STA304

Neil Montgomery

2016-03-03

stratified sampling: proportions and totals

back to the basic analysis

The analysis of a stratified sample comes down to weighted combination of a few simple random samples.

$$\bar{\ } \qquad \sum^{L} W \bar{\ }$$

$$
\begin{aligned}
y_{st} &= \sum_{i=1} w_i y_i \\
\hat{V}(\overline{y}_{st}) &= \sum_{i=1}^{L} W_i^2 \hat{V}(\overline{y}_i) \\
\hat{\tau} &= N\overline{y}_{st} \\
\hat{V}(\hat{\tau}) &= N^2 \hat{V}(\overline{y}_{st})
\end{aligned}
$$

(The last two lines are a little different in the book…but obviously equal.)

counts and proportions (stratified)

The weight concept W_i is the same.

So just replace $\overline{y}_i$ with $\hat{p}_i$ and use the proportion version of the $\hat{V}$ formula, which is:

$$\hat{V}(\hat{p}_i) = \frac{\hat{p}_i \hat{q}_i}{n_i - 1} \frac{N_i - n_i}{N_i}$$

where $\hat{q}_i = 1 - \hat{p}_i$.

proportion example - transformers

For example, in the transformer dataset we've been using (stratified by the `Size` variable), let's estimate the proportion manufactured by Nema and provide a 95% confidence interval. Here is an overall summary of the situation:

Size	N	W	n	p_Nema	V-hat-p_i	W*p_Nema	W^2*V-hat-p_i
100KVA	6732	0.2587340	155	0.3032258	0.0013317	0.0784548	0.0000891
50KVA	9882	0.3797994	228	0.3289474	0.0009458	0.1249340	0.0001364
75KVA	9405	0.3614666	217	0.2580645	0.0008620	0.0932817	0.0001126

The confidence interval is 0.2966705 ± 0.0367808

count example - transformers

The company wants permission to spend money to replace all the transformers that are over 50 years old. What should be budget be for this project? What error bound can we put on the budget?

We need to count of transformers over 50 years old. There is an `Age` variable in the data. Here is a summary of the situation:

Size	N	W	n	p_old	V-hat-p_i	W*p_old	W^2*V-hat-p_i
100KVA	6732	0.2587340	155	0.0709677	0.0004156	0.0183618	2.78e-05
50KVA	9882	0.3797994	228	0.1710526	0.0006076	0.0649657	8.76e-05
75KVA	9405	0.3614666	217	0.0967742	0.0003935	0.0349806	5.14e-05

The population size is 26019. So the estimated count is 3078.2582343 and the usual bound on the error of estimate is 672.2239136.

Convert to dollar amounts (for the budget) by multiplying by the unit cost.

counts/proportion stratified sample size

Use essentially the same formula as before, but the population variance is now $p_i q_i$. Given an allocation a_i:

$$n = \frac{\sum_{i=1}^{L} N_i^2 p_i q_i / a_i}{N^2 B^2/4 + \sum_{i=1}^{L} N_i p_i q_i}$$

Better is to divide through by N^2:

$$n = \frac{\sum_{i=1}^{L} W_i^2 p_i q_i / a_i}{B^2/4 + \frac{1}{N} \sum_{i=1}^{L} W_i p_i q_i}$$

And p_i is unknown and must be guessed, using the usual proportion guessing guidelines (use known information closest to 0.5).

example - count stratified sample size

The electricity regulator demands a bound on the error of estimating the true proportion of 50+ year old transformers to be no more than 500 units. What is the sample size required to fulfil this requirement?

First, change the bound from a "count" requirement to a"proportion" requirement, which is $500/26019 = 0.0192167 = B$.

The company decides on proportional allocation among the size ratings. We'll pretend that previous sample never happened and suppose that the company thinks between 10% and 20% of transformers are over 50 years old.

Then the formula reduces (noice-ly) to:

$$n = \frac{pq}{B^2/4 + \frac{1}{N}pq}$$

in which we can use the guess of $p = 0.2$. The sample size required is 1624.86.

optimal allocation

The optimal allocation formula also stays the same with σ_i replaced with $\sqrt{p_i q_i}$, becoming:

Textbook formula (bad for hand calculation):

Textbook formula (bad for hand calculation):

$$a_i = \frac{N_i\sqrt{p_i q_i}\Big/\sqrt{c_i}}{\sum_{k=1}^{L} N_k\sqrt{p_k q_k}\Big/\sqrt{c_k}}$$

Better for hand calculation is to divide by N to get:

$$a_i = \frac{W_i\sqrt{p_i q_i}\Big/\sqrt{c_i}}{\sum_{k=1}^{L} W_k\sqrt{p_k q_k}\Big/\sqrt{c_k}}$$

example count optimal allocation

This is really stretching the story, but suppose for some bizarre reason `100KVA` transformers are more costly to sample. Say they cost $10 each while the other two sizes cost $5 each. Here is a summary of the situation:

Size	N_i	W_i	p_i*q_i	c_i	W_isqrt(p_iq_i/c_i)
100KVA	6732	0.2587340	0.16	10	0.0327276
50KVA	9882	0.3797994	0.16	5	0.0679406
75KVA	9405	0.3614666	0.16	5	0.0646611