STA304

Neil Montgomery

2016-01-21

properties of $\overline{y}$

The variance of $\overline{y}$ is a little more involved due to the lack of independence:

$$\begin{aligned}
V(\overline{y}) &= V\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n^2} V\left(\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} V(y_i) + \sum_{i \neq j} Cov(y_i, y_j)\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\frac{-\sigma^2}{N-1}\right]
\end{aligned}$$

properties of $\overline{y}$

$$\begin{aligned} V(\overline{y}) &= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\frac{-\sigma^2}{N-1}\right] \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) \\ &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \end{aligned}$$

a "fun" digression back to $Cov(y_i, y_j)$

I said the derivation was dull and tricky. Here's a quicker method. From two slides ago:

$$V(\overline{y}) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(y_i) + \sum_{i \neq j} Cov(y_i, y_j)\right]$$

By symmetry in the sample, all the y_i and y_j have the *same* covariance, say, $Cov(y_1, y_2)$. So:

$$V(\overline{y}) = \frac{1}{n^2}\left[n\sigma^2 + n(n-1)Cov(y_1, y_2)\right]$$

This is true for all n. Including…$n = N$ (!). But what is $V(\overline{y})$ when $n = N$?

See, I told you that was going to be fun!

flashback: stats 101 confidence interval

- From the beginners' course it goes like this. You have a "population" $X \sim N(\mu, \sigma^2)$ and you gather a "sample" $\{X_1, X_2, \ldots, X_n\}$ "i.i.d." $N(\mu, \sigma^2)$ and then you get:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- But you don't know σ^2 so you guess it with S^2 and then:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- Unwrap this formula to get the $(1 - \alpha) \cdot 100\%$ confidence interval:

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \qquad \left(\text{jargon: } S/\sqrt{n} \text{ is the standard error of } \overline{X}\right)$$

back to this course

- The setup (and notation!!!) is quite a bit different but we're still stuck at the same step: $V(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

- In a non-shocking move that surprised nobody we'll estimate σ^2 with the *sample variance*:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

- But (this is a little tricky) the estimated variance of $\overline{y}$ will be defined as:

$$\hat{V}(\overline{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right),$$

  which is a random variable, with a distribution, a mean, a variance, etc.

the mean of $\hat{V}(\bar{y})$ — I

Preliminary result concerning $(n-1)s^2$ (for convenience):

$$\begin{aligned}
E\left((n-1)s^2\right) &= E\left(\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right) \\
&= E\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \quad \text{(standard trick)} \\
&= \sum_{i=1}^{n} E\left(y_i^2\right) - nE\left(\bar{y}^2\right) \\
&= nE\left(y_1^2\right) - nE\left(\bar{y}^2\right) \quad (y_1 \text{ as generic } y_i)
\end{aligned}$$

the mean of $\hat{V}(\bar{y})$ — II

Another standard trick: $V(X) = E\left(X^2\right) - E(X)^2$ therefore $E\left(X^2\right) = V(X) + E(X)^2$.

$$\begin{aligned}
E\left((n-1)s^2\right) &= nE\left(y_1^2\right) - nE\left(\bar{y}^2\right) \\
&= n\left[\left(V(y_1) + E(y_1)^2\right) - \left(V(\bar{y}) + E(\bar{y})^2\right)\right] \\
&= n\left[\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) + \mu^2\right)\right] \\
&= n\sigma^2\left[1 - \frac{1}{n}\left(\frac{N-n}{N-1}\right)\right] \\
&= n\sigma^2\left[\frac{n(N-1) - N + n}{n(N-1)}\right] = \frac{N}{N-1}(n-1)\sigma^2
\end{aligned}$$

the mean of $\hat{V}(\overline{y})$ — III

So:

$$E\left(s^2\right) = \frac{N}{N-1}\sigma^2.$$

Therefore:

$$\begin{aligned} E\left(\hat{V}(\overline{y})\right) &= E\left(\frac{s^2}{n}\left(\frac{N-n}{N}\right)\right) \\ &= \frac{N}{N-1}\,\frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) \\ &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = V(\overline{y}) \end{aligned}$$

the mean of $\hat{V}(\bar{y})$ — IV

Conclusion: $\hat{V}(\bar{y})$ is unbiased for $V(\bar{y})$.

In fact it was defined presicely to acheive unbiasedness.

Summary: $\bar{y}$ is a random variable with mean μ and variance $V(\bar{y})$, which depends on σ^2, so we will use the specically constructed $\hat{V}(\bar{y})$ to estimate it.

In other words $\sqrt{\hat{V}(\bar{y})}$ is the *standard error* of $\bar{y}$.

$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$ is very close to $\frac{s^2}{n}$ when n is small compared to N. This perhaps suggests a concept of "approximate independence".

$\frac{N-n}{N} = 1 - \frac{n}{N}$ is called the *finite population correction*.

the distribution of $\overline{y}$ — I

- The distribution of $\overline{y}$ is simple enough in theory—discrete uniform over all possible values of $\frac{1}{n}\sum_{i=1}^{n} y_i$, but since we don't actually know the values in the population $\{y_1, y_2, \ldots, y_N\}$ this isn't very helpful.
- Recall the *Central Limit Theorem*. Suppose $X_1, X_2, X_3, \ldots$ is an i.i.d. sequence of random variables with mean μ and variance σ^2. Then for all u:

  $$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u\right) = P(Z \leq u)$$

  where $Z \sim N(0, 1)$.
- And the convergence is really fast, so the practical impact is that for n "large enough" $\overline{X}$ is "approximately" Normal (this is what many people informally call the "Central Limit Theorem".)

the distribution of $\bar{y}$ — II

So we're saved. Apply the "Central Limit Theorem". As long as n is large, $\bar{y}$ will be approximately Normal. Obviously the larger n the better.

OK, well not quite (in theory), but it seems like we really are saved, except possibly from our own stupidity.

There are lots of "Central Limit Theorems" for various situations, including for sampling from finite populations.

As long as n is large enough, and not too close to N which is also large, the following will hold:

$$\bar{y} \sim^{approx} N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$$

the distribution of $\bar{y}$ — III

An equivalent approximation:

$$\frac{\bar{y} - \mu}{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}} \sim^{approx} N(0, 1)$$

Also true:

$$\frac{\bar{y} - \mu}{\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}} \sim^{approx} t_{n-1},$$

but we'll use the fact the $N(0, 1) \approx t_{n-1}$ for n large enough and dispense with having to use t tables as well.

confidence interval for μ

- Confidence intervals pretty much always look like this: *Estimate +/- "2" times Standard Error*, assuming we are sane and are producing a 95% interval and not trying to look clever by using 90% or 99%.
- Example from Stats 101: $\overline{X} \pm t_{n-1,0.025} \frac{s}{\sqrt{n}}$
- The technically correct value of "2" might be 1.96 (for N(0,1)), 2.021 (for t_{39}) and so on. But we'll just use 2 as a value for "2", because frankly these are all approximations anyway.
- We have all the building blocks. Our basic interval will simply be:

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

- *The Book* calls $2\sqrt{\hat{V}(\bar{y})}$ *the bound on the error of estimation*.

example 4.19

"A dentist was interested in the effectiveness of a new toothpaste. A group of N = 1000 schoolchildren participated in a study. Prestudy records showed there was an average of 2.2 cavities every six months for the group. After three months of the study, the dentist sampled n = 10 children to determine how they were progressing on the new toothpaste. Using the data in the accompanying table, estimate the mean number of cavities for the entire group and place a bound on the error of estimation."

The numbers of cavities in the 10 children (from the table) were:

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

The estimated mean is 2 and s^2 is 2.222, so the standard error is $\sqrt{\frac{2.222}{10}\frac{990}{1000}}$ = 1.483.