# Biostatistics 140.623
# Second Term, 2017-2018
## Problem Set 1
## Answer Key

**Vitamin A Supplementation to Prevent Children's Mortality in Nepal**

**Study Design - Sample Size Estimation**

1) Suppose you are interested in choosing an appropriate sample size for estimating the mortality rate within +/- 0.5% for children <3 years of age in a developing country in which vitamin A supplementation is not available.

a) Assuming a significance level of 0.05 and using available information from the Nepal data set, what sample size is required?

The sample size for estimating a proportion is $n = \frac{z^2 pq}{d^2}$. We have been told that d=0.005 and the z corresponding to a two-sided significance level of 0.05 is 1.96. In the Nepal study, 239 of the 8119 children <3 years of age who did not receive vitamin A supplementation died. We can use this proportion (239/8119 = 0.029) to estimate the mortality rate in a developing country in Africa which does not receive vitamin A supplementation.

Thus, $n = \frac{(1.96)^2(0.029)(1-0.029)}{(0.005)^2} = \frac{0.108}{0.000025} = 4327.02 \rightarrow 4328$

So, at least 4328 children are needed to estimate to within ±0.5 percentage points with 95% confidence the true mortality rate of children not receiving vitamin A.

b) Now, suppose no information is available from the 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.

If we didn't have the estimate of mortality rate in a population without vitamin A supplementation, we could use p=0.5 as a conservative estimate for the purpose of calculating sample size. This is conservative because p*q is never bigger than when p=0.5.

Then, $n = \frac{(1.96)^2(0.5)(1-0.5)}{(0.005)^2} = \frac{0.9604}{0.000025} = 38416$

Notice how much larger the necessary sample size is when we do not have a good estimate of the mortality rate in our intended study population.

**HOWEVER, p=0.5 is not a likely estimate of mortality in this population**, we can calculate sample size based on more plausible estimates of mortality rates. For example, we can use p ranging from 0.10 to 0.01, based on an investigation of reported mortality rates in this age range in different countries (obtained from DHS or WHO statistics).

When we estimate p=0.1, $n = \frac{(1.96)^2(0.1)(1-0.1)}{(0.005)^2} = \frac{0.3457}{0.000025} = 13829.76 \rightarrow 13830$

When we estimate p=0.05, $n = \frac{(1.96)^2(0.05)(1-0.05)}{(0.005)^2} = \frac{0.1825}{0.000025} = 7299.04 \rightarrow 7300$

When we estimate p=0.02, $n = \frac{(1.96)^2(0.02)(1-0.02)}{(0.005)^2} = \frac{0.0753}{0.000025} = 3011.8 \rightarrow 3012$

When we estimate p=0.01, $n = \frac{(1.96)^2(0.01)(1-0.01)}{(0.005)^2} = \frac{0.0380}{0.000025} = 1521.3 \rightarrow 1522$

**Table 1. Sample Size Calculated Under Different Assumptions of Mortality Rate**

| Estimated Mortality Rate | Sample Size Required to Estimate Mortality Rate to Within +/- 0.5% |
|---|---|
| 1% | 1522 |
| 2% | 3012 |
| 5% | 7300 |
| 10% | 13,830 |

Extra Supplemental Note: If we had been asked to determine the precision of the mortality rate estimates, we would need to begin with a fixed sample size. For example, we could use the sample size calculated in the first part of this problem, n = 4328. Rewriting the sample size formula mentioned in the first sentence of this solution gives the following formula for *d*.

$$d = z * \text{sqrt}( [p*q] / n)$$

2) Now suppose you have a chance to investigate vitamin A supplementation and children's mortality for children under 3 years of age. The `sampsi` command in `Stata` can be implemented to use 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 0.0294/0.0245 = 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 17,550 children per group according to the results below. A total sample size of 35,100 children would be required.

 

```
. sampsi 0.0294 0.0245, p(0.8)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            p1 =   0.0294
            p2 =   0.0245
         n2/n1 =   1.00

Estimated required sample sizes:

            n1 =    17550
            n2 =    17550
```

3) Verify Stata's calculations for part 2) using the formula provided in class. Expect your answer to be close in value to, but not exactly the same as, that provided by `Stata`. (`Stata` uses a continuity correction.)

The sample size in each group necessary to detect a difference in proportion of $\Delta$ is

$$n = \frac{\left[ z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} + z_{\beta}\sqrt{p_1q_1 + p_2q_2} \right]^2}{\Delta^2} \text{ where } \bar{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}.$$

(Notice, the $n_1$ and $n_2$ in the $\bar{p}$ calculation are the $n_1$ and $n_2$ from the population from which you obtained your sample estimates—not the $n_1$ and $n_2$ you are trying to find.)

From the Nepal study, we have estimates of $p_1$ and $p_2$:

$p_1$=proportion of vitamin A children dying = 206/8424 = 0.0245
$p_2$=proportion of placebo children dying = 239/8119 = 0.0294

Thus, assuming equal sample sizes, $\bar{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{n\ p_1 + n\ p_2}{n\ + n} = \frac{p_1 + p_2}{2} = 0.027$.

To achieve 80% power, $z_{\beta} = 0.84$. As usual, $z_{\alpha/2} = 1.96$.

So,

$$n = \frac{\left[1.96\sqrt{2(0.027)(0.973)} + 0.84\sqrt{(0.0245)(0.9755) + (0.0294)(0.9706)}\right]^2}{(0.0245 - 0.0294)^2} =$$

$$= 17{,}146$$

Notice, the sample size we calculated by hand does not agree exactly with `Stata`'s calculation. This is because `Stata` uses a continuity correction (see `Stata` Reference Manual for `sampsi` command).

 

4) Construct a table that displays the total sample sizes required under various assumptions of 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 1.5
    3. or 1.75.

Assuming that the **control mortality rate is the same**:

Since $RR = \frac{p_{placebo}}{p_{VitA}} = 1.2$. So,

$p_{VitA} = p_{placebo} / RR = 0.0294/1.2 = 0.0245$

Similarly, for RR = 1.5, then

$p_{VitA} = p_{placebo} / RR = 0.0294/1.5 = 0.0196$

Also, for RR = 1.75, then

$p_{VitA} = p_{placebo} / RR = 0.0294/1.75 = 0.0168$

Assuming that the **control mortality rate is 0.5% lower**, then

$p_{placebo} = 0.0294 - 0.005 = 0.0244.$

For RR = 1.2, then

$p_{VitA} = p_{placebo} / RR = 0.0244/1.2 = 0.0203$

Similarly, for RR = 1.5, then

$p_{VitA} = p_{placebo} / RR = 0.0244/1.5 = 0.0163$

Also, for RR = 1.75, then

$p_{VitA} = p_{placebo} / RR = 0.0244/1.75 = 0.0139$

Assuming that the **control mortality rate is 0.5% higher**, then

$p_{placebo} = 0.0294 + 0.005 = 0.0344.$

 

For RR = 1.2, then

$$p_{VitA} = p_{placebo} / RR = 0.0344/1.2 = 0.0287$$

Similarly, for RR = 1.5, then

$$p_{VitA} = p_{placebo} / RR = 0.0344/1.5 = 0.0229$$

Also, for RR = 1.75, then

$$p_{VitA} = p_{placebo} / RR = 0.0344/1.75 = 0.0197$$

**Please note: These calculations also could have been performed by fixing the Vitamin A mortality rate and varying the placebo (control) mortality rate. Either way is acceptable as long as it is explained and clarified.**

Now, we can calculate the sample size using `Stata`'s **`sampsi`** command. In order to calculate the sample size needed to detect a difference in proportions, we must supply the two hypothesized proportions. The default power is 90%, so if we want to use a different power level, we must specify it.

For RR =1.2:

```
. sampsi .0244 .0203, p(.8)
. sampsi .0294 .0245, p(.8)
. sampsi .0344 .0287, p(.8)
```

For RR =1.5:

```
. sampsi .0244 .0163, p(.8)
. sampsi .0294 .0196, p(.8)
. sampsi .0344 .0229, p(.8)
```

For RR =1.75:

```
. sampsi .0244 .0139, p(.8)
. sampsi .0294 .0168, p(.8)
. sampsi .0344 .0197, p(.8)
```

**Table 2: Sample Sizes Required Per Group by varying the Assumed Relative Risk (RR) and Assumed Mortality Rates in the Placebo Group, 2-sided $\alpha = 0.05$, Power = 80%**

| | $p_{placebo}$ | | |
|---|---|---|---|
| RR= $p_{placebo} / p_{Vit\ A}$ | 0.0244 | 0.0294 | 0.0344 |
| 1.2 | 20,889 | 17,550 | 15,111 |
| 1.5 | 5,013 | 4,017 | 3,474 |
| 1.75 | 2,861 | 2,387 | 2,045 |

From **Table 2**, we observe:

- For a fixed mortality rate in the placebo group, if we wish to detect a smaller relative difference in mortality rate between placebo and Vitamin A groups, the necessary sample size increases.
- For a fixed relative difference of interest (i.e., fixed relative risk), the necessary sample size decreases as the mortality rate in the placebo group increases.

 

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.

For RR =1.2:

```
. sampsi .0244 .0203, p(.9)
. sampsi .0294 .0245, p(.9)
. sampsi .0344 .0287, p(.9)
```

For RR =1.5:

```
. sampsi .0244 .0163, p(.9)
. sampsi .0294 .0196, p(.9)
. sampsi .0344 .0229, p(.9)
```

RR =1.75:

```
. sampsi .0244 .0139, p(.9)
. sampsi .0294 .0168, p(.9)
. sampsi .0344 .0197, p(.9)
```

**Table 3: Sample Sizes Required Per Group by varying the Assumed Relative Risk (RR) and Assumed Mortality Rates in the Placebo Group, 2-sided $\alpha$ = 0.05, Power = 90%**

| | $p_{placebo}$ | | |
|---|---|---|---|
| RR= $p_{placebo} / p_{Vit\ A}$ | 0.0244 | 0.0294 | 0.0344 |
| 1.2 | 27,800 | 23,357 | 20,111 |
| 1.5 | 6,628 | 5,430 | 4,593 |
| 1.75 | 3,767 | 3,142 | 2,692 |

Comparing **Tables 2 and 3**, we can see that:

- For a fixed mortality rate in the placebo group and fixed relative difference of interest, the necessary sample size increases if we desire larger power of the statistical test to detect the difference.

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.

For a new investigation of Vitamin A supplementation, we assumed that the placebo mortality rate was 2.94% ± .5% based on prior findings in the study conducted in Nepal. Our new prospective randomized community trial will require a sample size of 5,013 children per group (5,013 in the placebo group and 5,013 in the Vitamin A group). This total sample size of 13,300 10,026 children will provide 80% or more power to detect an increase in the risk of mortality of 50% or greater when comparing mortality in the placebo versus Vitamin A groups.

 