(0 1)-Boxes

In a (0 1)-box all the tickets are labeled with a 0 or a 1.

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

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

- The ***SD*** of a (0 1)-box is computed using the shortcut

$$SD_{box} = \sqrt{(\text{fraction of } \boxed{1}\text{s in box}) \cdot (\text{fraction of } \boxed{0}\text{s in box})}.$$

Simple random samples from a (0 1)-box.

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

- The ***expected percentage*** of 1s in the sample is equal to the percentage of 1s in the box.
- If the tickets are drawn ***with replacement***, then the ***standard error*** for the ***percentage*** of 1s 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 1s in the sample is

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

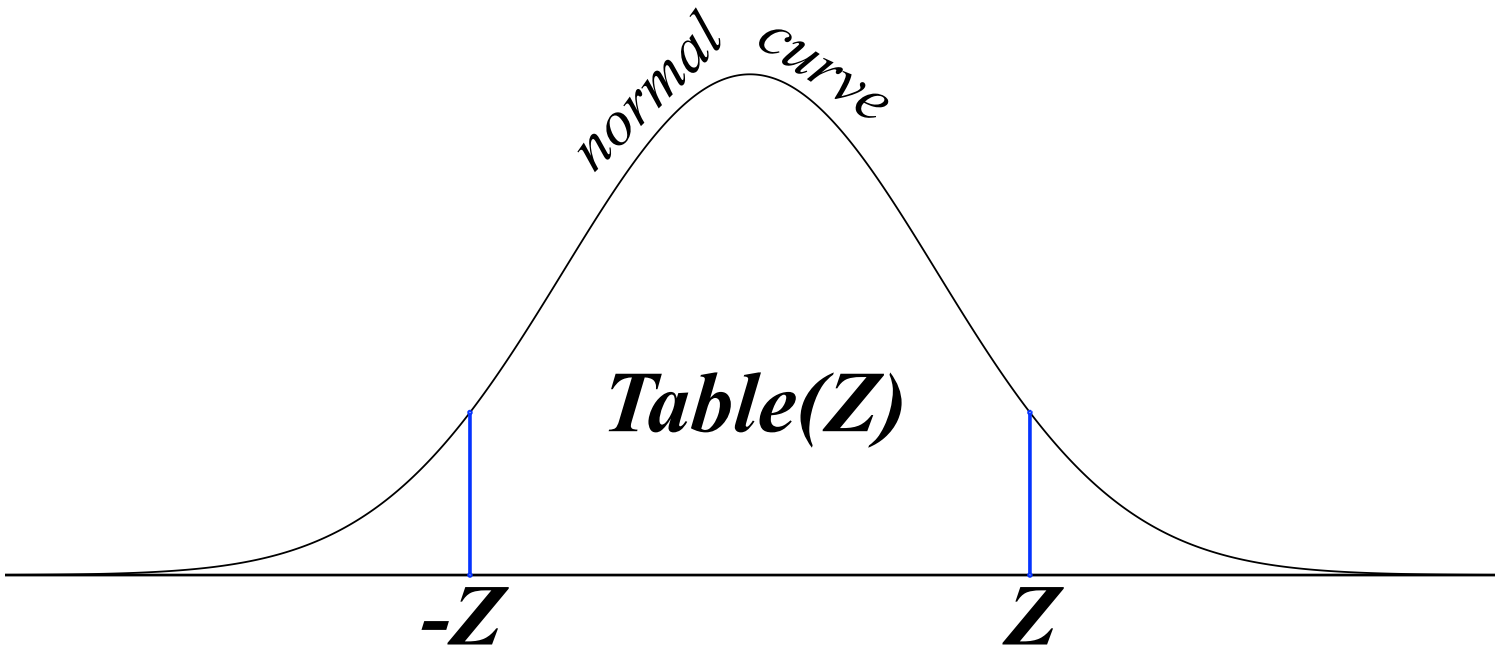
where the ***correction factor*** is $CF = \sqrt{\frac{N-n}{N-1}}$.

Comments:

- If n is small compared to N, then the correction factor (for simple random samples) has a negligible effect (and can be safely ignored).
- 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, after converting to *standard units*, is well approximated by the *normal curve*.
- In practice, 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, as given in the table at the back of the book and depicted in the following figure.



For example:

$$P(|(\text{observed }\%) - (\text{expected }\%)| < 2 \cdot SE_{\%}) \approx 0.95$$

and

$$P(|(\text{observed }\%) - (\text{expected }\%)| < 3 \cdot SE_{\%}) \approx 0.99.$$

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 margin of error. 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 box!

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 1000s of tickets. The number of $\boxed{1}$s in the sample is 285, what is the likely percentage of $\boxed{1}$s in the box?

(*) The sample percentage of $\boxed{1}$s is $\frac{285}{400} \times 100\% = 7.125\%$, so the percentage of $\boxed{1}$s in the box is *likely* to be *about* 71.25%. To make more precise sense of the words '*likely*' and '*about*', we use the $SE_{\%}$ and the normal approximation.

(*) The sample SD is $\sqrt{0.7125 \times 0.2875} \approx 0.45$, so the estimated $SE_{\%}$ is

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

(*) **Improved answer:** The percentage of $\boxed{1}$s in the box is *likely* to be $71.25\% \pm 2.25\%$. I.e., '*about*' can be taken to mean '*give or take one* $SE_{\%}$'.

(*) The normal approximation tells us that the chance is about 68% that the percentage of $\boxed{1}$s in the box is in the range $71.25\% \pm 2.25\%$. I.e., the normal approximation tells us that '*likely*' means that the chance is about 68% in this case

$\Rightarrow$ We can say that the interval $(69\%, 73.5\%)$ is a 68%-confidence interval for the percentage of $\boxed{1}$s in the box.

(*) We often want to be more sure of our estimates — it is more common to use 95%-confidence intervals:

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

$$\text{sample } \% \pm 2SE_{\%} = 71.25\% \pm 4.5\% = (66.75\%, 75.75\%).$$

Observation: A confidence interval depends on the sample data. A different sample will almost certainly yield a different interval. In fact, 100 different samples will probably produce 100 different intervals (though some of them will be very close to each other). The percentage of $\boxed{1}$s in the box is ***constant***. When we say that we are 95%-confident that the percentage of $\boxed{1}$s in the box is in the interval $(66.75\%, 75.75\%)$, what we are actually saying is that 95% of the intervals we construct this way will fall around the true box percentage.

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

(*) The sample percentage of Prop 101 supporters is

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

So the simple answer is to say that about 62% of California likely support the proposition. To give a more precise answer we need a box model.

(*) ($\boxed{0}\,\boxed{1}$) Box: Likely California voters. Tickets: $\boxed{1}$: favors Prop 101. Box SD=???

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

$$SE_{\%} = \frac{\text{box SD}}{\sqrt{3500}} \times 100\% \approx \frac{\text{sample SD}}{\sqrt{3500}} \times 100\% \approx \frac{0.485}{\sqrt{3500}} \times 100\% \approx 0.82\%$$

(*) Better answer: The percentage of likely California voters who support the proposition is likely to be in the range $62\% \pm 0.82\%$ (one $SE_{\%}$). A 95%-confidence interval for the percentage of (likely) voters who support the proposition is $(62\% \pm 1.64\%)$ (two $SE_{\%}$s).

Comments:

- The endpoints of a confidence interval change with each random sample that we draw. The ***population*** parameter that we are trying to estimate using this approach is ***constant***. Think of the interval as a horseshoe and the parameter as the stake in the ground. When we construct a 95% confidence interval for the parameter in question, it is like throwing a horseshoe at the stake, *with a 95% chance of hitting the stake each time.*

- ***Most importantly:*** *The methods outlined today are based on the assumption that the sample of the population from which the numbers are calculated is a* ***simple random sample****. If this assumption is violated, then the conclusions may not be valid. In particular the standard errors that we compute based on this assumption will be much too small.*

Next... Estimating averages.

(*) The expected value and standard error of the sum of n tickets drawn at random with replacement from a box of numbered tickets are

$$EV(\text{sum}) = (\text{Average of box}) \times n \quad \text{and} \quad SE(\text{sum}) = SD(\text{box}) \times \sqrt{n}.$$

The average of the draws is the sum of the draws divided by n, so...

(*) The expected value of the average of n tickets drawn at random with replacement from a box of numbered tickets is

$$EV(\text{Avg}) = \frac{(\text{Average of box}) \times n}{n} = \text{Average of box}.$$

Likewise

(*) The standard error for the average of the draws is

$$SE(\text{Avg}) = \frac{SE(\text{sum})}{n} = \frac{SD(\text{box}) \times \sqrt{n}}{n} = \frac{SD(\text{box})}{\sqrt{n}}.$$