Sampling and Sample Size

Goals

- Get an overview of sample size calculation
  - Binary worst-case method
  - Real Value Precision method
  - Critical Differences
  - Simulation
- Sampling in general

Why Sample Size is Odd

- How do I know how many observations I will need in order to identify population values?
- Ideally, you need to know the mean, variance and covariance of all the variables.
- If you know this, why do you need a survey?

Your sample size calculations are wrong to the extent that you need to give a survey.

The Worse Case

- Sample size calculations are often dependent on the worst case.
- Look at what you are trying to decide and spot the hardest hypothesis test.
  - Is it determining if a value is significantly different from zero?
  - Is it determining if two sub-groups given different answers?
  - Is there are legal or contractual required precision?

The Binary Worst Case

Suppose you are looking at the fraction of respondents that answer yes to a question. The standard error for a proportion is:

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$

- More observations, n, reduces the standard deviation.
- The numerator $P(1-P)$ is maximized at $P = .5$.
  - $.5(1-.5) = 0.25$
  - $.75(1-.75) = 0.1875$
- Assume the worst case when calculating required sample size.

The Binary Case: How Precise?

Precision is usually stated with two numbers. The probability of the estimate being within x percent of the population value.

- 90/10 Means you are 90% sure that the true population value is within 10% of the estimated value.
- 95/10 Means you are 95% sure that the true population value is within 10% of the estimated value.

The Rough Approximation

- Calculate the standard deviation of the sample proportion
  - ± 1.96 times that is a 95% confidence interval
  - ± 1.645 times that is a 90% confidence interval
  - ± 2.576 times that is a 99% confidence interval

Just multiply by two to approximate a 90% confidence interval.

Examples

- $\sqrt{\frac{.5(1-.5)}{100}} = 0.05$, precision $90/10$
- $\sqrt{\frac{.5(1-.5)}{400}} = 0.025$, precision $90/5$
- $\sqrt{\frac{.5(1-.5)}{1000}} = 0.0158114$, precision $90/3$
- $\sqrt{\frac{.5(1-.5)}{10000}} = 0.005$, precision $90/1$

Observations increases with the square of precision.

Tips

- This is why it is important to get a decision criteria first.
- Precision is expensive.
- If you just want to check if it is *different* than 40%, you only need 100 observations.
  - If you want to check if it is different than 45% you need 400
  - 47% needs 1,000
  - 49% needs 10,000

What About Real Valued Variables?

- These are things like age, GPA, etc.
- You need a guess at both the average and the variance.

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

The higher the standard deviation of the variable the higher the standard deviation of the mean. The more observations, the lower the standard deviation from the mean.

Suppose the standard deviation is 10

- $\frac{10}{\sqrt{100}} = 1$
- $\frac{10}{\sqrt{400}} = 0.5$
- $\frac{10}{\sqrt{1000}} = 0.3162278$
- $\frac{10}{\sqrt{10000}} = 0.1$

You can detect differences that are equal to twice that with the noted number of observations.

Ready to have your day ruined?

Finding the difference between two subgroups requires precision in both groups.

$$\sigma_{x_1 - x_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}$$

If you want to check if there is a difference

- Basically, you add the standard deviations for the subgroups.
- If you have subgroups, you need to make sure there are enough for precise estimate of both groups.

Wait . . . .what if I choose n_1 and n_2?

Probability Sampling

If you choose the proportions of group 1 and group 2 correctly, you can make it easier to detect differences in the two means.

$$\min_{n_1} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{N - n_1}}$$

Continued

$$\frac{\frac{\sigma_2^2}{(N-n_1)^2} - \frac{\sigma_1^2}{n_1^2}}{2\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{N-n_1}}} = 0$$

$$\frac{\sigma_2^2}{(N-n_1)^2} - \frac{\sigma_1^2}{n_1^2} = 0$$

$$\frac{\sigma_2^2}{(N-n_1)^2} = \frac{\sigma_1^2}{n_1^2}$$

$$\frac{\sigma_2}{\sigma_1} = \frac{n_2}{n_1}$$

Choose so that the observations are proportional to the standard deviation.

The Power Of Probability Sampling

- Allows you to get the most information possible out of our sample.
- More extreme: Experimental design allows you to pull out lots of information.

The Pitfall

To do a proper design you need:

- Means of the variables
- Variances of the variables
- Covariances between the variables

In short, if you know this, why survey?

Every Sampling Plan Starts with a Guess

- It should be reasonable based on knowledge of the respondents.
- The best place to look is at previous surveys.

You can do a few things

- If groups have different probability of responding, you can adjust the proportion in the sample to make the respondents more closely resemble the population.
- If you want to look at the differences between groups, you can adjust the proportion in the sample to maximize that precision.
- If you want overall more or less precision, you can adjust sample size.

Our First Survey Will Improve the Next

Terms

- Target Population: Who you want general findings about.
- Sample frame: The population in material form. A list of email address for all enrolled students for example
- Coverage rate: Fraction of population in sample frame. Some people will not be on the list. No email. No address.
- Coverage error: Not in the sample frame because of lag.
- Sample selection: How you pick.
- Sample: Who was picked.
- Completed sample/rate, completion rate, response rate: Fraction of people in the sample that completed the survey.
- Sampling error: Uncertainty because you pulled a sample instead of a census.
- Sampling weights: How you fixed a bias you purposefully put in sample selection or bias caused by different response rates.

Sample Methods Part 1

- Simple Random: Use a random number generator
- Systemic: Intercept survey, every 5th person that walks by. Odd numbered ID numbers.
- Stratified: Minimize variance within groups, maximize between groups, Works best when your strata have different responses.
- Proportional: Big units are more likely than small.

These all are generally fine and can be analyzed with weighting $\frac{P_{Pop}}{P_{Sample}}$.

More Problematic

- Quota: Pick who you want. Avoid
- Convenience: Lots of meta information on how you got people to answer. Avoid.
- Snowball: Get a group and ask them to refer to you. Requires bootstrap to analysis.
- Cluster Sampling: Randomly choose a sample of areas, blocks for example, and then randomize units within. Requires cluster stats or multi-level modeling.

These are either impossible to work with or in the case of Snowball and Cluster, much harder.

Practical

- Find an old survey from the same or similar population.
- You can use this a lot of ways.
  - Sample with replacement to observe the consequences of increasing and decreasing sample size.
  - Do as we did and create a probit model for participation. You can use this to stratify.

Sample with Replacement

Sample from the old survey data *with replacement*.

- See how changing the number sampled changes your variance estimates.
- There should be no change in the mean values

Using Old Data to Stratify

- Create a participation model, probit, that uses only the ex-ante data in both surveys.
- For every item in sample frame, estimate the probability of responding and create a, 1/p, weight.
- Normalize so that $\sum 1/p = N$
- Draw a stratified sample with this.

When the New Data Arrives.

Estimate a new participation model with same functional form and analyse with the new weights.

If you Don't have Old Data

- Find what kind of ex-ante data you have.
- The useful stuff is correlated with questions in the survey.
  - We used GPA, Race/Ethnicity and Total credits.
  - Income is common but we didn't have it.
- Increase probability of selection for:
  - Rich and Poor. They have low response rates.
  - Small sub-groups. You want to make sure there are enough observations for reasonable mean values.

Then Post-Sample Stratification

- Undo with $\frac{P_{Sample\ \ Frame}}{P_{Respondent}}$
- Analyze with these new weights.