Statistics 641, Fall 2009
Homework #4
Answers

1. Suppose we conduct a two period, two treatment (A and B), crossover trial in asthma patients. The primary outcome is FEV_1 at the end of each treatment period. (FEV_1 is *Forced Expiratory Volume in 1-second*, a measure of lung function that it is sensitive to airway constriction.) The final data can be found in the data file "data4a.csv" (in csv format, comma delimited). This dataset contains one record per subject with variables:

   - `seq`: Order of treatment allocation (AB or BA)
   - `baseline`: baseline FEV_1
   - `perI`: observed FEV_1 at the end of period I
   - `perII`: observed FEV_1 at the end of period II

   (a) Assuming no carryover effect, compute the estimate of the treatment effect, its standard error, and the t-statistic for the observed FEV_1.

   In R, read the data: `> D <- read.csv("xover.csv")`.
   We create a new variable which is the response on treatment B minus the response on treatment A: `> D$diff <- ifelse(D$seq=="AB", 1, -1)*(D$perII-D$perI)`
   (there are lots of other ways to do this). The average of the means within groups is the estimate of treatment difference, $\hat{\Delta}$:

```
> tapply(D$diff, D$seq, mean)
         AB          BA
 0.39600000 -0.04571429
> mean(tapply(D$diff, D$seq, mean))
[1] 0.1751429
```

   (Note that, because the groups are balanced, $\hat{\Delta}$ can also be estimated by the overall mean of `D$diff`).
   The standard error of the estimate is obtained from the pooled within-group variance:

```
> tapply(D$diff, D$seq, var)
       AB        BA
0.1972071 0.1863487
> mean(tapply(D$diff, D$seq, var))
[1] 0.1917779
```

   (Again, this works because the groups are balanced—you can just average the within-group variances. Otherwise, you need a weighted average, weighted by the degrees of freedom for each group.)
   Therefore, $\text{Var}(\hat{\Delta}) = 0.19178/(2 \times 35) = 0.0027397$, and the t-statistic is

$$\frac{0.17514}{\sqrt{0.0027397}} = 3.346$$

   Note that because we have balance, the same result can be achieved by creating a new variable, say z, that is $+1$ for sequence "AB" and -1 for sequence "BA". In the regression of `diff` against z, the intercept term is the effect of interest:

```
> D$z <- ifelse(D$seq=="AB",1,-1)
> summary(lm(diff~z, data=D))
Call:
lm(formula = diff ~ z, data = D)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17514    0.05234   3.346  0.00134 **
z            0.22086    0.05234   4.220 7.42e-05 ***
```

(b) Because of variability in baseline FEV_1, investigators often use percent change from baseline rather than the raw FEV_1 score. Repeat (a) using percent change from baseline. Given this result, comment on the necessity/advisability of using percent change in a crossover trial.

Create new variable corresponding to percent change from baseline:

```
> D$diff.perc <- D$diff/D$baseline*100
> summary(lm(diff.perc~z, data=D))
Call:
lm(formula = diff.perc ~ z, data = D)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.573      2.234   2.943  0.00445 **
z              7.400      2.234   3.313  0.00148 **
```

The t-statistic is smaller than in the previous part. People like to use percent change from baseline for two reasons: 1) if there is significant variability between subjects, change from baseline removes between subject variability (at least the additive component), and 2) since much of the variability in pulmonary function is related to body size, we expect larger changes in people large body size—dividing by baseline attempts to acount for this multiplicative coomponent. In a crossover design, however, using change from baseline accomplishes nothing, since the baseline values cancel when we compute the difference between periods. (We also know that unless between subject variability is sufficiently large, change from baseline is noisier than the follow-up value alone, so this doesn't necessarily work anyway.) Dividing by baseline is helpful only to the degree that baseline reliably predicts the magnitude of changes. In this case, either the magnitude of the changes is unrelated to baseline, or the noise in the baseline measurment was large relative to the systematic differences that in the end the signal to noise ratio (the t statistic) actually decreased.

(c) To guard against the possibility of a carryover effect, estimate the treatment difference, its standard error, and the corresponding t-statistic using only the period I FEV_1. Compare the standard error estimate to the standard error estimate from (a).

```
> summary(lm(perI~seq, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07829    0.09441  32.605   <2e-16 ***
seqBA        0.06457    0.13352   0.484     0.63
```

The mean difference is in period I is only 0.0646 (as opposed to 0.1751 from the crossover), and its standard error is 0.1335 (as opposed to 0.05234 from the crossover), and the t statistic is much smaller. The standard error is more than 2.5 times as large.

(d) Using both the period I and baseline values of FEV_1, find the best estimate of the treatment effect, its standard error and the t-statistic. Again, compare the standard error estimate to the standard error estimate from (a).

The best estimate is obtained using a regression model with baseline as the covariate.

```
> summary(lm(perI~seq+baseline, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.36227    0.16666   8.174 1.17e-11 ***
seqBA        0.10690    0.08060   1.326    0.189
baseline     0.63576    0.05803  10.955  < 2e-16 ***
```

Therefore, accounting for baseline, the estimate is 0.107 with a standard error of 0.0806. By optimally accounting for baseline, the standard error is reduced to just over 1.5 times that of the crossover.

(Note that the observed mean differences between A and B in the first period are much smaller than in the second period (.0646 versus .2857). This difference accounts for part of the decrease in the t-statistic when only period I is used. It is impossible to know from the data if the difference is random—due simply to between-subject variablity—or systematic—due to carryover effect. Therefore, unless we are certain that there cannot be significant carryover, we are forced to rely on the period I result alone).

2. Suppose that we have a binary outcome, and wish to show non-inferiority of treatment B relative to treatment A. In designing the trial we assumed that the failure rate in each treatment groups is $\pi_A = \pi_B = 0.30$. Given these rates, we consider that an increase in failure rate to 0.36 to constitute non-inferiority and enroll 1300 subjects in each treatment group.

   We can parameterize this margin of interest in (at least) two ways:

   - $\delta = 0.36 - 0.30 = .06$
   - $\delta = \log(OR) = \log(\pi_B/(1-\pi_B)) - \log(\pi_A/(1-\pi_A)) = .272$

Suppose at the trial's end we observe the following:

	failures	successes
A	273	1027
B	299	1001

(a) Construct a 95% confidence interval for $\pi_B - \pi_A$. Does this interval contain $\delta = 0.06$?

We have $\hat{\pi}_A = 273/1300 = .21$ and $\hat{\pi}_B = 299/1300 = .23$. Variances are $0.21 \times 0.79/1300 = 0.0001276$ and $0.23 \times 0.77/1300 = 0.0001362$ for groups A and B respectively. The 95% CI is $.23 - .21 \pm \sqrt{0.0001276 + 0.0001362} \times 1.96 = (-0.0119, 0.0519)$. This interval does not include 0.06, so we can conclude that B is not-inferior to A at the 95% confidence level.

(b) Construct a 95% confidence interval for $\log(OR)$. Does this interval contain $\delta = 0.272$.

If $\beta = \log(OR)$, then $\hat{\beta} = \log(299/1001) - \log(273/1027) = 0.1166$. The variance of $\hat{\beta}$ is $1/273 + 1/1027 + 1/299 + 1/1001 = 0.008980$ (delta method), so the 95% CI is $.1166 \pm \sqrt{0.008980} \times 1.96 = (-0.0691, 0.3023)$ This interval does include 0.272, so we cannot conclude that B is non-inferior to A at the 95% confidence level.

(c) Why do the results of (a) and (b) differ? Comment on the sensitivity of the non-inferiority hypothesis to the choice of scale (parameterization).

Unlike a null hypothesis of equality, the hypothesis of inferiority (true treatment difference larger than $\delta > 0$) depends on the parameterization. In the case of equality $\pi_B - \pi_A = 0$, $\log(\pi_B/\pi_A) = 0$ and $\log[\pi_B(1-\pi_A)/\pi_A(1-\pi_B)] = 0$ are all equivalent. In the non-inferiority case, we replace the "=0" with $\geq \delta$ (for properly defined δs), and they are no longer equivalent.

In this example, the observed rates $\hat{\pi}_A$ and $\hat{\pi}_B$ are much lower than expected, so the difference, $\hat{\pi}_B - \hat{\pi}_A$ is proportionally larger than expected and as the underlying rates decrease the variance (proportional to $\pi(1-\pi)$) decreases, shrinking the length of the confidence interval, making it easier to exclude δ for a fixed difference, $\pi_B - \pi_A$.

On the other hand, as the rates decrease, the expected cell counts (x) in the failure column decrease, increasing their contribution to the variance ($1/x$), whereas since the counts in the success column are already much larger, the corresponding decrease in the contribution to the variance due to increases in these cell counts do not offset the increases from the first column (i.e., $1/273 + 1/1027 > 1/390 + 1/910$). Hence, the variance of the observed log(OR) increases as the rates decrease, increasing the width of the confidence interval. Hence, it is more difficult to conclude non-inferiority on the log(OR) scale if the observed rates are lower than expected.

3. Suppose that we have a binary outcome, and wish to show superiority of treatment B relative to treatment B. In designing the trial we assumed that the failure rates are $\pi_A = 0.36$ and $\pi_B = 0.30$ in treatment groups A and B respectively.

   (a) Find the sample size required to detect the difference in rates above with 90% power at two-sided $\alpha = .05$.

   $\bar{\pi} = (.36 + .30)/2 = .33$, so

   $$N = \frac{(1.96 + 1.28)^2 \times .33 \times .67 \times 4}{(.36 - .30)^2} = 2579$$

   With equal sized groups, use N = 2580, or 1290 per group.

   (b) Suppose that the true control rate, π_A is different than expected. For the sample size found in (a),

      i. plot power as a function of true π_A for $.07 \leq \pi_A \leq 0.5$ under the assumption of constant risk difference $\pi_A - \pi_B = 0.06$ and

      ii. on the same figure, plot power as a function of true π_A under the assumption of constant log-odds ratio $\log(\pi_A/(1-\pi_A)) - \log(\pi_B/(1-\pi_B) = 0.272$.

   Why do these curves differ in the way that they do?

   Power can be found by solving the sample size equation for $Z_{1-\beta}$ and calculating the corresponding value of $1 - \beta$.

   $$Z_{1-\beta} = \frac{\sqrt{2780}(\pi_A - \pi_B)}{2\sqrt{\bar{\pi}(1-\bar{\pi})}} - Z_{1-\alpha/2}$$

   so

   $$1 - \beta = \Phi\left[\frac{\sqrt{2780}(\pi_A - \pi_B)}{2\sqrt{\bar{\pi}(1-\bar{\pi})}} - Z_{1-\alpha/2}\right]$$

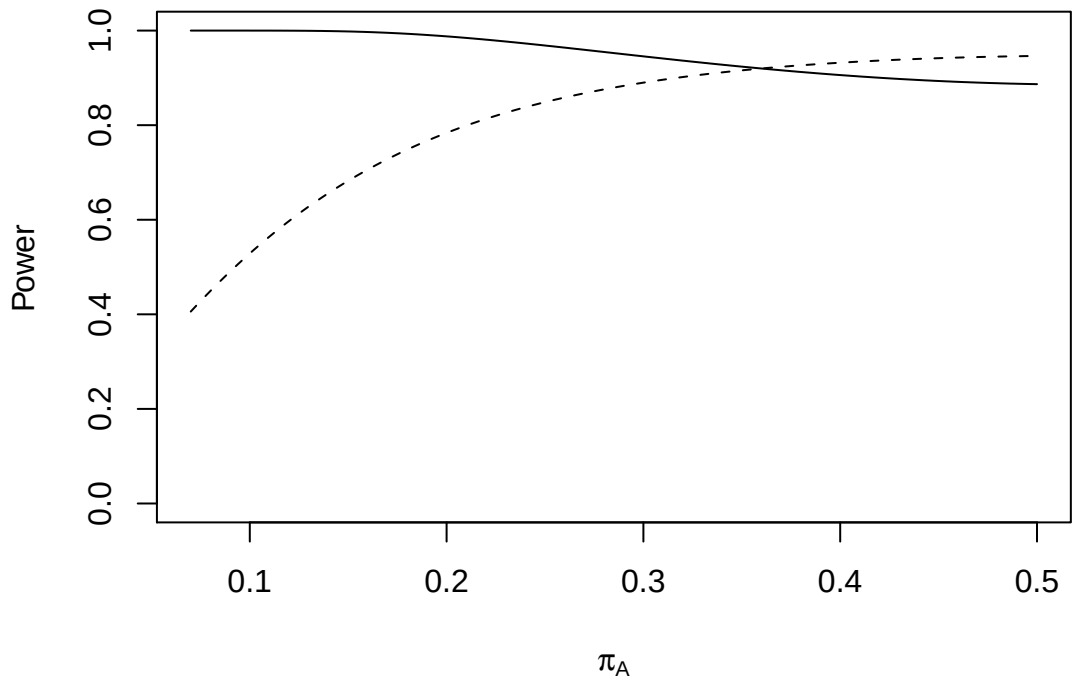
   where Φ is the standard normal CDF.

   If we choose π_A, and fix the risk difference to be 0.06, we have

   $$1 - \beta = \Phi\left[\frac{\sqrt{2780} \times 0.06}{2\sqrt{(\pi_A - .03)(1.03 - \pi_A)}} - Z_{1-\alpha/2}\right]$$

   If we choose π_A, and fix the odds-ratio to be $.3 \times .64/.7 \times .36 = 0.7619$, we have that

   $$\pi_B = \frac{0.7619\pi_A}{1 - .2381\pi_A}$$



As in the previous problem, for fixed risk difference, the variability in $\hat{\pi}_A - \hat{\pi}_B$ decreases as $\hat{\pi}_A$ decreases, so the standardized difference, $|\pi_A - \pi_B|/\sqrt{\hat{\pi}_A - \hat{\pi}_B}$ increases, with a corresponding increase in power. On the other hand, for fixed OR, the standardized difference, $|\log(OR)|/\sqrt{\text{Var}(\log(\widehat{OR}))}$ decreases with decreasing π_A, with a corresponding decrease in power.