Lecture Notes – Friday, 2/24/17– Estimating Sample Percentages

1 From number ...

When drawing at random with replacement from a **zero-one** box $\left|\overbrace{\boxed{1}\ \ldots\ \boxed{1}}^{M}\ \overbrace{\boxed{0}\ \ldots\ \boxed{0}}^{N}\right|$:

1. The ***expected number*** of $\boxed{1}$s in the sample is $E(\#) = np$, where

- n is the number of draws and
- $p = M/(M+N)$ is the proportion of $\boxed{1}$s in the box (so the percentage of $\boxed{1}$s in the box is $p \times 100\%$).

2. The ***standard error*** for the ***number*** of $\boxed{1}$s in the sample is

$$SE(\#) = \sqrt{n} \times SD(box) = \sqrt{n} \times \sqrt{p(1-p)}.$$

2 ... To percentage

The observed *percentage* of $\boxed{1}$s is the observed number of $\boxed{1}$s in the sample, divided by the number of draws, multiplied by 100%. I.e.,

$$\text{observed percentage of } \boxed{1}\text{s} = \frac{\text{observed number of } \boxed{1}\text{s}}{n} \times 100\%.$$

So...

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

$$E(\%) = \frac{E(\#)}{n} \times 100\% = \frac{np}{n} \times 100\% = p \times 100\% \quad (= \text{percentage of } \boxed{1}\text{s in the box.})$$

2. The ***standard error*** for the ***percentage*** of $\boxed{1}$s in the sample is

$$SE(\%) = \frac{SE(\#)}{n} \times 100\% = \frac{\sqrt{n} \times \sqrt{p(1-p)}}{n} \times 100\% = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \times 100\%$$

Comments:

(a) The normal approximation (central limit theorem) extends to this case as well. I.e., if the number of draws is large enough, then the probability histogram for the observed percentage of $\boxed{1}$s (scaled to standard units) is well-approximated by the normal curve.

(b) If $0 < p < 1$, then $\sqrt{p(1-p)} \leq 1/2$.[†] This means that the standard error for percentage is *always less than* $\frac{1/2}{\sqrt{n}} \times 100\% = \frac{50\%}{\sqrt{n}}$. *Always*, no matter what the distribution of $\boxed{1}$s and $\boxed{0}$s in the box.

3 Examples

Example 1. Suppose that $n = 1600$ marbles are drawn at random with replacement from a jar containing 36 red marbles and 64 blue marbles. What is the chance that the percentage of red marbles in the sample is between 34% and 38%?

- This is like drawing 1600 tickets (at random with replacement) from a box with 36 $\boxed{1}$s (red marbles) and 64 $\boxed{0}$s (blue marbles).
- Parameters: $p = 0.36$ and $n = 1600$.
- Expected percentage of red marbles: $E(\%) = 36\%$ = percentage of red marbles in the jar.

[†]To understand why, find the vertex of the parabola whose equation is $y = x(1-x)$.

- Standard error for the percentage of red marbles: $SE(\%) = \dfrac{SE(\#)}{1600} \times 100\% = \dfrac{\sqrt{0.36 \times 0.64}}{\sqrt{1600}} \times 100\% = 1.2\%$.

- **Normal approximation:**

$$\begin{aligned} P(34\% \leq \text{observed \% of red marbles} \leq 38\%) &\approx \text{area under normal curve between } \frac{34\% - 36\%}{1.2\%} \text{ and } \frac{38\% - 36\%}{1.2\%} \\ &= \text{area under normal curve between } -1.67 \text{ and } 1.67 \\ &\approx 90.3\% \end{aligned}$$

  $\Rightarrow$ The probability 90.3% comes from the normal table.

Example 2. When a fair coin is tossed n times, the law of averages tells us that the probability that we will observe about 50% *heads* approaches 100% as the number of tosses n increases. This claim can be justified by the normal approximation, as follows.[‡]

- Tossing a fair coin is like drawing from the box $\boxed{\boxed{1}\,\boxed{0}}$ at random with replacement: *heads* $\Leftrightarrow \boxed{1}$.

- The expected percentage of *heads* is $E(\%) = 50\%$

- The standard error for the percentage of *heads* in n tosses is

$$SE(\%) = \frac{SE(\#)}{n} \times 100\% = \frac{\sqrt{1/2 \times 1/2}}{\sqrt{n}} \times 100\% = \frac{100\%}{2\sqrt{n}} = \frac{50\%}{\sqrt{n}}.$$

- '*About* 50% *heads*' means that the difference between the observed percentage of heads and 50% is small. How small? As small as we want, if we are willing to toss the coin many times.

- Let's look at a specific example: the law of averages says that the probability

$$P(49.99\% < \text{observed percentage of heads in } n \text{ tosses} < 50.01\%)$$

  approaches 100% as the number of tosses n grows larger.

- We can use the normal approximation to estimate this probability and see how it depends on n:

$$\begin{aligned} P(49.95\% < \text{obs. \% of H in } n \text{ tosses} < 50.05\%) &\approx \text{AUNC between } \frac{49.95\% - 50\%}{SE(\%)} \text{ and } \frac{50.05\% - 50\%}{SE(\%)} \\ &= \text{AUNC between } \frac{-0.05\%}{50\%/\sqrt{n}} \text{ and } \frac{0.05\%}{50\%/\sqrt{n}} \\ &= \text{AUNC between } -\frac{0.05\sqrt{n}}{50} \text{ and } \frac{0.05\sqrt{n}}{50} \\ &= \text{AUNC between } -\frac{\sqrt{n}}{1000} \text{ and } \frac{\sqrt{n}}{1000} \end{aligned}$$

  (AUNC means 'area under the normal curve).

- Now, as n grows large so does $\sqrt{n}/1000$. For example, if $n = 9,000,000$, then $\sqrt{n} = 3000$, so $\sqrt{n}/1000 = 3$ and we can conclude that...

  if we toss a fair coin $9,000,000$ *times, the probability that the percentage of heads will be between* 49.95% *and* 50.05% *is about equal to the area under the normal curve between* -3 *and* 3. *I.e., the chance is about* 99.73% *that the percentage of heads in* $9,000,000$ *tosses is within* 0.05% *of* 50%.

  Comments:

(i) You can repeat this argument with any margin of error (0.05% above). The smaller the margin of error, the larger n will have to be.

(ii) You can also repeat this argument for a box of $\boxed{1}$s and $\boxed{0}$s where the proportion of $\boxed{1}$s is something other than 0.5.

[‡]You can skip this explanation if you want. It is a little technical, but not too bad. You should certainly give it a go if you are a math or science major.

4 Drawing without replacement

In practice — political polling for example — the 'draws from a box' are usually done without replacement. Though ideally, the 'tickets' are still drawn randomly.

Definition. A sample of n tickets drawn from a box without replacement is called a ***simple random sample*** if it is drawn in such a way that every possible set of n tickets in the box is just as likely to be drawn as any other.

Also, in practice, drawing a simple random sample is often difficult (e.g., prohibitively time-consuming and expensive) if not impossible. The math is more complicated for the types of sampling that is often used in practice (e.g., 'probability sampling' — see the end of Chapter 19 in the textbook), so we will focus on the case of simple random samples.[§] Technical details aside, the basic ideas are the same.

One more observation before returning to the math. When drawing without replacement, the sample size n cannot be bigger than the population size (the number of hypothetical tickets in the hypothetical box). Moreover as the sample size grows larger, it becomes more likely that the distribution of tickets in the sample will mirror the distribution of tickets in the box. For example, if we draw all the tickets in the box, the sample is the box so the distributions are identical. This is reflected in the fact that the standard errors for simple random samples are smaller than the standard errors for random samples *with replacement* (with the same sample size).

Math-Facts:
Suppose that a simple random sample of n tickets is drawn from a box of N tickets containing pN $\boxed{1}$s and $(1-p)N$ $\boxed{0}$s . Here p is the proportion of $\boxed{1}$s in the box, so $p \times 100\%$ is the percentage of $\boxed{1}$s in the box ($0 < p < 1$) and of course $n \leq N$.

- The ***expected percentage*** of $\boxed{1}$s in the sample is the same as the percentage of $\boxed{1}$s in the box. I.e.,

  $$E(\%) = p \times 100\%.$$

  This is the *same* as when drawing at random *with replacement*.

  $\Rightarrow$ The 'expected percentage' is the average of the 'box of percentages'. Imagine drawing every possible sample of n tickets from the original box (without replacement); for each such sample find the percentage of $\boxed{1}$s in that sample; and write each possible sample percentage on a ticket; and put these tickets in a new box. This is the 'box of percentages'.

- The ***standard error*** for the percentage of $\boxed{1}$s in the sample is

  $$SE(\%) = \sqrt{\frac{N-n}{N-1}} \times \overbrace{\frac{\sqrt{p(1-p)}}{\sqrt{n}} \times 100\%}^{\text{SE(\%) for drawing } \textit{with} \text{ replacement}} \quad .$$

  In other words, the standard error for percentage when drawing *without replacement* is equal to the standard error for percentage when drawing *with replacement*, multiplied by the ***correction factor***

  $$CF = \sqrt{\frac{N-n}{N-1}}$$

  Two things are worth noting about the correction factor:

  (i) As the sample size n grows, the numerator of the correction factor shrinks, so the standard error becomes smaller.

  (ii) On the other hand, if the sample size n is much smaller than the population size N (the usual case in political polling for example), then the correction factor is very close to 1 and can be safely ignored.

  E.g., if $N = 10,000,000$ and $n = 2500$, then

  $$CF = \sqrt{\frac{9,997,500}{9,999,999}} \approx 0.9999.$$

- The normal approximation works just as well in the case of simple random samples as in the case of drawing with replacement as long as the sample size is large enough (but still small compared to the size of the population).

[§] One of the main differences between using simple random samples and the types of sampling used in the 'real world' is that the standard errors for the real-world sampling methods tend to be bigger than the standard errors for simple random samples.

5 How big should the sample be?

When choosing a sample size, we try to make the ***likely*** size of the chance error ***as small*** as possible. The chance error is the difference between the observed percentage of $\boxed{1}$s in the sample and the expected percentage (the percentage of $\boxed{1}$s in the hypothetical box).

The standard error provides an estimate for the likely size of the chance error, and as we saw above, the chance error (for percentage) grows smaller as the sample size grows larger. When drawing without replacement, the standard error grows smaller a little bit faster (compared to the drawing with replacement) as the sample size grows because of the correction factor. However, when the size of the population is large this effect of the correction factor is negligible for most practical purposes until the sample size is at least 1% of the population or more.

This means that when choosing the size of sample from a large population, the size of the population can usually be ignored. The accuracy of the estimate will depend much more on the sample size than on the relative size of the sample size.

Example 3. Suppose that simple random samples of 2000 likely California voters (there are about $18,000,000$) and 2000 likely Vermont voters (there are about $500,000$) are polled, asking whether they support Bernie Sanders for King of America. The parameter recorded is the percentage of Bernie supporters in both samples, and suppose that his actual support in both states is 40%.

The standard error for percentage in the California poll will be

$$SE_C(\%) = \sqrt{\frac{18000000-2000}{18000000-1}} \times \frac{\sqrt{0.4 \times 0.6}}{\sqrt{2000}} \times 100\% \approx 1.0953\%$$

and the standard error for percentage in the Vermont poll will be

$$SE_V(\%) = \sqrt{\frac{500000-2000}{500000-1}} \times \frac{\sqrt{0.4 \times 0.6}}{\sqrt{2000}} \times 100\% \approx 1.0954\%.$$