($\boxed{0}$ $\boxed{1}$)-Boxes

In a ($\boxed{0}$ $\boxed{1}$)-box all the tickets are labeled with a 0 or a 1.

- The ***sum*** of all the tickets in a ($\boxed{0}$ $\boxed{1}$)-box is equal to the ***number*** of $\boxed{1}$s in the box.

- The ***average*** of a ($\boxed{0}$ $\boxed{1}$)-box equals the ***fraction*** of $\boxed{1}$s in the box, or equivalently, the ***percentage*** of $\boxed{1}$s in the box.

- The ***SD*** of a ($\boxed{0}$ $\boxed{1}$)-box is computed using the shortcut

$$SD_{box} = \sqrt{p \cdot (1 - p)},$$

where p is the fraction of $\boxed{1}$s in box and $(1 - p)$ is the fraction of $\boxed{0}$s in box.

Simple random samples from a ($\boxed{0}$ $\boxed{1}$)-box.

A ***simple random sample*** of n tickets drawn from a ($\boxed{0}$ $\boxed{1}$)-box of N tickets is a random sample drawn ***without replacement***.

- The ***expected percentage*** of $\boxed{1}$s in the sample is equal to the percentage of $\boxed{1}$s in the box.

- If the tickets are drawn ***with replacement***, then the ***standard error*** for the ***percentage*** of $\boxed{1}$s in the sample is

$$SE_{\%} = \frac{SD_{box}}{\sqrt{n}} \times 100\%.$$

- When the tickets are drawn ***without replacement***, then the ***standard error*** for the ***percentage*** of $\boxed{1}$s in the sample is

$$SE_{\%} = CF \times \frac{SD_{box}}{\sqrt{n}} \times 100\%,$$

  where the ***correction factor*** is $CF = \sqrt{\frac{N-n}{N-1}}$.

When should we include the correction factor?

(*) For simple random samples (random samples without replacement), it is always correct to include the correction factor when calculating the SE. However if the sample size n is very small compared to the population size N, then the correction factor has a negligible effect (and can be safely ignored).

Example: If $N = 4000$ and $n = 400$, then

$$CF = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{3600}{3999}} \approx 0.949,$$

so the correction factor will have a small but noticeable effect on the $SE_{\%}$, and should be included in the calculation. On the other hand, if $N = 400000$ and $n = 400$, then

$$CF = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{399600}{399999}} \approx 0.9995,$$

so the correction factor will have a negligible effect on the $SE_{\%}$, and we don't need to include it in the calculation.

Normal approximation

- When a simple random sample is drawn from a ($\boxed{0}\ \boxed{1}$)-box, the observed percentage of $\boxed{1}$s in the sample differs from the expected percentage of $\boxed{1}$s by some ***chance error***. This chance error is generally no larger than one or two $SE_{\%}$s.

- If the sample size is ***large enough***, then the probability histogram for the *sample percentages of* $\boxed{1}$s is well approximated by the *normal curve* (after converting to *standard units*).

- This means that if the sample size is ***large enough***, then

$$P(|(\text{observed }\%) - (\text{expected }\%)| < Z \cdot SE_{\%}) \approx Table(Z),$$

  where $Table(Z)$ is the area under the normal curve from $-Z$ to Z (which can be found in the table at the back of the book).

- How large is ***large enough***? If p is the fraction of $\boxed{1}$s in the population (box) and n is the sample size, then the normal approximation is reasonably accurate when both $np \geq 10$ and $n(1-p) \geq 10$.

Example.

Suppose that a simple random sample of 400 tickets is drawn from a ($\boxed{0}$ $\boxed{1}$)-box of 5000 tickets containing 3000 $\boxed{1}$s and 2000 $\boxed{0}$s.

What percentage of $\boxed{1}$s are we likely to see in the sample?

- The expected percentage of $\boxed{1}$s in the sample is 60% (same as the box percentage).
- The standard error is $SE_{\%} = \sqrt{\frac{4600}{4999}} \times \frac{\sqrt{0.6 \cdot 0.4}}{20} \times 100\% \approx 2.35\%$.
- The sample percentage of $\boxed{1}$s is likely to be in the range $60\% \pm 2.35\%$, or between 57.65% and 62.35%. The margin of error here is 1 $SE_{\%}$, and the probability that the sample percentage falls in this range is about 68%.
- If we want a higher probability that the sample percentage falls into the predicted range, we can increase the range. The probability that the sample percentage of $\boxed{1}$s falls in the range $60\% \pm 4.7\%$ (55.3% to 64.7%) is about 95%, since the margin of error is now $2SE_{\%}$.

From the sample to the box...

The estimate

$$\boxed{P\big(\text{population } \% - 2SE_{\%} < \text{sample}\% < \text{population } \% + 2SE_{\%}\big) \approx 95\%}$$

remains accurate even when we don't know the composition of the population!

The boxed estimate above can be rewritten as

$$P\left(|\text{ population } \% - \text{sample}\%| < 2SE_{\%}\right) \approx 95\%$$

and this can be rewritten as

$$\boxed{P\big(\text{sample } \% - 2SE_{\%} < \text{population } \% < \text{sample } \% + 2SE_{\%}\big) \approx 95\%}$$

I.e., we can use the sample percentage to find a *likely* range of values for the population percentage!

The interval $((\text{sample } \%) - 2 \cdot SE_{\%}, (\text{sample } \%) + 2 \cdot SE_{\%})$ is called a ***95% confidence interval*** for the population percentage.

Problem:

If we don't know the composition of the box, then we don't know the SD of the box, so we can't find the $SE_{\%}$!

Solution:

Use the ***sample*** proportions of $\boxed{1}$s and $\boxed{0}$s to estimate the proportions in the box and use these estimates to approximate the SD of the box. If the sample size is big enough, this approximation will be reasonably good.

Example.

A simple random sample of 400 tickets is drawn from a box of $\boxed{1}$s and $\boxed{0}$s containing more than 100,000 tickets. The number of $\boxed{1}$s in the sample is 285 — find 95%-confidence interval for the percentage of $\boxed{1}$s in the box.

(*) The sample percentage of $\boxed{1}$s is $\frac{285}{400} \times 100\% = 71.25\%$.

(*) The sample SD is $\sqrt{0.7125 \times 0.2875} \approx 0.45$,

(*) The estimated $SE_{\%}$ is

$$SE_{\%} = \frac{\text{SD(box)}}{\sqrt{400}} \times 100\% \approx \frac{\text{SD(sample)}}{\sqrt{400}} \times 100\% \approx \frac{0.45}{20} \times 100\% = 2.25\%.$$

(*) A 95%-confidence interval for the percentage of $\boxed{1}$s in the box is

$$(\text{sample } \% \pm 2 \cdot SE_{\%}) = (71.25\% \pm 2 \cdot 2.25\%) = (71.25\% \pm 4.5\%)$$

Note: We don't use the correction factor here, because 400 is very small compared to 100000+ (and we don't know N).

Does this make $SE_{\%}$ bigger or smaller?

What does "95%-confidence" *mean*?

(*) A confidence interval depends on the sample data. Different samples generally produce different sample data — in this case, different sample percentages.

(*) This means that 100 different samples will produce 100 different 95%-confidence intervals — though most of them will be very similar to each other, some perhaps identical.

(*) The percentage of $\boxed{1}$s in the population (box) is unknown but ***fixed***. The intervals we construct vary with the samples.

(*) The term "95%-confidence" means that about 95% of all the intervals we construct using this method will contain the true (but unknown) population percentage.

Example.

A simple random sample of 3500 California voters is surveyed and 2170 of those surveyed say that they support Proposition 101. What is the percentage of all California voters that support this proposition?

Example.

A simple random sample of 3500 California voters is surveyed and 2170 of those surveyed say that they support Proposition 101. What is the percentage of all California voters that support this proposition?

The sample percentage of Prop 101 supporters is

$$(2170/3500) \times 100\% = 62\%.$$

The simple answer is that about 62% of California voters are likely to support the proposition. To give a more precise answer — in the form of a 95%-confidence interval — we need a box model.

Example.

A simple random sample of 3500 California voters is surveyed and 2170 of those surveyed say that they support Proposition 101. What is the percentage of all California voters that support this proposition?

The sample percentage of Prop 101 supporters is

$$(2170/3500) \times 100\% = 62\%.$$

The simple answer is that about 62% of California voters are likely to support the proposition. To give a more precise answer — in the form of a 95%-confidence interval — we need a box model.

(i) Box: California voters. $\boxed{1}$: voter who favors Prop 101.

(ii) Box SD $\approx$ Sample SD $= \sqrt{0.62 \times 0.38} \approx 0.485$

(iii) $SE_{\%} = \dfrac{\text{box SD}}{\sqrt{3500}} \times 100\% \approx \dfrac{0.485}{\sqrt{3500}} \times 100\% \approx 0.82\%$

(iv) A 95%-confidence interval for the percentage of California voters who support Prop 101 is $62\% \pm 1.64\% = (60.36\%, 63.64\%)$.

Observation

In practice, when surveying large populations the accuracy of the prediction depends on the primarily on the sample size, <u>not</u> the <u>relative size</u> of the sample.

What does this mean?

(*) The accuracy of the prediction is given by the margin of error, which is the $SE_{\%}$.

(*) $SE_{\%} \approx CF \times \dfrac{SD_{\text{sample}}}{\sqrt{\text{sample size}}} \times 100\%$

(*) $CF = \sqrt{\dfrac{\text{pop size} - \text{sample size}}{\text{pop size} - 1}}$

(*) If the population size is much bigger than the sample size (which is the usual case), then $CF \approx 1$ and

$$SE_{\%} \approx \frac{SD_{\text{sample}}}{\sqrt{\text{sample size}}} \times 100\%$$

which depends only on the sample size.