in surveys
Errors in percentages
Inference from samples
Confidence intervals
Accuracy of averages
Population

everything of interest

Sample

what we collect data about

chance error in the sample.

— depends mostly on the size of the sample
— very little on the size of the population.

<table>
<thead>
<tr>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshmen</td>
</tr>
<tr>
<td>Sophomore</td>
</tr>
<tr>
<td>Junior</td>
</tr>
<tr>
<td>Senior</td>
</tr>
</tbody>
</table>

\[ \text{Total} = 235 \]

Sample random sample.

The % of Juniors in a sample.

Generate a simple random sample.

Note: if sample size is population size can approximate sampling without replacement as if we sampled with replacement.
Sample: 5th row. More likely to be biased.
3 out of 12 Juniors.

11 people \[ \text{people with birthdays on 3/4/5th of month.} \]
3 Juniors. \[ \frac{3}{11} \times 100 = 27\% \]
asume this is a CRD.

How big a chance error away from 32% should we expect?

<table>
<thead>
<tr>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Juniors</th>
<th>everyone else</th>
</tr>
</thead>
</table>

Sample of size 11

Expected # Juniors = \( \frac{\text{size of sample}}{\text{size of box}} \times \text{average of box} \)

\[ = \frac{11 \times 75}{235} = 3.5 \]

\[ \text{SE} = \sqrt{\frac{\text{size of sample}}{\text{SD box}}} \]
\[ SD_{box} = (1 - 0) \sqrt{\frac{75}{255} \times \frac{160}{235}} \]

\[ = 0.466 \]

\[ SE = \sqrt{11} \times 0.466 = 1.5^- \]

Expect 3.5 ± 1.5 Juniors

% of Juniors in our sample.

\[ \text{expected } \% = \frac{\text{expected } \#}{\text{sample size}} \times 100 \]

\[ = \frac{3.5}{11} \times 100 = 32\% - \text{same as population } \% \]

\[ \text{chance error in } \% = \frac{SE}{\text{sample size}} \times 100 = 14\% \]

32\% ± 14\% of our sample to be Juniors

How does the chance error change with sample size?
Sample of size 44.

\[
\text{expected # Juniors} = 44 \times \frac{75}{235} = 14
\]

\[
\text{chance error on } \frac{\text{# Juniors}}{\text{sample size}} = \sqrt{\frac{\text{sample size} \times \text{SD box}}{\text{# Juniors}}} = \sqrt{44 \times 0.466} = 3.1
\]

\[
\text{expected % Juniors} = \frac{14}{44} \times 100 = 32\%
\]

\[
\text{chance error in %} = \frac{3.1}{44} \times 100 = 7\%
\]

Sample is \(\frac{1}{4}\) smaller

chance error in % is \(2\times\) smaller.

This does not depend on the size of the population — provided that the sample is small enough that we can consider the sample to be approximately drawn with replacement.
What happens when the sample is a significant fraction of the population?

\[
\begin{array}{c|c}
75 & 160 \\
\hline
1 & 10
\end{array}
\]

=> slightly less variability.

SE when drawing without replacement = SE when drawing with replacement 

correction factor = \[
\sqrt{\frac{\text{# tickets in box} - \text{# draws}}{\text{# tickets} - 1}}
\]

Sample of 11 out of 235

\[
c_f = \sqrt{\frac{235 - 11}{235 - 1}} = 0.98.
\]

\[\geq 1 \Rightarrow \text{sampling with replacement}\]

Sample of 44 out of 235

\[
c_f = \sqrt{\frac{235 - 44}{234}} = 0.9. \text{ starts to make a difference.}
\]

↑ reduces the chance error we computed earlier.
Corrected chance error \[ = \frac{3.1 \times 0.9}{44} \times 100 = 6.4\% \]

(used to be 7%)

Alternatively

If I fix the sample size, how big must the population be before the correction is negligible.

Sample size = 2500

<table>
<thead>
<tr>
<th>Population size</th>
<th>Correction factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.707</td>
</tr>
<tr>
<td>10,000</td>
<td>0.866</td>
</tr>
<tr>
<td>100,000</td>
<td>0.987</td>
</tr>
<tr>
<td>500,000</td>
<td>0.997</td>
</tr>
</tbody>
</table>

The chance error is 0% depends entirely on the size of the sample.

Known population \[ \rightarrow \] Fluctuations in sample.

Inference.
Inference.

From size and composition of sample, we can say how accurate the estimate of the population parameter will be.

All UCSC students.

Parameter: % freshpeople.

Fresh | Not fresh | Sample
1 | 0 | ?

We don't know the population parameter, so we don't know how many tickets of each type to put in the box.

Assume we have a SRS of size 250.
Assume we have 70 freshpeople in the sample.
What's the SE?

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

\[
\text{SD box} = (1 - \theta) \sqrt{\frac{\text{fraction with } 1's}{\text{fraction with } 0's}}
\]

Use the fractions in the sample as our best guess of the population proportions.

\[
\text{SD box} = (1 - \theta) \sqrt{\frac{70}{250} \times \frac{180}{250}} = 0.449.
\]

SE or \( \pm \) fresh. in sample = \( \sqrt{\text{# draws} \times \text{SD box}} \)

\[
= \sqrt{250} \times 0.449 = 7.1
\]

as a % of sample. \( \frac{7.1}{250} \times 100 = 2.8\% \).

\( \frac{70}{250} \) freshpeople. as % is \( \frac{70}{250} \times 100 = 28\% \).

from our sample of 250 students, we estimate

\[\% \text{ of freshpeople in campus} = 28\% \pm 2.8\%.\]
Sample % = population % + chance error

29 = 28 + 0
28 = 26 + 2
28 = 24 + 4
28 = 30 - 2
28 = 31 - 3

we know to what size chance error to expect.

we convert these to standard units.

Recall that 95% of the time the chance error is within ±2 SE.

Define confidence interval around the sample %.

sample % ± 1 SE is 68% C.I. 25.2 + 30.8
sample % ± 2 SE is 95% C.I. 22.4 + 33.6
We can be about 95\% confident that the \% of freshmen on campus is between 22.4\% and 33.6\%.

Note we have NOT said that the chance of the \% of freshmen being between 22.4 and 33.6 is 95\%.

Because probability is defined as the frequency of occurrence of an event. The population \% is either between 22.4 and 33.6 \% or it isn't.

It doesn't change, no matter how many times you measure the population \%.

The chance involved is in the sample not the parameter.

Registrar's web site:

\[ \text{parameter value} = \frac{4506}{15125} = 29.8\% \]

...there is no variability in this value.
Sample of size 250

to fresh. 95% CI. for population %

\[ 22.4 \rightarrow 33.6 \]

the population % lies in the CI.

The 95% CI covers the population %.

Draw another sample of size 250

62 fresh. in this sample.

% fresh. \( \frac{62}{250} \times 100 = 24.8\% \)

\[
\begin{array}{c|c|c|c}
\text{Sample} & \text{Freq.} & \text{Observed} & 95\% \text{ CI} \\
\hline
1 & 1 & 1 & \\
0 & 10 & 0 & \\
\end{array}
\]

\[ SE = \frac{1}{250} \times SD_{\text{box}} \]

\[ SD_{\text{box}} = \sqrt{250} \times 0.43 \]

\[ SE \text{ on # fresh.} = \sqrt{250} \times 0.43 \]

\[ SE \text{ on % fresh.} = \frac{\sqrt{250} \times 0.43 \times 100}{250} = 2.7\% \]

\[ 95\% \text{ CI} = 24.8 \pm 2 \times 2.7 \text{ mean } \pm 2SE \]

\[ 24.8 \pm 5.4\% \]

\[ 19.4 \rightarrow 30.2\% \]

\[ 22.4 \rightarrow 33.6 \]

different sample gives a different CI.

This CI covers the parameter value.
95% CI
95% of the CIs from different samples cover the population parameter.