STATS STARTER

Sampling

Sample Size Calculation: Precision-Based

Difficulty ★★

Brought to you by:
Joe Zemmels,
CSAFE Graduate Student

Introduction

Sampling: the process of selecting a subset of items from a population to learn about a quantity that describes a characteristic of the population.

Sampling an entire population can be costly, time-consuming, or dangerous for the sampler. If we're willing to tolerate some uncertainty, then we can often construct a sample that is considerably smaller than the population. We will discuss one approach for calculating sample size with the goal of estimating a population characteristic within a set precision.*

Concept Overview

Testing a lot of seized bags for suspected illicit drugs can be costly and time-consuming. Searching an entire hard drive for suspected Child Abuse Material (CAM) can be both time-consuming and distressing for the examiner [1]. Sampling can be applied in these situations to avoid analyzing the entire population of items.

An important consideration when using sampling is the number of units one should take from the population. For example, is testing 2 of 100 bags of suspected Cocaine too few accurately estimate the total number of bags containing Cocaine? The answer to this question relies largely on the goals of the analysis. Two mathematically-principled approaches for calculating sample size are *precision-based* and *model-based*.

When should I use a precision-based approach?	**When should I use a model-based approach?**
Desire to estimate a population characteristic within some tolerance. Examples: • Out of 100 bags, estimate the number containing Cocaine within 1 bag from the truth. • Estimate the proportion of 10,000 images on a hard drive containing CAM within 5% of the truth.	Determine whether the number/proportion of positive cases in a population exceeds some threshold. Examples: • Determine whether at least 90 out of 100 bags contain Cocaine. • If at least 10% of a hard drive contains CAM, then the defendant will be charged with a greater sentence.

We will consider the precision-based approach on the next page. You can find more information on the model-based approach in the *Sample Size Calculation: Model-Based* Stats Starter.

* Many techniques and guidelines exist for performing both random and non-random sampling [2,3,4,5]. Here, we assume that a random sampling procedure has already been selected for use and a sample size needs to be calculated.

Theory

Depending on whether estimation of a whole number (e.g., total number of bags) or a proportion (e.g., proportion of images on a hard drive) is of interest, there are two formulas used in precision-based approach to sample size calculation We will first define the terms used in the formulas:

- D: the allowable difference or tolerance between the estimate and the true population value.
- σ_0^2: a prior estimate of the population variance.
- p_0: a prior estimate of the population proportion.
- n: the estimated sample size. This is the unknown (or dependent) variable in the formula.

Prior estimates of the population variance or proportion are most commonly obtained using results from previous, similar casework. The precision-based sample size formulas are given by:

Whole Number Estimation

$$n = \frac{9\sigma_0^2}{D^2}$$

Proportion Estimation

$$n = \frac{9p_0(1-p_0)}{D^2}$$

Application

Consider an example of estimating the proportion of images on hard drive containing CAM. Suppose we want our estimate to be 1% away from the truth, but we *don't* have a good prior estimate of the population proportion. A conservative approach is to find the value of p_0 that maximizes $p_0(1-p_0)$. This happens to be $p_0 = 0.5$. Using the formula above, the sample size is:

$$n = \frac{9 * 0.5 * (1-0.5)}{(0.01)^2} = 22,500$$

We can see that the sample size is quite large. It's important to note that we intentionally chose a conservative prior population proportion estimate and a small tolerance. Both of these values are likely to differ in actual casework. In-general, precision-based approaches result in larger sample sizes than, for example, model-based approaches.

Learn More

To learn more, we recommend the CSAFE Sampling Short Course. For more information on the formulas presented here, refer to the following references.

[1] Brian Jones, Syd Pleno, Michael Wilkinson, The use of random sampling in investigations involving child abuse material, Digital Investigation, Volume 9, Supplement, 2012, Pages S99-S107, ISSN 1742-2876, https://doi.org/10.1016/j.diin.2012.05.011.
[2] European Network of Forensic Science Institutes. ENFSI DWG Qualitative Sampling Calculator – Revision July 2017. 20 July 2018. Microsoft Excel File. https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-DWG-Qualitative-Sampling-Calculator-Revision-July-2017.xls2007-2016.xlsm
[3] ASTM E122-17, Standard Practice for Calculating Sample Size to Estimate, With Specified Precision, the Average for a Characteristic of a Lot or Process, ASTM International, West Conshohocken, PA, 2017, www.astm.org
[4] ASTM E2548-16, Standard Guide for Sampling Seized Drugs for Qualitative and Quantitative Analysis, ASTM International, West Conshohocken, PA, 2016, www.astm.org
[5] United Nations Office on Drugs and Crime. Guidelines on Representative Drug Sampling. March 2011. https://www.unodc.org/documents/scientific/Drug_Sampling.pdf

Funding

CSAFE is a publicly funded organization headquartered at Iowa State University. The National Institute of Standards and Technology (NIST) is one of the center's providers, supporting CSAFE as a nationally recognized Center of Excellence in Forensic Sciences, NIST Award #70NANB15H176 and #70NANB20H019.

forensicstats.org
csafe@iastate.edu