Biost 518: Applied Biostatistics II
Emerson, Winter 2006

Homework #3 Key
January 28, 2003

Written problems due at the beginning of class, Friday, January 27, 2003.

All questions relate to the planning of a phase II cancer prevention study of DFMO and its ability to suppress the level polyamines in the colonic mucosa. We will focus in particular on the spermidine levels, and will summarize the distribution of spermidine in a treatment group using the mean μ. We consider below several different approaches which differ in the definition of the "treatment effect" θ. I note here (and again below), that several of the options we consider would be considered highly inappropriate for a real study.

We desire to calculate the sample size required to detect a hypothesized effect of DFMO on the mean spermidine level. We intend to use a one-sided level α hypothesis test, and we want to have power β to reject the null hypothesis H_0: θ.= θ_0 when the "design" alternative H_1: θ.= θ_1 is true.

Recall from lecture that the most common formula used in sample size calculations is

$$N = \frac{\delta_{\alpha\beta}^2 V}{\Delta^2}$$

where

- N is the total sample size to be accrued to the study,
- V is the average variability contributed by each subject to the estimate of the treatment effect θ (for each problem below, I provide the formula for V),
- $\delta_{\alpha\beta}$ is a "standardized alternative" which would allow a standardized one-sided level α hypothesis test to reject the null hypothesis with probability (power) β (note that many textbooks use notation in which the power is denoted 1-β), and
- Δ is some measure of the distance between the null and alternative hypotheses.

Often clinical trials are conducted with a stopping rule which allows early termination of the study on the basis of one or more interim analyses of the data. When such a "group sequential test" is to be used, the value of the standardized alternative $\delta_{\alpha\beta}$ must be found using special computer software. On the other hand, when a "fixed sample study" (i.e., one in which the data are analyzed only once) is to be conducted, the standardized alternative for a one-sided test is given by

$$\delta_{\alpha\beta} = z_{1-\alpha} + z_{\beta}$$

where z_p is the pth quantile of the standard normal distribution. In Stata, the value of z_p can be found by using the function `invnorm( )`. For instance, if α = 0.025, the value of $z_{0.975}$ can be found from the Stata command

```
disp invnorm(0.975).
```

(Stata would then display `1.959964.`)

The formula for Δ depends on the statistical model used, but is usually either

- $\Delta = \theta_1 - \theta_0$ (used for inference in "additive models" for means and proportions, and sometimes medians), or
- $\Delta = log(\theta_1 / \theta_0)$ (used for inference in "multiplicative models" for geometric means, odds, and hazards, and sometimes means and medians),

1. **(Obtaining estimates for use in sample size calculations)** When making inference about spermidine levels using means (and differences of means), the formula for V will typically involve the standard deviation σ of measurements made within a treatment group. When using paired observations, the formula for V may also involve the correlation ρ between two measurements made on the same individual some time apart. We will derive estimates of σ and ρ from a pilot study of DFMO. The following estimates should be used as needed to answer all other questions. Using the DFMO data set available on the class web pages:

    a. What is the standard deviation of spermidine measurements made at baseline (time of randomization) across all subjects? (We can ignore dose, because these measurements were made prior to receiving drug.)

Ans: s = 1.553 micromoles/mg protein. *(I keep 4 significant digits for use in intermediate calculations).*

    b. What is the correlation of spermidine measurements made at baseline (time 0) and after 12 months of study on subjects in the placebo group? (We use only the placebo group to avoid having to adjust for a treatment effect.)

Ans: r = 0.3935. *(I keep 4 significant digits for use in intermediate calculations).*

2. **(A single arm study of spermidine after 12 months of treatment and effect of different levels of power)** Suppose we choose to provide DFMO at a single dose to N subjects. We use as our measure of treatment effect the mean spermidine level at the end of treatment. Suppose from previous study we know that in the untreated state the mean spermidine level is 3.25 micromoles/mg protein, and we want to detect whether treatment with DFMO will result instead in an average spermidine level of 2.50 micromoles/mg protein. We intend to perform a hypothesis test in which

    - the one-sided level of significance is α = 0.025,
    - the measure of treatment effect is $\theta = \mu_{D,12}$ (the mean spermidine level in the patients treated with DFMO after 12 months of treatment),
    - the average variability contributed by each subject to the estimated treatment effect (the sample mean) is $V = \sigma^2$, and
    - the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

    a. What sample size will provide 80% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.80} = 0.842$; $\delta_{\alpha\beta} = 1.960 + 0.8416 = 2.802$.

To find Δ: $\theta_0 = 3.25$; $\theta_1 = 2.50$; $\Delta = 2.50 - 3.25 = -0.7500$.

To find V: $V = \sigma^2 = 1.553^2 = 2.412$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 2.802^2 \times 2.412 / 0.7500^2 = 33.7$, so round up to 34.

b. What sample size will provide 90% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.90} = 1.282$; $\delta_{\alpha\beta} = 1.960 + 1.282 = 3.242$.

To find Δ: $\theta_0 = 3.25$; $\theta_1 = 2.50$; $\Delta = 2.50 - 3.25 = -0.7500$.

To find V: $V = \sigma^2 = 1.553^2 = 2.412$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.242^2 \times 2.412 / 0.7500^2 = 45.1$, so round up to 46.

c. What sample size will provide 95% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.95} = 1.645$; $\delta_{\alpha\beta} = 1.960 + 1.645 = 3.605$.

To find Δ: $\theta_0 = 3.25$; $\theta_1 = 2.50$; $\Delta = 2.50 - 3.25 = -0.7500$.

To find V: $V = \sigma^2 = 1.553^2 = 2.412$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.605^2 \times 2.412 / 0.7500^2 = 55.7$, so round up to 56.

d. What sample size will provide 97.5% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 3.25$; $\theta_1 = 2.50$; $\Delta = 2.50 - 3.25 = -0.7500$.

To find V: $V = \sigma^2 = 1.553^2 = 2.412$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 2.412 / 0.7500^2 = 65.9$, so round up to 66.

e. What sample size will guarantee that a 95% confidence interval for θ would not include both the null and alternative hypotheses?

Ans: The same sample size that provides 97.5% power guarantees that the width of the 95% confidence interval will be the distance between the null hypothesis and the design alternative hypothesis. So a sample size of 66 will suffice.

f. Why is this a very bad study design scientifically?

Ans: This study presumes that we can absolutely trust that the mean under the null hypothesis was determined with infinite precision (i.e., based on an infinite sample size). Furthermore, it presumes that the patient population used in this next study will be exactly the same sort of patients that were used to determine the null hypothesis. To the extent that the patient population is different due to underlying clinical or subclinical disease, due to changes in diet, due to other environmental influences such as ancillary treatments, or due to genetic profile, this would make the previously observed mean irrelevant. This would also be the case if there have been changes in the way polyamines are measured in the laboratory. All of this is a pretty tall order, so such single arm trials are extremely problematic.

3. **(A single arm study of spermidine after 12 months of treatment and the effect of dichotomizing the data)** Suppose we choose to provide DFMO at a single dose to *N* subjects. We use as our measure of treatment effect the proportion of subjects having

spermidine level below 2.50 micromoles/mg protein at the end of treatment. Suppose from previous study we know that in the untreated state the mean spermidine level is 3.25 micromoles/mg protein and that the data is approximately normally distributed. We are guessing that the treatment treatment with DFMO will result instead in an average spermidine level of 2.50 micromoles/mg protein. We intend to perform a hypothesis test in which

- the one-sided level of significance is $\alpha = 0.025$,
- the desired statistical power is $\beta = 0.975$,
- the measure of treatment effect is $\theta = p_{D,12}$ (the proportion of subjects treated with DFMO who have spermidine levels lower than 2.50 micromoles/mg protein after 12 months of treatment),
- the average variability contributed by each subject to the estimated treatment effect (the sample proportion) is $V = \theta(1-\theta)$ (most often, we would compute this under the alternative hypothesis in this setting),
- the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

a. Using the estimated standard deviation obtained in problem 1 and assuming normally distributed spermidine levels, what proportion of subjects would you expect to have measurements lower than 2.50 micromoles/mg protein if the true mean were 3.25 micromoles/mg protein? (This can serve as your null hypothesis for the test of proportions.)

Ans: We want to determine the probability that a $N(\mu = 3.25, \sigma^2 = 1.553^2)$ random variable would be less than 2.50. Using the properties of the normal distribution, this is the same as the probability that a N(0,1) random variable would be less than (2.50 – 3.25)/1.553 = -0.4829. Using the Stata command "`display norm(-0.4829)`" we find that under the null hypothesis (and the assumption of normality) we would expect $p_{D,12} = \theta_0 = 0.3146$.

b. Using the estimated standard deviation obtained in problem 1 and assuming normally distributed spermidine levels, what proportion of subjects would you expect to have measurements lower than 2.50 micromoles/mg protein if the true mean were 2.50 micromoles/mg protein? (This can serve as your alternative hypothesis for the test of proportions.)

Ans: We want to determine the probability that a $N(\mu = 2.50, \sigma^2 = 1.553^2)$ random variable would be less than 2.50. Using the properties of the normal distribution, this is the same as the probability that a N(0,1) random variable would be less than (2.50 – 2.50)/1.553 =0.0000. Using the Stata command "`display norm(0)`" we find that under the alternative hypothesis (and the assumption of normality) we would expect $p_{D,12} = \theta_1 = 0.5000$.

c. What sample size will provide 97.5% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0.3146$; $\theta_1 = 0.5000$; $\Delta = 0.3146 - 0.5000 = -0.1854$.

To find V: $V = \theta_1(1-\theta_1) = 0.2500$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 0.2500 / 0.1854^2 = 111.8$, so round up to 112.

d. What advantages or disadvantages does this study design have over the study design used in problem 2?

Ans: To the extent that it is clinically most important to lower spermidine levels below 2.50 micromoles/mg protein, this study design answers the most relevant scientific question. However, if such a scientific threshold did not exist, then we have clearly lost information about how DFMO might tend to lower spermidine levels across the population. This loss of information is reflected in the higher sample size requirements when dichotomizing the data: 112 versus 66.

e. Why is this a very bad study design scientifically?

Ans: For the exact same reasons as given in problem 2, we should not use a single arm study.

4. **(A single arm study of change in spermidine over 12 months of treatment)** Suppose we choose to provide DFMO at a single dose to *N* subjects. We use as our measure of treatment effect the difference between mean spermidine level at the end of treatment and at the beginning of treatment (because we are using means, we know that the difference in means is the same as the mean change). The null hypothesis is that the mean change is 0 micromoles/mg protein, and we want to detect whether treatment with DFMO will result in an average decrease of 0.75 micromoles/mg protein (this hypothesis corresponds to the same difference hypothesized in problem 2). We intend to perform a hypothesis test in which

   - the one-sided level of significance is $\alpha = 0.025$,
   - the desired statistical power is $\beta = 0.975$,
   - the measure of treatment effect is $\theta = \mu_{D,12} - \mu_{D,0}$ (the mean spermidine level in the patients treated with DFMO after 12 months of treatment minus the mean spermidine level in those same patients prior to treatment), and
   - the average variability contributed by each subject to the estimated treatment effect (the sample mean change) is $V = 2\sigma^2(1-\rho)$.
   - the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

   a. What sample size will provide 97.5% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 2\sigma^2(1 - \rho) = 2 \times 1.553^2 \times (1 - 0.3935) = 2.926$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 2.926 / 0.7500^2 = 79.9$, so round up to 80.

b. What advantages or disadvantages does this study design have over the study design used in problem 2?

Ans: It is common for naïve researchers to believe that they always gain precision by subtracting off baseline values. In this case, however, such a method of adjusting for baseline measurements decreased precision as evidenced by the increased sample size needed for this design: 80 when using the difference versus 66 when using only the follow up values. The

approach using the change does have the advantage of not relying on the need to know that the mean spermidine value was 3.25. It replaces that requirement with the need to know that spermidine values would not change over time due to aging, due to secular trends in the diet or other environmental variables, or due to drift in laboratory measurement procedures.

c. What would the correlation between measurements made on the same subject have to be in order to have this "pre/post" comparison more efficient than the study design used in problem 2?

Ans: Only V is different between problems 2 and 4. By examining the formulas for V, we see that V will be equal in the two approaches when $\rho = 0.5$. When $\rho < 0.5$, the approach based on using only the follow-up will be more precise; when $\rho > 0.5$, the approach based on the change in measurements will be more precise.

d. Why is this a very bad study design scientifically?

Ans: This study presumes that we can absolutely trust that in the absence of DFMO treatment, there will be no systematic change in spermidine levels. As noted in the answer to part b, factors such as aging, secular trends in diet, or drift in laboratory procedures can make such an assumption inappropriate. (Recall that we have seen such factors cause a statistically significant change in the placebo group in the beta carotene dataset.) All of this is a pretty tall order, so such single arm trials are extremely problematic.

5. **(A two arm study of mean spermidine after 12 months of treatment)** Suppose we randomly assign *N* subjects in a double blind fashion to receive either DFMO at a single dose or placebo. We use a randomization ratio of *r* subjects on DFMO to 1 subject on placebo. We use as our measure of treatment effect the difference between mean spermidine level at the end of treatment for patients on DFMO and mean spermidine level at the end of treatment for patients on placebo. The null hypothesis is that the difference in means is 0 micromoles/mg protein, and we want to detect whether treatment with DFMO will result in an average spermidine level that is 0.75 micromoles/mg protein lower than might be expected on placebo (this hypothesis corresponds to the same difference hypothesized in problem 2). We intend to perform a hypothesis test in which

   - the one-sided level of significance is $\alpha = 0.025$,
   - the desired statistical power is $\beta = 0.975$,
   - the measure of treatment effect is $\theta = \mu_{D,12} - \mu_{P,12}$ (the mean spermidine level in the patients treated with DFMO after 12 months of treatment minus the mean spermidine level in in the patients treated with placebo after 12 months of treatment),
   - the average variability contributed by each subject to the estimated treatment effect (the difference in sample means) is $V = \sigma^2(1/r+2+r)$, and
   - the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

   a. What sample size will provide 97.5% power to detect the design alternative when *r=1*?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_{\beta} = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = \sigma^2(1/r+2+r) = 4 \times 1.553^2 = 9.647$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 9.647 / 0.7500^2 = 263.5$, so round up to 264.

b. What sample size will provide 97.5% power to detect the design alternative when *r=2*?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = \sigma^2(1/r+2+r) = 4.5 \times 1.553^2 = 10.85$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 10.85 / 0.7500^2 = 296.5$, so round up to 297 (or 298 so there can be equal numbers in each treatment group).

c. What sample size will provide 97.5% power to detect the design alternative when *r=5*?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = \sigma^2(1/r+2+r) = 7.2 \times 1.553^2 = 17.37$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 17.37 / 0.7500^2 = 474.5$, so round up to 475 (or 476 so there can be equal numbers in each treatment group).

d. What advantages or disadvantages does this study design have over the study design used in problem 2?

Ans: This design uses a concurrent control group. This is the standard for credible scientific research.

6. **(A two arm study of change in mean spermidine after 12 months of treatment)** Suppose we randomly assign *N* subjects in a double blind fashion to receive either DFMO at a single dose or placebo. We use a randomization ratio of 1 subject on DFMO to 1 subject on placebo. We use as our measure of treatment effect the mean change in spermidine level at the end of treatment for patients on DFMO and mean change in spermidine level at the end of treatment for patients on placebo. The null hypothesis is that the difference in means is 0 micromoles/mg protein, and we want to detect whether treatment with DFMO will result in an average change in spermidine level that is 0.75 micromoles/mg protein lower than might be expected on placebo (this hypothesis corresponds to the same difference hypothesized in problem 2). We intend to perform a hypothesis test in which

   - the one-sided level of significance is $\alpha = 0.025$,
   - the desired statistical power is $\beta = 0.975$,
   - the measure of treatment effect is $\theta = (\mu_{D,12} - \mu_{D,0}) - (\mu_{P,12} - \mu_{P,0})$ (the mean change in spermidine level in the patients treated with DFMO after 12

months of treatment minus the mean change in spermidine level in the patients treated with placebo after 12 months of treatment), and

- the average variability contributed by each subject to the estimated treatment effect (the difference in sample means) is $V = 8\sigma^2(1-\rho)$.
- the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

a. What sample size will provide 97.5% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 8\sigma^2(1-\rho) = 8 \times 1.553^2 \times (1 - 0.3935) = 11.70$

To find N: $N = \delta_{\alpha\beta}^2 \, V / \Delta^2 = 3.920^2 \times 11.70 / 0.7500^2 = 319.6$, so round up to 320.

b. What advantages or disadvantages does this study design have over the study design used in problem 5?

Ans: By randomization, both treatment arms would tend to have the same mean spermidine at baseline. Hence, the difference in final spermidines should estimate the exact same quantity as the difference in change of spermidine levels. Because the scientific question is the exact same, we are free to consider only the statistical precision in choosing the better of the two analysis approaches. In this case, the spermidine levels in a person are not all that highly correlated over the time period of the trial. Thus it turns out that the more statistically precise approach would be to ignore the baseline measurements altogether. This is a quite counter-intuitive result, but an important one to remember: Taking differences between measurements does not always gain you precision (or any accuracy) when the subjects are comparable at baseline, and in a randomized trial we are safe in assuming the subjects were comparable at baseline.

7. **(A two arm study of mean spermidine after 12 months of treatment using Analysis of Covariance)** Suppose we randomly assign N subjects in a double blind fashion to receive either DFMO at a single dose or placebo. We use a randomization ratio of 1 subject on DFMO to 1 subject on placebo. We use as our measure of treatment effect the mean change in spermidine level at the end of treatment for patients on DFMO and mean change in spermidine level at the end of treatment for patients on placebo. We decide to analyze our data using linear regression in which we model the mean spermidine level after 12 months of treatment (SPD12) including as predictors a binary variable measuring treatment assignment (TX) and a continuous variable measuring the baseline spermidine level for each individual (SPD0):

$$E\left(SPD12_i \mid TX_i, SPD0_i\right) = \beta_0 + \beta_1 \times TX_i + \beta_2 \times SPD0_i$$

The null hypothesis is that treatment with DFMO is not associated with any difference in the mean spermidine level, and we want to detect whether treatment with DFMO will result in an average spermidine level that is 0.75 micromoles/mg protein lower than might be expected on placebo (this hypothesis corresponds to the same difference hypothesized in problem 2). We intend to perform a hypothesis test in which

- the one-sided level of significance is $\alpha = 0.025$,

- the desired statistical power is $\beta = 0.975$,
- the measure of treatment effect is $\theta = \beta_1$ (see part a),
- the average variability contributed by each subject to the estimated treatment effect (the difference in sample means) is $V = 4\sigma^2(1-\rho^2)$,
- the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

a. What is the scientific interpretation of the slope parameter β_1?

Ans: β_1 is the difference in mean spermidine level after treatment between an individual treated with DFMO and an individual who had the same baseline spermidine value but was treated with placebo.

b. What sample size will provide 97.5% power to detect the design alternative?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 4\sigma^2(1-\rho^2) = 4 \times 1.553^2 \times (1 - 0.3935^2) = 8.153$

To find N: $N = \delta_{\alpha\beta}^2 \, V / \Delta^2 = 3.920^2 \times 8.153 / 0.7500^2 = 222.7$, so round up to 224.

c. For what values of the within subject correlation will this analysis be more efficient than the analysis in problem 5?

Ans: Only V is different between problems 5a and 7. By examining the formulas for V, we see that V will be equal in the two approaches only when $\rho = 0$. Otherwise, the approach based on the analysis of covariance will be more precise.

d. For what values of the within subject correlation will this analysis be more efficient than the analysis in problem 6?

Ans: Only V is different between problems 6 and 7. By examining the formulas for V, we see that V will be equal in the two approaches only when $\rho = 1$. Otherwise, the approach based on the analysis of covariance will be more precise.

Some additional general comments about the analysis of covariance approach:

- **These uniformly better efficiency of the Analysis of Covariance (ANCOVA) estimator relies on the fact that the mean baseline measurements are the same for the two treatment groups. In a randomized clinical trial, we know that to be the case. In an observational study, that might not be the case, and we would then have to worry about whether we were answering the same scientific question in problems 5, 6, and 7. Generally, in an observational study, the analyses in these three problems would be answering different scientific questions.**
- **Had we used the change in spermidine as our response variable in the above regression model adjusting for treatment assignment and baseline spermidine, we would have obtained the exact same slope estimate for β_1. Thus the key thing is to adjust for the baseline in a regression model.**

- **For those of you who do (or have to in Stat 513) care about such things, the slope can be shown to be basically the same as the Uniform Minimum Variance Unbiased Estimator (UMVUE) for this problem.**
- **Given that the Analysis of Covariance (ANCOVA) estimator is the most efficient estimator in the clinical trial setting, is there ever a time we would want to use anything else? Yes. When the measurement is very expensive (in money, as with MRI, or in patient safety, as with some biopsies), we could consider whether it is better to have 1 measurement on 2N people and use only the final measurement as in problem 5, or to have 2 measurements on N people and use the ANCOVA analysis as in problem 7. The deciding point should be when V from problem 5 is double V from problem 7. This occurs when $(1-\rho^2)=0.5$ or $\rho = 0.707$. So, if you are constrained by the number of measurements you can afford, you will have more precision if you only obtain the follow-up measurement when $\rho < 0.707$, and if you obtain both baseline and follow-up measurements when $\rho > 0.707$.**

I didn't ask this question, but: Given that the analysis of covariance approach is the most efficient, is there ever a time we should use something else in a randomized clinical trial?

e. Suppose we choose instead to use a sample size of 30. What power do we have to detect the design alternative of a 2.50 micromole/mg protein difference in mean spermidine levels?

Ans: To solve this problem, we need to rearrange the sample size formula to solve for the standardized alternative, and then find z_β.

$$N = \frac{\delta_{\alpha\beta}^2 V}{\Delta^2} \quad \Rightarrow \quad \delta_{\alpha\beta} = \Delta\sqrt{\frac{N}{V}}$$

$$\Rightarrow \quad z_\beta = \Delta\sqrt{\frac{N}{V}} - z_{1-\alpha} = 2.50\sqrt{\frac{30}{8.153}} - 1.96 = 2.836$$

And using the Stata command "`display norm(2.836)`", we find that the power is 99.77%.

f. Suppose we choose instead to use a sample size of 30. For what alternative do we have 97.5% power?

Ans: To solve this problem, we need to rearrange the sample size formula to solve for the difference in the hypotheses, and then find $\Delta = \theta_1 - \theta_0 = \theta_1$.

$$N = \frac{\delta_{\alpha\beta}^2 V}{\Delta^2} \quad \Rightarrow \quad \Delta = \delta_{\alpha\beta}\sqrt{\frac{V}{N}} = 3.92\sqrt{\frac{8.153}{30}} = 2.04$$

8. **(A subgroup analysis and test for interaction in a two arm study of mean spermidine after 12 months of treatment)** Suppose we randomly assign *N* subjects in a double blind fashion to receive either DFMO at a single dose or placebo. We use a randomization ratio of 1 subject on DFMO to 1 subject on placebo. We use as our measure of treatment effect the difference between mean spermidine level at the end of treatment for patients on DFMO and

mean spermidine level at the end of treatment for patients on placebo. The null hypothesis is that the difference in means is 0 micromoles/mg protein, and we want to detect whether treatment with DFMO will result in an average spermidine level that is 0.75 micromoles/mg protein lower than might be expected on placebo (this hypothesis corresponds to the same difference hypothesized in problem 2). We want to perform tests separately for each of two equal size subgroups (say, males and females) in the population. We intend to perform a hypothesis test in which

- the one-sided level of significance is $\alpha = 0.025$,
- the desired statistical power is $\beta = 0.975$,
- the measure of treatment effect is $\theta = \mu_{D,12} - \mu_{P,12}$ (the mean spermidine level in the patients treated with DFMO after 12 months of treatment minus the mean spermidine level in in the patients treated with placebo after 12 months of treatment),
- the average variability contributed by each subject to the estimated treatment effect (the difference in sample means) is $V = 4\sigma^2$, and
- the comparison between alternative and null hypotheses is $\Delta = \theta_1 - \theta_0$.

a. What sample size is needed in each subgroup to provide 97.5% power to detect the design alternative, if each hypothesis test can be performed using the 0.025 level of significance? So what is the total sample size required in this setting? (Note that this last quantity could have been obtained from the general formula by using $V = 8\sigma^2$, where we multiplied the subgroup average variability by 2 to account for needing the sample size in each subgroup.)

<u>Ans</u>: From the answer to problem 5a, we find that we need 264 subjects in each subgroup, and therefore 528 subjects total. This can also be derived as follows:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 8\sigma^2 = 8 \times 1.553^2 = 19.29$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 19.29 / 0.7500^2 = 526.9$, so round up to 528.

b. If two subgroups are tested, we are in fact giving ourselves two opportunities to declare DFMO beneficial. If our type I error rate is 0.025 on each test, then our experimentwise error might be nearly double that (0.049375, which is derived by considering the error rate of 0.025 for making a mistake in males plus 0.025 for making a mistake in females minus 0.025^2 for making a mistake in both males and females at the same time). Because of this, usual statistical practice in general (and for regulatory agencies in particular) might demand that you provide an adjustment for the "multiple comparisons" by using one-sided level $\alpha = 0.0125$ tests in each subgroup (in these two independent subgroups, we could actually use $\alpha = 0.01257912$). What sample size is needed in each subgroup to provide 97.5% power to detect the design alternative in each subgroup if we make such a multiple comparison adjustment to control the experimentwise type I error? What would be the total sample size required in this setting?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.9875} = 2.241$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 2.241 + 1.960 = 4.201$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 4\sigma^2 = 4 \times 1.553^2 = 9.647$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 4.201^2 \times 9.647 / 0.7500^2 = 302.7$, so round up to 304 for each subgroup. We would then require 608 subjects overall.

c. Suppose now that we actually hypothesize that DFMO is associated with a 2.50 micromole/mg protein difference in mean spermidine levels in males, but that females are unaffected by DFMO. We wish to test for such an effect modification by sex (this is a single hypothesis test, so no need for multiple comparison adjustments). Because we would merely be comparing the difference of treatment effect in males (where the average variability is $V = 4\sigma^2$ as given above) and females (where the average variability is again $V = 4\sigma^2$ as given above), and because the estimated treatment effects are independent, we know that the average variability for the difference of the estimated treatment effects will just be the sum of the average variability for each subgroup estimate, and then we would multiply by 2 because we will have to have the sample size in both males and females. Hence, the average variability needed to detect this interaction could be based on the standard formula with $V = 16\sigma^2$. What sample size is required to establish the existence of this interaction with 95% confidence (97.5% power)?

Ans:

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -2.500$; $\Delta = -2.500 - 0 = -2.500$.

To find V: $V = 16\sigma^2 = 16 \times 1.553^2 = 38.59$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 38.59 / 2.500^2 = 94.9$, so round up to 96.

NOTE: The above answer is the appropriate for the question I asked using a design alternative of -2.50. The following answer is appropriate for the question I meant to ask: I meant to use a difference of -0.75 as the design alternative, in order that you could see the huge sample size needed to detect the interaction.)

To find $\delta_{\alpha\beta}$: $z_{1-\alpha} = z_{0.975} = 1.960$; $z_\beta = z_{0.975} = 1.960$; $\delta_{\alpha\beta} = 1.960 + 1.960 = 3.920$.

To find Δ: $\theta_0 = 0$; $\theta_1 = -0.7500$; $\Delta = -0.7500 - 0 = -0.7500$.

To find V: $V = 16\sigma^2 = 16 \times 1.553^2 = 38.59$

To find N: $N = \delta_{\alpha\beta}^2 V / \Delta^2 = 3.920^2 \times 38.59 / 0.7500^2 = 1054.2$, so round up to 1056.

Some general comments on subgroups and tests for interactions:

- **Note the difference between the scientific question asked in the subgroup analyses in parts a and b and the test for interaction in part c. A subgroup analysis is asking whether the treatment works in a subset of the population. The test for interaction is asking whether the treatment works differently between two subgroups. Answering the latter question requires a very large sample size.**

- **The adjustment made for multiple comparisons in part b is called a "Bonferroni correction". It just divides your level of significance by the number of analyses you perform. This protects against mutually exclusive analyses (i.e., cases where a type I error could not be made simultaneously in two analyses). This is very conservative as you get to large number of analyses, but as can be seen above, it is not all that conservative with respect to two independent analyses: We would use a level 0.0125 test with Bonferroni, and a level 0.01258 test for two independent analyses. The latter choice would have corresponded to a critical value of 2.239 and have allowed a sample size of 302.4—very little different from the Bonferroni correction.**
- **Sometimes in clinical trials we have a single subgroup that we would investigate when the treatment is not effective in the whole sample. In these cases, we can perform an adjustment that is much improved over the Bonferroni by using the same methods we might use in group sequential monitoring of the trial.**