Statistics 641, Fall 2014
Homework #6
Solutions

1. The dataset `data6.csv` contains data collected from a crossover study with 40 subjects per sequence. The variables in the dataset are:

`seq`	Assigned treatment sequence
`y`	Response
`id`	Subject id
`period`	Period
`z`	Treatment ("A" or "B")

   (a) Calculate the means within each treatment group separately for periods 1 and 2. Using these means, calculate the estimate of the treatment difference assuming no carryover.

```
> m <- with(data, tapply(y, list(z, period), mean))
> m
        1       2
A  9.6625 22.7025
B 10.1000 22.9200
# difference in "AB" group:
> m[2,2]-m[1,1]
[1] 13.2575
# difference in "BA" group:
> m[2,1]-m[1,2]
[1] -12.6025
# mean difference:
> (m[2,2]-m[1,1] + m[2,1]-m[1,2])/2
[1] 0.3275
```

   Note that the estimate of period effect is:

```
> (m[2,2]-m[1,1] - (m[2,1]-m[1,2]))/2
[1] 12.93
```

   (b) Fit a regression model that estimates the treatment difference and its standard error.

```
> summary(lm(y ~ z + period + id, data=data))

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept) -3.409e+00  6.695e-01    -5.091 2.41e-06 ***
zB           3.275e-01  1.412e-01     2.320 0.022947 *
period       1.293e+01  1.412e-01    91.602  < 2e-16 ***
ids02        1.750e+00  8.927e-01     1.960 0.053535 .
...
```

```
ids80       -1.400e+00  8.927e-01    -1.568 0.120878
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

The estimate of treatment effect is 0.3275 with SE 0.1412. The p-value is 0.0229.

Alternative using mixed-effects model from package `lme4`:

```
> lmer(y ~ z + period + (1|id), data=data)
Linear mixed model fit by REML
...
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 9.2589}  3.04284
 Residual             0.79697  0.89273
Number of obs: 160, groups: id, 80

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -3.2125     0.4129   -7.78
zB            0.3275     0.1412    2.32
period       12.9300     0.1412   91.60

Correlation of Fixed Effects:
       (Intr) zB
zB     -0.171
period -0.513  0.000
```

(c) Fit a regression model to estimate the effect of treatment using only period 1. (This is equivalent to a parallel group trial in which subjects are assigned only one of "A" or "B.")

```
> summary(lm(y ~ z, data=data, subset=period==1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6625     0.4734  20.412   <2e-16 ***
zB            0.4375     0.6695   0.654    0.515
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) Comment on the differences between the analyses in parts (b) and (c).

The standard errors of the two estimates are quite different

- Cross-over analysis: SE = 0.1412
- Parallel group (period 1 only): SE = 0.6695

This suggests that there is high correlation between the period 1 and period 2 observations from each subject. Because the cross-over model is based on within-subject differences, the subject-level effects are accounted for and the variability is significantly reduced providing greater power.

Even though the point estimate of the difference is larger in period 1 analysis, the increased variance results in a statistically in-significant difference.

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 $p_A = p_B = 0.30$. 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 the non-inferiority margin in (at least) two ways:

   - $\delta = 0.36 - 0.30 = 0.06$
   - $\delta = \log \dfrac{p_B(1-p_A)}{(1-p_B)p_A} = \log(0.36/0.64) - \log(0.30/0.70) = .272$

   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 $p_B - p_A$. Does this interval contain $\delta = 0.06$?

We have $\widehat{p}_A = 273/1300 = .21$ and $\widehat{p}_B = 299/1300 = .23$. 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 $\widehat{\beta} = \log(299/1001) - \log(273/1027) = 0.1166$. The variance of $\widehat{\beta}$ 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 $p_B - p_A = 0$, $\log(p_B/p_A) = 0$ and $\log[p_B(1-p_A)/p_A(1-p_B)] = 0$ 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 $\widehat{p}_A$ and $\widehat{p}_B$ are much lower than expected, so the difference, $\widehat{p}_B - \widehat{p}_A$ is proportionally larger than expected and as the underlying rates decrease the variance (proportional to $p(1-p)$) decreases, shrinking the length of the confidence interval, making it easier to exclude δ for a fixed difference, $p_B - p_A$.

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. Heart patients have a greater risk of a second heart attack (MI) or death immediately following the first MI, then they will later on. Suppose (simplistically) that with standard therapy, the hazard rate λ is constant .10/year during the first 6 months following MI and constant .04/year thereafter. Suppose further that a new treatment is expected to reduce these rates by 25%. We wish to perform a study of patients enrolled immediately following an MI with (uniform) recruitment and followup of either:

- (i) 1.5 year recruitment, 4 year followup
- (ii) 2 year recruitment, 3.5 year followup

(a) Calculate the required number of events to achieve 90% power at $\alpha = 0.05$. Assume equal numbers of patients in each treatment group.

Schoenfeld's formula gives the number of events required to achieve the desired power is (assume 90% power and $\alpha = 0.05$).

$$\frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\xi_1 \xi_2 \log(r)^2} = \frac{3.24^2}{(1/2)^2 \log .75^2} = 507.4$$

(b) Compute the required sample size (number of subjects) for each of the two designs. *Hint: You need to compute the probability that a subject will experience an event during the trial. The hazard function is piecewise linear.*

To compute the sample size, we need to find the number of subjects required to reach this number of events. There are a couple approaches, the simplest is to use the average hazard. The hazards for the active treatment group is 0.75/year for the first 6 months and 0.03/year thereafter. The average of the control and active hazards is 0.875/year during the first 6 months and 0.035/year thereafter. The cumulative hazard is

$$\Lambda(t) = \begin{cases} 0.0875t & \text{if } t < 0.5 \\ 0.0875 \times .5 + 0.035(t - .5) = 0.02625 + 0.035t & \text{if } t \geq 0.5 \end{cases}$$

The probability that a subject experiences an event before time t is $1 - e^{-\Lambda(t)}$.

If the total length of follow-up is F and the length of the recruitment period is R, then (because $F - R > .5$) the probability of an event is

$$\begin{aligned} \bar{\rho} &= \frac{1}{R} \int_{F-R}^{F} 1 - e^{-\Lambda(u)} du \\ &= 1 - \frac{e^{-.02625}}{R} \int_{F-R}^{F} e^{-.035u} du \\ &= 1 + \frac{e^{-.02625}}{0.035R} \left(e^{-.035 \times F} - e^{-.035 \times (F-R)} \right). \end{aligned}$$

For the two scenarios we have

i. $\bar{\rho} = 0.1305$

ii. $\bar{\rho} = 0.1073$

Therefore, the required number of subjects for each of the two scenarios is:

i. $507.4/0.1305 = 3887$

ii. $507.4/0.1073 = 4727$

4. 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 $p_A = 0.36$ and $p_B = 0.30$ 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 $\alpha = .05$.

   $\bar{p} = (.36 + .30)/2 = .33$, 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, p_A is different than expected. For the sample size found in (a),

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

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

   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 $1 - \beta$.

$$Z_{1-\beta} = \frac{\sqrt{2580}(p_A - p_B)}{2\sqrt{\bar{p}(1-\bar{p})}} - Z_{1-\alpha/2}$$

so

$$1 - \beta = \Phi\left[\frac{\sqrt{2580}(p_A - p_B)}{2\sqrt{\bar{p}(1-\bar{p})}} - Z_{1-\alpha/2}\right]$$

where Φ is the standard normal CDF.

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

$$1 - \beta = \Phi\left[\frac{\sqrt{2580} \times 0.06}{2\sqrt{(p_A - .03)(1.03 - p_A)}} - Z_{1-\alpha/2}\right]$$

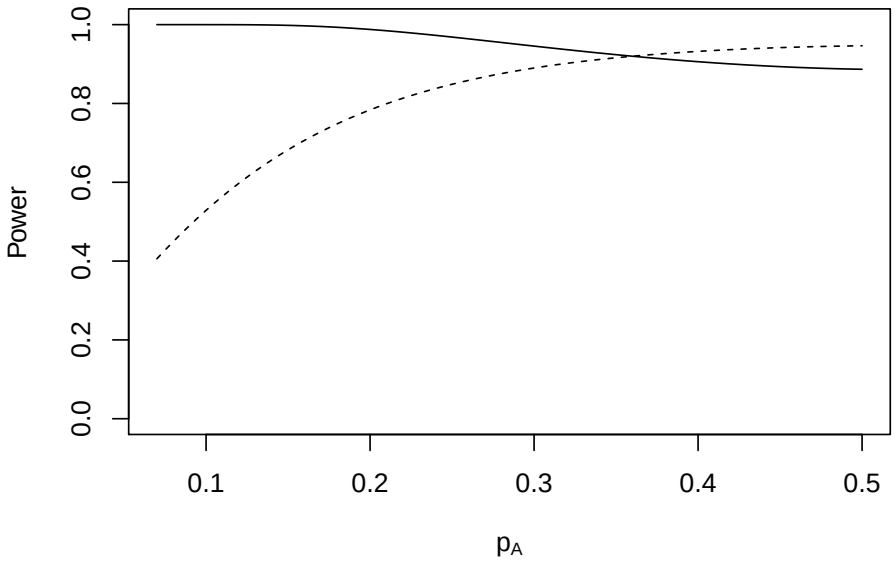
If we choose p_A, and fix the odds-ratio to be $.3 \times .64/.7 \times .36 = 0.7619$, we have that

$$\frac{p_B}{1-p_B} = \frac{0.7619 p_A}{1 - p_A} \qquad \text{and solving for } p_B, \qquad p_B = \frac{0.7619 p_A}{1 - .2381 p_A}$$

R code:

```
> pA <- 7:50/100
> pBi <- pA -.06
> pBii <- exp(-.272)*pA/(1+(exp(-.272)-1)*pA)
```

```
> pBari <- pA-.03
> pBarii <- (pA+pBii)/2
> plot(pA, pnorm(sqrt(2580)*.06/2/sqrt(pBari*(1-pBari))-1.96),
+   type="l",ylim=c(0,1), ylab="Power",xlab=expression(pi[A]))
> lines(pA, pnorm(sqrt(2580)*sqrt(pBarii*(1-pBarii))*.272/2-1.96),
+   lty=2)
> abline(h=.9,lty=3)
> legend("bottomright", bty="n", lty=1:2, c("Constant Difference in Rates",
+   "Constant Odds Ratio"))
```



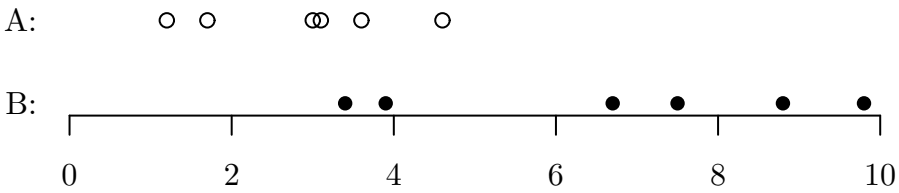
As in the previous problem, for fixed risk difference, the variability in $\widehat{p}_A - \widehat{p}_B$ decreases as $\widehat{p}_A$ decreases, so the standardized difference, $|p_A - p_B|/\sqrt{\widehat{p}_A - \widehat{p}_B}$ increases, with a corresponding increase in power. On the other hand, for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

5. Suppose that we observe the following continuous responses for groups A and B.

   A: 3.0, 1.7, 4.6, 3.6, 1.2, 3.1
   B: 3.9, 6.7, 7.5, 3.4, 8.8, 9.8

   Let $\widehat{\mu}$ be then difference in means, $\widehat{\mu} = \overline{X}_B - \overline{X}_A$. Compute a one-sided p-value under the randomization distribution generated by the random allocation rule ($N_A = N_B$). (Note that you don't need to generate the entire randomization distribution for $\widehat{\mu}$.) Compare to p-values from a two-sample t-test and Wilcoxon rank-sum test (since these are two-sided by default, you'll need to divide the p-values from these tests by 2).

First, we have $\widehat{\mu} = 6.683 - 2.867 = 3.817$. A plot of the data shows that most of the observations in group B are well to the right of the observations in group A.



There are only small number of treatment allocations that produce a more extreme mean difference than the one observed, specifically, we need only consider allocations such that the mean in group B is larger than the observed mean, 6.683. These are enumerated below:

Allocated to B	mean
3.9, 4.6, 6.7, 7.5, 8.8, 9.8	6.883
3.6, 4.6, 6.7, 7.5, 8.8, 9.8	6.833
3.4, 4.6, 6.7, 7.5, 8.8, 9.8	6.800
3.1, 4.6, 6.7, 7.5, 8.8, 9.8	6.750
3.0, 4.6, 6.7, 7.5, 8.8, 9.8	6.733^*
3.6, 3.9, 6.7, 7.5, 8.8, 9.8	6.717
3.4, 3.9, 6.7, 7.5, 8.8, 9.8	6.683

Therefore, there are 7 allocations yielding a mean difference at least as extreme as the one observed. There are a total of

$$\binom{12}{6} = 924$$

possible allocations, so the one-sided p-value is $7/924 = .0076$.

t-test and Wilcoxon rank-sum test:

```
> u <- c(3.9, 6.7, 7.5, 3.4, 8.8, 9.8) ; v <- c(3.0, 1.7, 4.6, 3.6, 1.2, 3.1)
> t.test(u,v)
Welch Two Sample t-test
data:  u and v
t = 3.2591, df = 7.203, p-value = 0.01334
...
> wilcox.test(u,v)
Wilcoxon rank sum test
data:  u and v
W = 33, p-value = 0.01515
alternative hypothesis: true location shift is not equal to 0
```

One-sided p-values are $0.01334/2 = 0.0067$ (t-test) and $0.01515/2$=0.0076, and these are comparable to that above. In fact, by default, `wilcox.test` uses the randomization distribution in small samples, and this p-value is identical to that above. This suggests that the number of sequences with rank-sum at least as large as that in the original sample is the same as the number with mean difference at least as large. Interestingly, these are *not* the same sequences. For the actual allocation, the rank sum for group B is 54, but the rank sum using the starred (*) allocation in the table above is 53 (less extreme), whereas the rank sum allocating 3.6, 3.9, 4.6, 7.5, 8.8, 9.8 to B has rank-sum 54 (at least as extreme), but group mean $6.367 < 6.683$.