STA304

Neil Montgomery

2016-01-25

simple random sampling continued

"bound on the error of estimation" - I

My own preference is to express estimates in terms of 95% confidence intervals, such as (in the case of estimating μ with $\hat{\mu} = \bar{y}$:

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})}$$

which comes from the following approximation:

$$P\left(-2 \leq \frac{\bar{y} - \mu}{\sqrt{\hat{V}(\bar{y})}} \leq 2\right) \approx 0.95$$

"bound on the error of estimation" - II

But there's nothing wrong with reëxpressing the probability as:

$$P\left(|\bar{y} - \mu| < B\right) \approx 0.95$$

calling B a *bound on the error of estimation* with $B = 2\sqrt{\hat{V}(\bar{y})}$

This bound is with probability 0.95. The book calls it "the bound" but it is really "a bound".

Textbook questions tend to ask for an estimate along with such a bound. It's equivalent to finding the 95% confidence interval.

We'll make free use of either approach.

discussion: is $\bar{y}$ the "best" estimator *under SRS*?

It depends on the "rules", and is (was?) an area of statistical research.

A common rule is: among all the unbiased estimators, pick the one with the smallest variance. However, under SRS $\bar{y}$ is the *only* unbiased estimator.

You can get lower variance, but to do so you'll have to move away from SRS. Much of the course will be spent on other sampling designs.

estimation of population total

Recall the population total τ is related to the population mean μ through the obvious $\tau = N\mu$. The following formulae and results are then *immediate* from the ones we got last week:

$$\begin{aligned}
\hat{\tau} &= N\overline{y} \\
E\left(\hat{\tau}\right) &= N\mu = \tau \qquad \text{(unbiased)} \\
V(\hat{\tau}) &= N^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad \text{(nice theory)} \\
\hat{V}(\hat{\tau}) &= N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right) \qquad \text{(useful)}
\end{aligned}$$

along with the usual $\hat{\tau} \pm 2\sqrt{\hat{V}(\hat{\tau})}$ or $B = 2\sqrt{\hat{V}(\hat{\tau})}$.

disappointing example (continued)

Consider again the dentist, his toothpaste, and that population of $N = 1000$ schoolchildren. The numbers of cavities in the 10 children (from the table) were:

```
##  [1] 0 4 2 3 2 0 3 4 1 1
```

For estimating μ we had $\bar{y} = 2$ with standard error $\sqrt{\hat{V}(\bar{y})} = 1.483$. A *bound on the error of estimation* is then $B = 2.966$.

The total number of cavities amongst the childen is then simply estimated as $\hat{\tau} = 2000$ with standard error $\sqrt{\hat{V}(\hat{\tau})} = 1483$

sample size selection (means and totals)

An important part of a sampling plan is to choose the sample size.

There are two (arbitrary) choices to make:

1. How close to the true mean would we like to be (probably)?
2. With what probability would we like to be that close?

The first is related to the bound B on the error of estimation, and the second is related to confidence level. We'll fix a confidence level of 95%. The bound B is strictly a matter of choice.

sample size selection for estimating a mean - I

We would like to be within B of the true mean with probability 0.95. The sample size formula is based on:

$$P\left(|\bar{y} - \mu| < B\right) \approx 0.95$$

and we know $B = 2V(\bar{y}) = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$ is a solution to this equation.

(Note that the "2" comes from $z_{0.025} = 1.96$ which solves the equation $P(Z \leq z_{\alpha/2}) = 0.95$ and if someone really wanted a different level of confidence, the next formula will be slightly different.)

sample size selection for estimating a mean - II

Solving for n gives:

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

There is a practical problem. We don't know σ^2. And we can't do the old "let's use the sample to estimate it" thing, because we don't yet have a sample. Some possible practical solutions:

- use s^2 from a previous similar sample
- perform a "pilot sample" - a small preliminary sample conducted exactly for this (and possible other) preliminary estimates
- use a rough estimate based on prior knowledge of the range of "most" values

rough estimate of σ

A Normal distribution (perhaps common in practice) has 95% ("most values") of its probability between two standard deviations of its mean, i.e. $\mu \pm 2\sigma$. Or expressed another way, about 95% of the probability is contained within a range that is 4σ wide.

If we have a good feeling that "most values" (say, 95% or so) lie inside a certain *range*, we can use the following guesstimation:

$$\begin{aligned} 4\sigma &\approx \text{range} \\ \sigma &\approx \frac{\text{range}}{4} \end{aligned}$$

cavity example (continued)

Suppose now a public health department wants to further study the disgraceful state of local children's teeth. What sample size should be used to estimate the mean number of cavities to within a 0.1 bound on the error of estimation (95% confidence level implied, as always)?

Solution 1 - use the data from the previous study undertaken by the dentist in 4.19. The estimate for σ is $\sqrt{\hat{V}(\bar{y})} = 1.483$. Plug this into the formula to get:

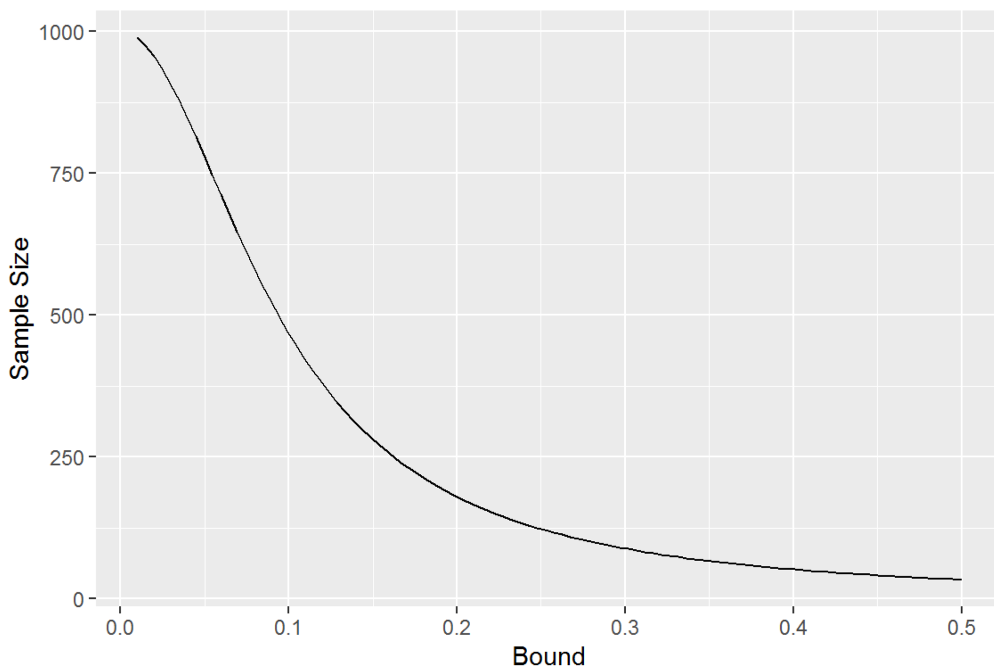
$$n = \frac{N\sigma^2}{(N-1)B^2/4 + \sigma^2} = \frac{(1000)(2.199289)}{(999)(0.1)^2/4 + 2.199289} = 468.254$$

Solution 2 - use the dentist's "gut" feeling that "most" children have between 0 and 5 cavities. Use $\sigma \approx (5-0)/4 = 1.25$ in the formula. This time you get $n = 384.852$

cavity example - some additional commentary

The range/4 guesstimator was perhaps not ever going to work very well, since the guesstimation is based on a Normal distribution.

Also, here is a plot of the dependency of $B = 0.1$ on the sample size required:



sample size selection for estimating a total

As usual, the methods for population total follow from the methods for population mean. In this case the desired , call it now B_τ, is simply divided by N and used as in the previous formula with $B = \frac{B_\tau}{N}$.