Biostatistics 140.623

Third Term, 2017-2018 Problem Set 1 (with R)

Martin Skarzynski

02/01/2018

Vitamin A Supplementation to Prevent Children's Mortality in Nepal

Study Design - Sample Size Estimation

Learning Objectives:

Students who successfully complete this project section will be able to:

- Calculate the sample size necessary for estimating mortality in children < 3 years of age with a desired level of precision.
- Estimate the sample size for a new study of vitamin A and mortality in children < 3 years of age.

Data Set :

The Nepal data set is located in the .csv data file named nepal621.csv.

Methods :

1) Suppose you are interested in choosing an appropriate sample size for estimating the 16- month mortality rate for children younger than 3 years of age in a developing country in which vitamin A supplementation is not currently available.

a. Use the available information from the Nepal data set to choose a sample size so that you estimate this rate to within +/- 0.5%.

```
# install.packages("readr")
# install.packages("dplyr")
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
nep  <- read_csv("nepal621.csv")
```

```
## Parsed with column specification:
## cols(
##   sex = col_character(),
##   age = col_character(),
##   trt = col_character(),
##   status = col_character()
## )
```

```
dim(nep)
```

```
## [1] 27121     4
```

```
head(nep)
```

```
## # A tibble: 6 x 4
##   sex   age   trt     status
##   <chr> <chr> <chr>   <chr>
## 1 Male  3-4   Placebo Alive
## 2 Male  1-2   Vit A   Alive
## 3 Male  3-4   Placebo Alive
## 4 Male  3-4   Vit A   Alive
## 5 Male  <1    Vit A   Alive
## 6 Male  3-4   Placebo Alive
```

```
nep %>%
 filter(age!="3-4") %>%
 filter(trt!="Vit A") %>%
 summarize(N_Alive = sum(status=="Alive"),
 Perc_Alive = round(N_Alive/n(),4)*100,
 N_Died = sum(status=="Died"),
 Perc_Died = round(N_Died/n(),4)*100,
 Total=n(),
 Perc_SD=sd(status=="Died")*100)
```

```
## # A tibble: 1 x 6
##   N_Alive Perc_Alive N_Died Perc_Died Total Perc_SD
##     <int>      <dbl>  <int>     <dbl> <int>   <dbl>
## 1    7880       97.1    239      2.94  8119    16.9
```

```
samp  <- function(s,d){
(1.96**2)*(s**2)/(d**2)
}
```

```
samp(s=16.9,d=0.5)
```

```
## [1] 4388.798
```

The sample size required estimate the 16-month mortality rate for children younger than 3 years of age in a developing country in which vitamin A supplementation is not currently available to within +/- 0.5% is 4389.

b. Now, suppose **no information** is available from this Nepal study. Determine what sample size would be required for each of a range of plausible values of the mortality rate. Summarize your sample size findings in a **table**.

True Difference	Standard Deviation	Sample Size per Group
0.25	12	8851
0.25	20	24586
0.5	12	2213
0.5	20	6147
0.75	12	983
0.75	20	2732

2) Now suppose you have a chance to investigate the effect of vitamin A supplementation on the mortality of children under 3 years of age. The power.prop.test() command in R can be used with the results of the Nepal trial to choose the size of the vitamin A and control groups (assuming equal sample sizes for both groups) for the new study. Confirm from the data set that the 16-month mortality in the placebo group is 0.0294 and the 16- month mortality in the Vitamin A group is 0.0245 for the Nepal

study. The estimated relative risk of death in the placebo group as compared to the Vitamin A group is $0.029/0.0245 = 1.2$. Assuming a significance level of 0.05 and power of 80%, the sample size needed in the new study to detect a relative risk of 1.2 is **17,144 children per group** according to the results on the next page. A total sample size of 34,288 children would be required.

```
> power.prop.test(n=NULL, p1=0.0294, p2=0.0245, sig.level=0.05, power=0.8,
alternative="two.sided")

    Two-sample comparison of proportions power calculation

        n = 17143.9
        p1 = 0.0294
        p2 = 0.0254
        sig.level = 0.05
        power = 0.8
        alternative = two.sided

NOTE: n is number in *each* group
```

```
power.prop.test(n=NULL, p1=0.0294, p2=0.0245, sig.level=0.05, power=0.8,
alternative="two.sided")
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 17143.9
##              p1 = 0.0294
##              p2 = 0.0245
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
nep %>%
 filter(age!="3-4") %>%
 group_by(trt) %>%
 summarize(N_Alive = sum(status=="Alive"),
 Perc_Alive = round(N_Alive/n(),4)*100,
 N_Died = sum(status=="Died"),
 Perc_Died = round(N_Died/n(),4)*100,
 Total=n(),
 Perc_SD=sd(status=="Died")*100)
```

```
## # A tibble: 2 x 7
##   trt     N_Alive Perc_Alive N_Died Perc_Died Total Perc_SD
##   <chr>     <int>      <dbl>  <int>     <dbl> <int>   <dbl>
## 1 Placebo    7880       97.1    239      2.94  8119    16.9
## 2 Vit A      8218       97.6    206      2.45  8424    15.4
```

3) Verify R's calculations for part 2) by hand using the method learned in class. Expect your answer to be close in value to, but not exactly the same as, that provided by R, due to rounding in hand calculations. (Stata uses a continuity correction that R doesn't, so the value calculated from Stata will also be different than the one you calculated by hand and in R.)

```
samp2  <- function(p1,p2){
    pm <- (p1+p2)/2
    qm <- abs(pm-1)
    q1 <- abs(p1-1)
    q2 <- abs(p2-1)
    d <- p1 - p2
    size <- (1.96*sqrt(2*pm*qm)+0.84*sqrt(p1*q1+p2*q2))**2/(d**2)
    size
}
```

```
samp2(p1=0.0294, p2=0.0245)
```

```
## [1] 17124.5
```

Using the method learned in class, I obtained a sample size of 17125. This answer is very close to the one returned by the `power.prop.test` function.

4) Construct a **table** that displays the total sample sizes required under various assumptions about the mortality rate in the control group and the relative risk of interest. Assume a significance level of 0.05 and 80% power. Comment on what you observe.

Vary the assumptions by: a. Assuming that the control group mortality rate (risk) is: 1. the same as that observed in Nepal placebo group of children < 3 years of age 2. or .5% lower 3. or .5% higher b. Assuming that the relative risk of death for children in the control group as compared to children receiving vitamin A is hypothesized to be: 1. 1.2 (the same as the relative risk that was estimated for Nepali children in this age group 2. or 1. 3. or 1.75.

Mortality Rate in the Control Group	Relative Risk of Interest	Sample Size per Group
0.0294	1.20	17125
0.0244	1.00	Inf
0.0344	1.75	1896
0.0294	1.20	17125
0.0244	1.00	Inf
0.0344	1.75	1896

As the relative risk approaches 1, the sample size gets larger. Using the method above, I obtained Inf (short for infinity) when the relative risk was 1. The greater the relative risk, the smaller the sample size required.

5) Construct another **table** that displays the total sample sizes required under the same varying assumptions of the mortality rate in the control group and the relative risk of interest. This time, assume a significance level of 0.05 and 90% power. Comment on what you observe.

```
samp3  <- function(p1,p2){
    pm <- (p1+p2)/2
    qm <- abs(pm-1)
    q1 <- abs(p1-1)
    q2 <- abs(p2-1)
    d <- p1 - p2
    size <- (1.96*sqrt(2*pm*qm)+1.28*sqrt(p1*q1+p2*q2))**2/(d**2)
    size

}
```

Mortality Rate in the Control Group	Relative Risk of Interest	Sample Size per Group
0.0294	1.20	22929
0.0244	1.00	Inf
0.0344	1.75	2538
0.0294	1.20	22929
0.0244	1.00	Inf
0.0344	1.75	2538

To have a higher power (90% instead of 80%), the sample size required is higher. Just as before, the greater the relative risk, the smaller the sample size required.

6) Select a design based upon your findings from parts 4 and 5 above. Write a brief paragraph that presents and justifies your choice. Be numerate.

Based on my findings from parts 4 and 5 above, I would aim to recruit between 18000 and 23000 children to investigate the effect of vitamin A supplementation on the mortality of children under 3 years of age. This sample size will allow me to detect a relative risk similar to the one that was estimated for Nepali children in the less-than-3-year-old age group with a power between ~80-90% and a significance level of 0.05. These values are reasonable because they are based on the Nepal Vitamin A study, which is likely to be relevant for future studies regarding the effect of Vitamin supplementation on child mortality. It may be a good idea to increase the sample size to account for potential losses-to-follow-up in each treatment group. Losses-to-follow-up may occur, for example, if the parents of some of the children decide not to participate anymore.