\[
\begin{array}{c|c|c}
 & F & M \\
\hline
115 & 91 \\
55.8\% & 44.2\%
\end{array}
\]

\[
\text{mean} = \frac{115}{206} = 0.558.
\]

\[
\text{SD} = (1-0)\sqrt{\frac{115}{206} \times \frac{91}{206}} = 0.4966 \pm 0.5
\]

Sample of size 10 (of whom 7 were women).

Expected \# F = 10 \times 0.558 = 5.58.

\[
\text{SE} = \sqrt{10} \times 0.4966 = 1.57.
\]

In a sample of size 10, expect 5.6 \pm 1.6 women.

(\text{so to get a value of 7 is not unreasonable}).

\underline{Percentages}

Expected \% of F in the sample is \frac{5.58}{100} \times 100 = 55.8\%.

\underline{Chance error in \%} = \frac{\text{SE}}{\text{sample size}} \times 100 = 15.7\%.

\Rightarrow \text{Expect our sample (of size 10) to be 55.8} \pm 15.7\% \text{ female.}
What happens if we take a sample of size 40?

Expected # f in sample = 40 \times 0.558 = 22.3

Expected % f in sample = \frac{22.3}{40} \times 100 = 55.8%.

Expected % does not change with the sample size.

Chance error in # f

\[ SE = \sqrt{40 \times 0.4996} = 3.16. \]

Chance error in % f

\[ \frac{3.16}{40} \times 100 = 7.9%. \]

Size of the sample has increased by a factor of 4.

Chance error in % has decreased by a factor of 2.

As the sample size increases, the chance error in % decreases.

Irrespective of the size of the population, provided that the sample is small enough that we can consider the sample to be approximately drawn with replacement.
Chance error in % depends on the sample size (and not population size) when sample is small in relation to population size.

What happens when the sample is a significant fraction of the population:
the contents of the box get smaller on each draw => slightly less variability

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

Sample of size 10 out of 206

correction factor = \sqrt{\frac{206 - 10}{206 - 1}} \approx 0.98.

Sample of size 40 out of 206

correction factor = \sqrt{\frac{206 - 40}{206 - 1}} = 0.90 \quad \text{starting to become important}

so the chance error in % is reduced from the value we found earlier:
\[ \frac{3.16 \times 100 \times 0.9}{40} = 7.1\% \]
Alternatively.

If we fix the sample size, how large must the population be before the correction is negligible?

Opinion polls often use \( n = 2500 \)

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Corrected Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</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>

⇒ get same chance error in your opinion poll in California as in Wyoming.

So far we've assumed we've known what was in the box (population), and looked at the characteristics of a sample.

Now: given information about a sample, what can we say about the population? – "Inference"
From the size and composition of the sample, can we say how accurate the estimation of the population parameter will be?

Population
- all USC students

want to know % fresh persons.

when studying variability we built a box model.

Don't know how many tickets of each type to put in the box.
Assume that we have a sample:
size 250, 70 freshmen.

If we repeated the sampling, how much variability would we expect?

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

\[ \text{SD of box} = (1-\rho) \sqrt{\frac{\text{fraction of tickets with 1}}{250} \times \frac{\text{fraction of tickets with 0}}{250}} \]

\[ \text{unknown.} \]

What do we do? — use the sample fractions

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

\[ SE \text{ on } \# \text{ freshmen} = \sqrt{250 \times 0.449} = 7.1 \]

\[ SE \text{ on } \% = \frac{7.1}{100} \times \frac{100}{250} = 2.8\% \]

Estimate population \% on the sample 0% \[ (\frac{70}{250} \times 100 = 28\%) \]

\[ \Rightarrow \text{\% of freshmen on campus is } 28\% \pm 2.8\%. \]
\[
\frac{\text{sample } \%}{28} = \frac{\text{population } \%}{28} + \frac{\text{chance error } \%}{0}
\]

\[
\frac{\text{sample } \%}{28} = \frac{\text{population } \%}{26} + \frac{\text{chance error } \%}{2}
\]

\[
\frac{\text{sample } \%}{28} = \frac{\text{population } \%}{24} + \frac{\text{chance error } \%}{4}
\]

\[
\frac{\text{sample } \%}{28} = \frac{\text{population } \%}{32} + \frac{\text{chance error } \%}{-4}
\]

What can we say about the population percentage?

For each value of the population % listed, we can think about the resulting chance error in terms of how many SEs it is.

We can define a confidence interval around the sample % and use the Normal approximation to define the confidence level.

\[
\text{sample } \% \pm 1 \times \text{SE is } 68\% \text{ confidence interval}
\]

\[
\text{sample } \% \pm 2 \times \text{SE is } 95\% \text{ confidence interval}
\]

\[
\pm 3 \times \text{SE } 99.7\% \text{ CI.}
\]
"we can be about 95% confident that the % of fresh persons on campus is between 22.4 % and 33.6 %.

Q: what's the probability that the (population) % of fresh persons on campus lies in the interval 27.4 - 33.6 %

Note: we have not said that the chance of the % of fresh, 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 is not.

However often you measure the entire population this will not change.

due to the chance is in the sampling, not in the parameter.
The parameter takes a fixed value:

\[
\frac{\text{# fresh students}}{\text{total # students}} = \frac{45.06}{15125} = 29.8\%.
\]

Our sample of 70 out of 250 allowed us to compute the confidence interval.

95% CI for population % is 22.4 - 33.6%.

In this case, the 95% CI covers the population %.

Now let's draw another sample of size 250.

This time we have 62 fresh.

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

\[
\text{SE SD of box} = (1-0) \sqrt{\frac{62}{250} \times \frac{18.8}{250}} = 0.43
\]

SE on # fresh = \(\sqrt{250} \times 0.43\)

SE on % fresh = \(\sqrt{250} \times 0.43 \times 100 = 2.7\%\).

95% CI is 24.8 ± \(\frac{2}{250} \times 2.7\)

19.4 ± 30.2%. Again, this covers the population %.
202 different samples, each of size 250

# of fresh people in sample

# samples (frequency)
Each line represents the CI based on a different sample (95% CI).
95% of the CI's generated from different samples do cover the population %.