STA304

Neil Montgomery

2016-01-28

proportions (and counts)

categorical data

A significant amount of sample data is categorical in nature, as opposed to fundamentally numerical. But the line can be blurry.

- "Failed" vs. "Not Failed"
- "Liberal", "Conservative", "NDP", "Green", "BQ"
- Gas pipes come in 1.0, 1.5, and 1.75 inch sizes
- Likert scale questions (thousands of papers arguing about these!)

In this course we'll only consider the case of two outcomes, which is not as limiting as it sounds.

LifeProTip: use *proportions* and not *percentages*—the latter are merely formatting conventions for human viewing.

analogy/notation

Recall the actual population:

$$\{e_1, e_2, \ldots, e_N\}$$

which we have usually conflated with (numerical) *measurements taken* on thepopulation:

$$\{y_1, y_2, \ldots, y_N\}$$

For the case of proportions and counts, we can simply (arbitrarily) code the two outcomes as 0 and 1 and then use all the previous theory.

But just to mess with you we'll change the notation. Rather than μ (the population mean) we'll use p (the population proportion of 1s) although the definition is identical: $p = \frac{1}{N}\sum_{i=1}^{N} y_i$. For convenience we'll also define $q = 1 - p$ (the population proportion of 0s)

the population variance (not in book!)

The population variance is *defined* (like before) as: $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - p)^2$. This turns out to be:

$$\begin{aligned}
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - p)^2 &= \frac{1}{N} \left(\sum_{i=1}^{N} y_i^2 - Np^2 \right) \\
&= \frac{1}{N} \left(\sum_{i=1}^{N} y_i - Np^2 \right) \\
&= \frac{1}{N} \left(Np - Np^2 \right) \\
&= p(1 - p) = pq
\end{aligned}$$

the sample and its properties

So we don't know what p is. We plan to gather a sample $\{y_1, y_2, \ldots, y_n\}$. Under simple random sample this sample has some familiar properties.

- The y_i have the same distribution with pmf as follows

$$y_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p = q \end{cases}$$

  (stats people would prefer the more useful: $p(y) = p^y(1-p)^{1-y}$ *for* $y \in \{0, 1\}$*)*

- $E(y_i) = p$ and $V(y_i) = pq$

- But the y_i are *not* independent…

- We won't use this, but what is the *distribution* of k, the number of 1s (some may know)?

estimating p under SRS

Use the usual estimator with a new name:

$$\overline{y} = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

If I need to I might also declare k to be the "number of 1s" in the sample and use $\hat{p} = k/n$.

For convenience we'll also define $\hat{q} = 1 - \hat{p}$.

$\hat{p}$ is unbiased for p (easy exercise)

the true variance of $\hat{p}$

Since there's nothing special about 0s and 1s for population values we can recycle the old methods:

$$\begin{aligned} V(\hat{p}) &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \text{ (copied from before)} \\ &= \frac{pq}{n}\left(\frac{N-n}{N-1}\right) \text{ (since } \sigma^2 = pq) \end{aligned}$$

And, $s^2 = \frac{1}{n-1}\sum_{i=1}^n \left(y_i - \hat{p}\right)^2$ is still used to estimate $\sigma^2 = pq$. It turns out that $s^2 = \frac{n\hat{p}\hat{q}}{n-1}$, so we end up with this unbiased estimator for $V(\hat{p})$:

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}$$

Due to symmetry $\hat{V}(\hat{q}) = \hat{V}(\hat{p})$.

(counts)

Population total remains: $\tau = \sum_{i=1}^{N} y_i$. No special notation this time.

Results: use $Np = \tau$, so essentially multiply by N as appropriate.

example

A gas distribution company has employees and contractors who install and repair equipment. It is important that the work be performed up to standard. The company has a Quality Assurance (QA) program that audits samples of work done. The QA program selects a SRS of $n = 412$ out of $N = 12251$ tasks to audit. A task "conforms" if it meets all "technical" *and* "paperwork" standards.

In the SRS $k = 377$ tasks were found to conform. Estimate the proportion of conforming tasks. Estimate the number of non-conforming tasks that will require follow-up work.

We have $\hat{p} = 0.915$ with standard error $\sqrt{\hat{V}(\hat{p})} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}} = 0.014$ and a 95% C.I. is $[0.888, 0.942]$.

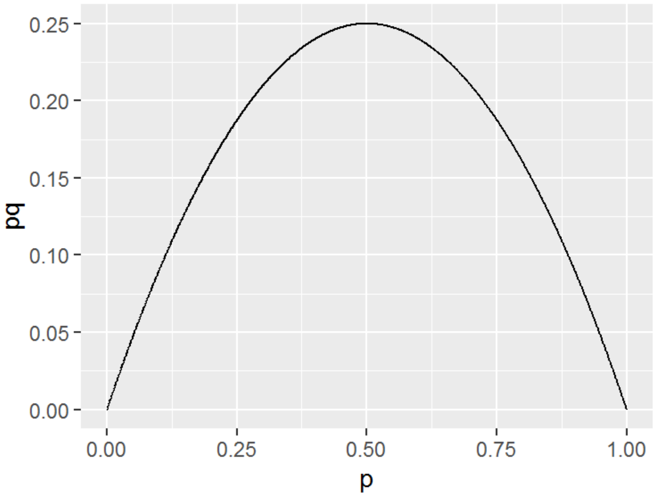
A 95% C.I. for the number of non-conforming tasks (simply Nq, estimated by $N\hat{q}$) is then $[709.487, 1371.993]$.

sample size selection for proportions

For a bound B on estimation error and assuming 95% confidence level, the recycled formula becomes:

$$n = \frac{N\sigma^2}{(N-1)B^2/4 + \sigma^2} = \frac{Npq}{(N-1)B^2/4 + pq}$$

along with the classic "but we don't know" (…in this case…) "p" problem.



How to handle this problem requires an understanding of $pq = p(1-p)$ as a function of p.

handling the unknown p problem

The basic rule is: use as a "guess" the closest value to 0.5 that prior knowledge gives you. Examples:

- You have good information that the true value is between 0.2 and 0.3. **Use 0.3 in the sample size formula**.
- You have good information that the true value is between 0.75 and 0.85. **Use 0.75 in the sample size formula**
- You have no information. **Use 0.5**.
- You have good information that the true value is between 0.4 and 0.6. **Use 0.5**.

example

For the next audit cycle the QA department wants a closer estimate of the amount of follow-up work that will have to be done. Last time the estimate was within $N\sqrt{\hat{V}(\hat{q})} = 331.253$. This time they want to be within 200. What should the sample size be this time?

The sample size formula $n = \frac{Npq}{(N-1)B^2/4+pq}$ is in terms of a bound B for the error in estimating p (or q). We're interested in the total Nq. So in the formula put B/N in place of B.

There's also the matter of the unknown p. We can use the 95% C.I. from before as our "good information": $[0.888, 0.942]$. The closest to 0.5 is 0.888.

example - answer

So the formula becomes:

$$\begin{aligned} n &= \frac{Npq}{(N-1)(B/N)^2/4 + pq} \\ &= \frac{12251(0.888)(0.112)}{(12250)(0.000067) + 0.888(0.112)} \\ &= 1330.59 \end{aligned}$$

caution: rare items

The gas company adopts a policy to estimate proportions with a bound of $B = 0.05$ on the error of estation (95% confidence level assumed).

Suppose the gas company operates $N = 10000$ of a certain type of valve. They want to estimate the proportion that are "stuck" open (which is very dangerous).

So they use the sample size formula with the maximal $p = 0.5$ to obtain a required sample size of $n = 384.652$.

They collect a sample of that size and find $k = 1$ (!) stuck valve in the sample.

With some concern they calculate a confidence interval and get $[-0.00232, 0.007175]$. They contact a statistician for assistance.