Sample Size / Power Considerations

Today we will briefly discuss sample size and power calculations for your studies. First we will review the basic sample size and power estimation procedures and then think about how to extend these to regression settings.

We will use data provided by Alex Krist to illustrate our calculations. The study is looking at outcomes relating to Prevnar, a vaccine recommended for infants and toddlers which guards against some pneumococcal bacteria that can cause life-threatening meningitis and blood infections. We will consider as the outcome of interest an indicator of whether children receive the recommended 3 vaccinations of Prevnar at any age (prevany: 1 if received all three vaccinations, 0 otherwise). The available predictors are:

- Aprvn1: age in months of the child at the first Prevnar vaccination

- AlwaysPCV: an indicator of whether the medical office had the vaccine available at the time the child presented for vaccine (1 if Prevnar was present at all office visits, 0 if there was at least one visit with no Prevnar)

How you determine your sample size will depend on the goal of your analysis:

1. The purpose of the study is to estimate the prevalence of children receiving the recommended 3 vaccinations of Prevnar to within some specified percentage of the true prevalence with 95% confidence (i.e. to within 5 percent of the true prevalence).

2. The purpose of the study is to compare(test) the difference in the prevalence of children receiving the recommended 3 vaccinations of Prevnar among children who had access to the vaccine at all visits verses those who did not.

3. The purpose of the study is to compare(test) the difference in the prevalence of children receiving the recommended 3 vaccinations of Prevnar among children who had access to the vaccine at all visits verses those who did not after adjusting for the age in months at the first Prevnar vaccination.

Sample Size Based on ESTIMATION:

We assume that the prevalence estimate is approximately normally distributed and we use the standard formula for a 95% confidence interval to solve for the required sample size:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm m$$

We then specify *m* and solve for n:

$$n = \frac{1.96^{2}\,\hat{p}(1-\hat{p})}{m^{2}}$$

For instance the following table provides the required sample size to estimate the true prevalence to within *m* with 95% confidence for a variety of guesses for *p*.

Guess of true prevalence	Margin of Error (*m*)		
	2 percent	5 percent	10 percent
0.10	865	139	35
0.20	1537	246	62
0.30	2017	323	81
0.40	2305	369	93
0.50	2401	385	96

Sample size based on TESTING THE DIFFERENCE IN TWO PREVALENCES:

Now we assume that the two estimates of the prevalence are approximately normally distributed. The null hypothesis is H0: p1 = p1 and the alternative is that H1: p1 ≠ p2.
So what do we need to specify to calculate our sample size?

1. The significance level of your test (α)
2. The power that you would like to achieve to detect the difference of interest (1-β)
3. The clinically important difference in p1 – p2 = Δ that you'd like to be able to detect
4. Initial guesses for p1 and p2.

Then the sample size required to detect a difference of Δ in the two prevalences at the α-level with power 1-β is determined by:

$$n = \left[z_{1-\alpha/2}\sqrt{2\overline{pq}} + z_{1-\beta}\sqrt{p_1 q_1 + p_2 q_2} \right]^2 \Big/ \Delta^2$$

where $\overline{p} = \frac{p_1 + p_2}{2}$, $\overline{q} = 1 - \overline{p}$

The table below presents required sample sizes to detect a difference of Δ in the two prevalences at the 0.05 level with 80% power, assuming a variety of prevalence values for the children who had access to Prevnar at all vaccination visits.

Prevalence among children with access to the vaccine at all visits	Scientifically Significant Difference (Δ)		
	0.05	0.10	0.15
0.25	1134	270	113
0.30	1291	313	134
0.35	1417	349	151
0.40	1511	376	165

Sample size based on TESTING THE DIFFERENCE IN TWO PREVALENCES after adjusting for additional covariates:

Now, to the more realistic problem where you want to compare the odds of completing all three vaccinations for children with access to the vaccine at all visits vs. those who did not have access to the vaccine at all visits, after adjusting for the age a first vaccination. In this case, our analysis involves building a logistic regression model:

$$Log\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = \beta_0 + \beta_1 I(access\ ) + \beta_2 Age$$

where the coefficient of interest is β_1 which is the log odds ratio of completing all three vaccinations comparing children with access at all vaccine visits to those who did not have full access to the vaccine, after adjusting for the age a first vaccination.

To determine a sample size for this problem is difficult since the sample size will depend on the variability of β_1 which is a function of the sample size but also the distribution of the other covariates.

I will present a method for determining the sample size (or power) using a simulation study. I will illustrate this method using the Prevnar data provided by Alex. First I will fit the model to the data to determine the actual value of β_1 .

```
. logistic prevany aprvn1 alwaysPCV

Logistic regression                               Number of obs   =        107
                                                  LR chi2(2)      =      13.23
                                                  Prob > chi2     =     0.0013
Log likelihood = -67.321306                       Pseudo R2       =     0.0895

------------------------------------------------------------------------------
     prevany | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      aprvn1 |   .7960748   .0599193    -3.03   0.002     .6868869    .9226192
   alwaysPCV |   1.145634   .5064119     0.31   0.758     .4817077    2.724632
------------------------------------------------------------------------------
```

After adjusting for the age of the child at the first vaccination, we estimate that the odds of complete vaccination among children with full access to the vaccine are approximately 15% greater than the odds for children without full access to the vaccine. This difference in the odds is not statistically significant, but may be clinically important. We would like to design a larger study that has power to detect this finding.

For this purpose, power is defined to be: Pr(rejecting H0: β_1 =0 | β_1 = log(1.15)).

So if we can generate datasets where β_1 = log(1.15), then we can fit the regression model and test H0: β_1 =0, and our estimate of the power is the percentage of the datasets where we reject.

How do we generate datasets where β_1 = log(1.15)? We can use the available pilot data for this purpose since we know that in the pilot data β_1 = log(1.15). We will take random samples (with replacement, also known as bootstrap samples) of various sizes larger than the pilot sample size and estimate the power for those sample sizes.

First, lets demonstrate the idea of the bootstrap sampling and power calculation using a sample size of 100 (just under the sample size of 107 for the pilot study). The "bootstrap" command in Stata will take 500 random samples with replacement ("rep(500)") from the data and for each sample Stata will perform the specified analysis ("logistic prevany aprvn1 alwaysPCV"). You then request the quantities that you want to save from the analysis, I have requested to save the regression coefficient of interest ("b = _b[alwaysPCV]") and corresponding standard error ("se = _se[alwaysPCV]"). These will be written to a new dataset so that in the end you will have 500 values of b and se that correspond to each bootstrap sample.

```
. bootstrap "logistic prevany aprvn1 alwaysPCV" b = _b[alwaysPCV] se = _se[alw
> aysPCV], rep(500) size(100) saving(sam100)

command:      logistic prevany aprvn1 alwaysPCV
statistics:   or         = _b[alwaysPCV]
              se         = _se[alwaysPCV]

Bootstrap statistics                              Number of obs    =       107
                                                  Replications     =       500

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
           b |   500  .1359579 -.0072825  .4748829  -.7970584   1.068974   (N)
             |                                       -.790884   1.010838   (P)
             |                                      -.7683083    1.03025  (BC)
          se |   500  .4420365  .0313302  .0288445   .3853647   .4987082   (N)
             |                                       .4300197   .5389925   (P)
             |                                       .4126742   .4589331  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

The estimated power is calculated as follows:

```
. use "C:\DATA\sam100.dta", clear
(bootstrap: logistic prevany aprvn1 alwaysPCV)

. gen z = b/se

. gen reject = 1 if abs(z) > 2
(471 missing values generated)

. tab reject

     reject |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         29      100.00      100.00
------------+-----------------------------------
      Total |         29      100.00
```

The estimated power to detect the odds ratio of 1.15 in a sample of 100 children is 29/500 = 5.8%.

Now, we'd like to increase that power, so we will need to consider larger sample sizes. Unfortunately, Stata does not perform bootstrap samples of sizes greater than the original sample size, so the data will be imported into R (another statistical package, which is FREE!) and the power analysis will be performed there. The R commands are provided below if you are interested….

```
data = read.table("c:/Elizabeth/Regression Course/Materials 2004/power.csv",sep=",",header=T)
dim(data)
[1] 168   3
names(data)
[1] "aprvn1"    "prevany"   "alwaysPCV"

# Create a program to perform the simulation:
power = function(DD=data,reps=500,size=150,ss=743) {
        set.seed(ss)
        out=NULL
        for(i in 1:reps) {
                junk.data = DD[sample(x=seq(1:length(data[,1])),size=size,replace=T),]
                junk = glm(prevany ~ alwaysPCV + aprvn1,data=junk.data,family=binomial)
                out = c(out,summary(junk)$coeff[2,4])
        }
        count = ifelse(out<0.05,1,0)
        return(sum(count)/reps)
}

power(size=150)
power(size=500)
power(size=1000)
```

We don’t have much luck in this dataset (great variability and small effect size), but the estimated powers for the sample sizes are listed below:

Sample Size	Estimated Power (Percent)
100	5.8
150	7.0
500	9.0
1000	10.2

Although this example didn’t work out very well, you can use these ideas and apply them to your studies.