GALLUP CONSULTING

World Bank: Listen to Latin America Project L2L

Power Analysis - Report

In statistics one is trying to make inferences about an entire population with only information about a sample of it. One concern is being able to detect differences between groups in the population given differences observed between those same groups in the sample. These groups are based on different treatments or conditions, such as those interviewed via SMS and those interviewed using CATI. The differences of interest might be the proportion of who respond "yes" to a particular question or the average number of years of education of the respondents. A test of statistical significance is performed in order to see if differences in these proportions or mean between the sample groups are large enough to infer that these differences actually exist between the groups in the population.

In a test of statistical significance there is a null hypothesis that predicts that there is no actual difference in proportions or means between the groups in the population, given the difference observed between the sample groups. There is also an alternative hypothesis that predicts that there in fact is a difference between the groups in the population. The power of a statistical test is the probability that the test will accurately detect differences between the two population groups, given differences observed between the sample groups. In other words, the test will reject the null hypothesis when the null hypothesis is false. The power is also known as the sensitivity of the test.

There are two types of error that can occur when conducting a test of statistical significance. Type I error, generally denoted by the symbol α, is the probability of rejecting the null hypothesis when it is true. This is a false positive because the test determined that there were differences between the population groups when in fact there were not. Type II error, generally denoted by β, is the probability of accepting the null hypothesis when it is false (see the table below). This is a false negative because the test determined that there were no differences between the population groups when in fact there were. This means that β can also be thought of as the false negative rate. The power of a test is equal to $1 - \beta$. So as β decreases, meaning the probability of a type II error decreases, the power of the test increases.

In statistical tests the p-value is the probability that the difference between two groups, captured by the test statistic calculated, has a value as extreme or more extreme than the one found given that the null hypothesis is true. So if there is no actual difference between the proportions in the population groups, what is the probability the difference observed between the sample groups' proportions would be as extreme as it is? Typically $\alpha = 0.05$ and the null hypothesis is rejected for estimated p values less than or equal to 0.05. This means that if the probability of observing a difference between sample group proportions is .05 or less, it is concluded that there is in fact a difference between the proportions in the population groups; the null hypothesis that there is no difference is rejected. A type I error occurs then when the p-value is less than .05, so the null hypothesis is rejected, but the null hypothesis is in fact true.

		State of World	
		H_0 True	H_0 False
Research Result	H_0 True	OK	Type II Error (β)
	H_0 False	Type I Error (α)	OK

The power of a test is determined by the sample size of the test, the statistical significance of the test (α) and the effect size. The effect size is the size of difference between two groups. This could be measured by the difference in proportions or means of each of the groups. For example, as mentioned above, this could be the difference in the proportion responding "yes" to an item or the difference in the mean years of education. Alternatively the effect size of a test comparing multiple groups at once, such as an analysis of variance (ANOVA), would be similar to the F value of each treatment in the ANOVA. Power tests are often calculated in order to ascertain what the minimum effect size, or difference between the groups in the sample, would need to be in order to detect actual differences between the groups in the population. As the sample size increases, the power of the test increases because differences between the population groups can be detected by smaller effect sizes, or differences between the sample groups.

The power calculations (Lenth, 2006-9) below compare incentive groups within each data collection mode group in Peru and Honduras. Within each data collection mode group, $0, $1 and $5 incentives are offered. For the sake of this exercise comparisons are made between the $0 incentive group and one other treatment group. As the n sizes are nearly identical for each treatment group, $1 or $5, the results of the power analysis are the same

for either treatment group used. Again the significance level $\alpha = 0.05$ is used and the power was as close to .8 as possible given the software limitations.

Peru SMS Group		
Treatment	Sample Size	Proportions
$0	233	0.465
$1 or $5	233	0.6
	Difference	0.135
	Significance Level (α)	0.05
	Power (1 - β)	0.8094

Peru IVR and CATI Groups		
Treatment	Sample Size	Proportions
$0	133	0.465
$1 or $5	133	0.645
	Difference	0.18
	Significance Level (α)	0.05
	Power (1 - β)	0.8119

Honduras SMS/IVR/CATI Group		
Treatment	Sample Size	Proportions
$0	200	0.465
$1 or $5	200	0.61
	Difference	0.145
	Significance Level (α)	0.05
	Power (1 - β)	0.8043

Honduras SMS Group		
Treatment	Sample Size	Proportions
$0	300	0.465
$1 or $5	300	0.585
	Difference	0.12
	Significance Level (α)	0.05
	Power (1 - β)	0.818

Given two groups of 233 the minimum effect size for a statistical test, with power of approximately .8 and an alpha of .05, is 13.5%. Under the same conditions, the minimum effect size given two groups of 133 is 18%, between two groups of 200 is 14.5% and between two groups of 300 is 12%.

Disregarding incentive groups, comparisons between modes can be made in each of the countries. In Peru this would be between the IVR, CATI and SMS groups. The IVR and CATI groups both have an initial n size of 400, while the SMS group has an initial n size of 700. In Honduras this would be between the group receiving the CATI, IVR and SMS modes versus the group receiving only the SMS mode. The former has an initial n size of 600 while the latter has an n size of 900. The significance level $\alpha = 0.05$ is used and the power was as close to .8 as possible given the software limitations.

Peru IVR vs. CATI Groups		
Treatment	Sample Size	Proportions
IVR	400	0.5
CATI	400	0.602
	Difference	0.102
	Significance Level (α)	0.05
	Power ($1 - \beta$)	0.8089

Peru IVR or CATI vs. SMS Groups		
Treatment	Sample Size	Proportions
IVR or CATI	400	0.5
SMS	700	0.59
	Difference	0.09
	Significance Level (α)	0.05
	Power ($1 - \beta$)	0.8071

Honduras SMS, IVR or CATI vs. SMS Groups		
Treatment	Sample Size	Proportions
IVE, CATI or SMS	600	0.5
SMS	900	0.575
	Difference	0.075
	Significance Level (α)	0.05
	Power ($1 - \beta$)	0.801

Given two groups of 400 the minimum effect size for a statistical test, with power of approximately .8 and an alpha of .05, is 10.2%. Under the same conditions, the minimum effect size given one group of 400 and another of 700 is 9%, and between one group of 600 and another of 900 is 7.5%.

In order to compare both incentive groups and mode groups simultaneously, a two-way ANOVA would be performed to test if there was a difference between any of the incentive groups, any of the mode groups and between the groups formed by the interaction of the two, while taking both treatments and their interaction into account. This is done by computing an F statistic, which tells us the variability in the dependent variable between levels of one treatment divided by the variability in the dependent variable not explained by any of the treatments (or their interactions). So this can be thought of as the variance of dependent variable explained by the treatment in relation to the variance of the dependent variable that cannot be explained at all. The more variance in the dependent variable explained by the differences between the levels of the treatment, the more likely that there are actually differences in the population groups that these levels represent, i.e. that there are statistically significant differences in the mean of the dependent variable between the levels of the treatment group. For the purposes of this study it could be thought of as the variability in the mean years of education explained by the mode used divided by the variability in the mean years of education not explained by the mode, incentive group or the interaction between the two. In other words, is there a statistically significant difference between the mean number of years of education for each mode, taking into account the incentives used and the interaction between mode and incentive?

In order to calculate the power of such an ANOVA an effect size for the ANOVA must be calculated. In this case the effect size is similar to the F statistic but not identical. Instead of the variability in the dependent variable explained by the treatment in relation to the variability in the dependent variable not explained at all, the effect size is the variability in the dependent variable explained by the treatment in relation to the total variability in the dependent variable. Thus an effect size must separately be calculated for each treatment in the ANOVA aside from the default F tests performed.

In the Peru experiment there is the mode treatment with three levels and the incentive treatment with three levels, as well as the interaction between the two. Note that there are actually four modes, but all respondents receive the first mode thus removing it as a treatment level. Power is computed separately for each of these treatments and for their interaction. Given a minimum cell size of 133 the minimum effect sizes that would result in a statistically significant result are displayed below. Effect sizes of this kind less than or equal to .1 are considered “small” so the following results should be looked upon favorably.

	Levels	N Size per level*	Power	Alpha	Minimum Effect Size
Mode	4	399	0.8	0.05	0.09555
Incentive	3	532	0.8	0.05	0.09525
Mode*Incentive	12	133	0.8	0.05	0.1309

*Note that this assumes all cells to have the n size of the smallest cell for conservative calculation purposes

The above is simply a cross-sectional look at time one (T2) of the experiment. This is actually a panel of respondents meaning that there are several data collection phases per respondent. With a panel comes attrition and so for a subsequent data collection phase, say T3, we will assume 50% attrition. This would reduce the per cell n-size to 66. Given that, the minimum effect sizes that would result in a statistically significant result are displayed below. As expected, the minimum effect size increases as the n size decreases, keeping the alpha and power constant.

	Levels	N Size per level*	Power	Alpha	Minimum Effect Size
Mode	4	198	0.8	0.05	0.1358
Incentive	3	264	0.8	0.05	0.1353
Mode*Incentive	12	66	0.8	0.05	0.1862

*Note that this assumes all cells to have the n size of the smallest cell for conservative calculation purposes

It is important to note that since this data is collected from a panel, comparisons within the same respondent overtime are correlated. A comparison of T2 vs. T3 is not the same as comparing 1500 respondents at T2 to 750 new respondents at T3; they are the same 750 respondents in both time periods plus an additional 750 respondents at T2 who attrited. The effect of this is to reduce the power of the test as the effective sample size is reduced due to the correlation of the data.

The above is of particular note in Honduras, as the first sub-panel of respondents all receives SMS, IVR and CATI. This cannot be analyzed as 600 SMS respondents versus 600 new IVR respondents versus 600 new CATI respondents. Independence of responses is an assumption of ANOVA and thus that test cannot be used. In this case a mixed-model would be used that treats the respondent as a random effect in order to account for the correlation among responses from each respondent. As this is a much more complicated regression based model there is not software to calculate the power of such a test but it will be less than the power of a test without correlated responses. Despite the lowered power, this sampling plan will allow each respondent to serve as their own control group.

Reference

Lenth, R. V. (2006-9). Java Applets for Power and Sample Size [Computer software]. Retrieved *August* 25, 2011 from http://www.stat.uiowa.edu/~rlenth/Power.