Sampling for Variance in a Population

Stacey, Brian

Southern New Hampshire University

18 March 2015

Sampling for Variance in a Population

Brian Stacey

Southern New Hampshire University

Determining confidence intervals for a population μ estimate from a sample $\bar{x}$ with a either a known or an unknown population σ use similar methods. The more data one has to begin with the better will be the estimate, so knowing the standard deviation of the population will provide a better estimate of the mean. In our case we have no data to begin with so must collect the data to build our estimates. This eliminates the possibility of estimating the confidence interval of μ with a known σ, which is the better of the estimates, and has a narrower range. The data collection and calculation of s and $\bar{x}$ will be necessary in our case. Once those two estimators are calculated, finding the confidence interval is an easy task.

$$\bar{x} \pm t\alpha/_2 \frac{s}{\sqrt{n}}$$

Where $t\alpha/_2$ is obtained from a Student's t distribution table since we are looking for a 95% confidence, α=.05 (α and confidence being complements), and α/2=.025. The Student's t tables are arranged by degrees of freedom and confidence, we will also need degrees of freedom; in this case simply n-1. Assuming 50 stores, d.f.=49, thus making $t\alpha/_2$=2.010.

The student's t distribution is very similar to the normal distribution, however is not as peaked and has fatter tails, this results in an overestimation if used in place of a z-score, a useful thing when we don't have exact values for our calculations.

Choosing the number of outlets to sample to obtain the required data is only half of the sampling problem. Deciding which of the outlets to use is as important as how many. Fundamentally this is a choice about variance and bias.

Bias is the difference between our result and the true population parameter (in this case mean) that is a result of choosing poor sample points, e.g. convenience over true randomness; whereas variance is the difference between our estimate and the parameter resulting from randomly occurring differences in the population that is evident in our sample. Think of bias as accuracy and variance as precision. An estimate may have any combination of high or low bias and/or variance.

Variance can be whittled away with increased sample size. Imagine rolling a single die, if you roll the die an infinite number of times the results will be evenly split between 1-6 (a uniform distribution), however if you only roll the die six times the probability that all six numbers will appear (each once) is only 1.5%; our sample lies somewhere between trying once and trying an infinite number of times. The larger the sample size the more indicative of the population it will be.

Bias can only be reduced through proper sampling technique, and even then may still occur. In our case, we can eliminate the possibility of non-participation as a source for bias, and must instead focus on initial selection. For our purposes the best option for eliminating bias in sample selection is to select a random set of our stores. We will assign a number to each store and have Excel select a set of n random numbers, the numbers selected will correspond to the stores and those will be the stores we obtain our data from. The number required comes from:

$$n = (\frac{z\sigma}{E})^2$$

Where E=desired error, and z is the appropriate z score for a 95% confidence interval (1.96). Obviously a smaller desired sample size requires a larger allowable error which results in a larger interval ($\overline{x} \pm z\frac{\sigma}{\sqrt{n}}$). Since we don't know the population standard deviation we must estimate it. There are several methods of doing so and most rely on having data of some sort. Assuming that we do not have any data to begin with, we will take a small initial sample and calculate the sample standard deviation (s) from that sample. We then use s in place of σ in our sample size calculation. This sounds like circular logic, and to a degree it is, but it is only to provide us a starting point. If we had pre-existing previous data we could also use that to estimate the upper and lower bounds and set $\sigma=[(b-a)^2/12]^{1/2}$ for a uniform distribution or $\sigma=(b-a)/6$ for a normal distribution.

References

Chiang, A. (1984). *Fundamental methods of mathematical economics* (3rd ed.). New York: McGraw-Hill.

Fortmann-Roe, S. (2012, June 1). Bias and Variance. Retrieved March 15, 2015, from http://scott.fortmann-roe.com/docs/BiasVariance.html

Gutierrez, D. (2014, October 22). Ask a Data Scientist: The Bias vs. Variance Tradeoff – inside BIGDATA. Retrieved March 15, 2015, from http://insidebigdata.com/2014/10/22/ask-data-scientist-bias-vs-variance-tradeoff/

Halls-Moore, M. (2015, February 25). The Bias-Variance Tradeoff in Statistical Machine Learning - The Regression Setting. Retrieved March 15, 2015, from http://www.quantstart.com/articles/The-Bias-Variance-Tradeoff-in-Statistical-Machine-Learning-The-Regression-Setting