Statistics 641, Fall 2013
Take Home Final Exam
Solutions

1. Suppose that we have a binomial response with expected failure probabilities $\pi_0 = 0.2$ and $\pi_1 = 0.1$ for treatments 0 (control) and 1 (experimental) respectively. Calculate the required sample sizes to achieve 90% power (type II error rate of 10%) at level $\alpha = .01$ assuming

   (a) Equal allocation to groups 0 and 1

   $$N = \frac{\bar{\pi}(1-\bar{\pi})(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\xi_0\xi_1(\pi_0 - \pi_1)^2}$$

   and with equal allocation $\bar{\pi} = (\pi_0 + \pi_1)/2 = 0.15$.

   $$\frac{4 \times .15(1-.15)(2.58+1.28)^2}{(0.2-0.10)^2} = 759.9$$

   so $N = 760$.

   (b) 2:1 randomization to groups 0, 1 respectively

   $\bar{\pi} = (2\pi_0 + \pi_1)/3 = 0.167$ and

   $$\frac{9 \times .167(1-.167)(2.58+1.28)^2}{2(0.2-0.10)^2} = 931.2$$

   so $N = 932$.

   (c) 1:2 randomization to groups 0, 1 respectively

   $\bar{\pi} = (\pi_0 + 2\pi_1)/3 = 0.133$ and

   $$\frac{9 \times .133(1-.133)(2.58+1.28)^2}{2(0.2-0.10)^2} = 774.8$$

   so $N = 775$.

   (d) Why are these different?

   N is sensitive to $\bar{\pi}$, which in turn depends on ξ_j. When more subjects are allocated to the treatment with the lower rate, fewer subjects are required to detect a fixed difference in probabilities.

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 the non-inferiority margin in (at least) two ways:

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

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 $\pi_B - \pi_A$. 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 we conduct a 2-arm randomized trial in three stages of 20 subjects each. The outcome is binomial with expected failure probabilities of $p_C = .6$ in the control group (C) and $p_E = .2$ in the experimental group (E). For purposes of this problem, the overall type I error rate will be $\alpha = 0.1$. For each stage we observe a two-by-two table of the following form:

	Failure	Success	Total
Control	x_k	$\cdot$	10
Experimental	$\cdot$	$\cdot$	10
	m_k	$\cdot$	20

Let $y_k = \sum_{u=1}^{k} x_k$ (cumulative number of failures in group C). At each stage we will stop and reject $H_0\colon p_C = p_E$ if $y_k \geq b_k$ for some b_k.

(a) Compute the contribution to the expected information from each stage (will be the same for all stages since they're the same size). Compute the total expected information (the sum from the three stages.)

Conditional on m_k, x_k is hypergeometric so information is

$$\frac{10 \times 10 m_k(20 - m_k)}{20^2 \times 19} = \frac{m_k(20 - m_k)}{76}. \tag{1}$$

(I also accept the unconditional variance, $m_k(20-m_k)/80$. This just re-scales information and has no effect on the information fractions.)

If $p_C = 0.6$ and $p_E = .2$, $\bar{p} = 0.4$, so $E[m_k] = 20 \times 0.4 = 8$. Therefore, expected information for stage k is

$$\frac{m_k(20 - m_k)}{76} = \frac{8 \times 12}{76} = 1.263$$

Full information is $3 \times 1.263 = 3.789$.

(b) Suppose that we observe $m_1 = 7$, $m_2 = 12$ and $m_3 = 9$. Compute the cumulative information fractions, t_1 and t_2, after stages 1 and 2 based on the total expected information from part (a).

Using equation (1) above,

$$\frac{m_k(20 - m_k)}{76} = \begin{cases} \dfrac{7(20-7)}{76} = 1.197 & \text{for } k = 1 \\ \dfrac{12(20-12)}{76} = 1.263 & \text{for } k = 2 \end{cases}$$

The cumulative information fractions are $t_1 = 1.197/3.789 = 0.316$ and $t_2 = (1.197 + 2.163)/3.789 = 0.649$.

(c) Using a one-sided α-spending function of the form $g(t) = .1t^2$ and the information fractions from part (b), compute the cumulative boundary crossing probabilities for stages 1 and 2.

$g(t_1) = .1 \times 0.316^2 = 0.00998$ and $g(t_2) = .1 \times 0.649^2 = 0.0422$.

(d) Using the functions `dhyper` and/or `phyper`, find b_1 so that the $\Pr\{y_1 \geq b_1\} \leq g(t_1)$

Let $f_k(x) = \Pr\{x_k = x\}$ and $F_k(x) = \Pr\{x_k \leq x\}$. Given $m_1 = 7$, $x_1 \leq 7$ and is hypergeometric, equivalent to drawing 10 balls from an urn with 7 black and 13 white. We need $\Pr\{y_1 \geq b_1\} = 1 - F(b_1 - 1) \leq 0.00998$. We have the following values for F:

```
> 1-phyper(0:7, 7, 13, 10)
[1] 0.998452012 0.971362229 0.825077399 0.500000000 0.174922601
[6] 0.028637771 0.001547988 0.000000000
```

Since $1 - F_1(5) > 0.00998$ and $1 - F_1(6) < 0.00998$, the smallest such b_1 is $b_1 = 7$.

(e) Similarly, find b_2 and b_3 so that $\Pr\{y_2 \geq b_2\} \leq g(t_2) - g(t_1)$ and $\Pr\{y_3 \geq b_3\} \leq g(t_3) - g(t_2)$.

Analogous to the calculations for a phase II trial (lecture 16, page 4 or lecture 23, pages 16-18), given $m_2 = 12$, $x_2 \leq 13$ and is hypergeometric. We want $\Pr\{y_2 \geq b_2\} \leq 0.0422 - 0.00998 = 0.03222$

$$\Pr\{y_2 \geq b_2\} = \sum_{u=0}^{6} f_1(u)(1 - F_2(b_2 - 1 - u))$$

Try $b_2 = 12, 13$ (unless you get lucky, you probably need to try more of these, but these are the important ones):

```
> sum(dhyper(0:6, 7,13,10)*(1-phyper(11-0:6,12,8,10)))    ## b2=12
[1] 0.09668775
> sum(dhyper(0:6, 7,13,10)*(1-phyper(12-0:6,12,8,10)))    ## b2=13
[1] 0.02455177
```

The smallest b_2 is 13.

To find b_3, we need

$$\Pr\{y_3 \geq b_3\} = \sum_{u=0}^{13} p_2(u)(1 - F_3(b_3 - 1 - u))$$

where $p_2(\cdot)$ is the sub-distribution of y_2, given the stopping boundary at stage 1 (see lecture 23, middle page 17 for phase II version)

$$p_2(y) = \Pr\{y_2 = y\} = \sum_{u=0}^{6} f_1(u) f_2(y - u)$$

Slick way:

```
> p2 <- sapply(0:12, function(y) sum(dhyper(0:6, 7,13,10)*dhyper(y-0:6,12,8,10)))
```

Less slick way (even though because $m_2 = 12$ we know $y_2 \geq x_2 \geq 2$, so we could use '`for(i in 2:12)`' in the loop):

```
> p2 <- rep(0,13)
> for(i in 0:12) p2[i+1] <-sum(dhyper(0:6,7,13,10)*dhyper(i-0:6,12,8,10))
```

Note that, because `p2` starts at $y_2 = 0$, entry i in this vector is $p_2(i-1)$.

Now we need b_3 such that $\Pr\{y_3 \geq b_3 - 1\} \leq 0.1 - 0.0422 = 0.0578$. As above, try $b_3 = 18, 19$

```
> sum(p2*(1-phyper(16-0:12, 9,11,10)))  ## b3=17
[1] 0.0749899
> sum(p2*(1-phyper(17-0:12, 9,11,10)))  ## b3=18
[1] 0.01964804
```

Thus, $b_3 = 18$.

(f) Compute Z based on the score test for $y_1 = 4, 5, 6$. Compare to the stopping boundaries based on normal data, the spending function $g(t)$, and observed information fraction. Repeat for $y_2 = 11, 12, 13$ and 14, and $y_3 = 16, 17, 18$ and 19. Compare the stopping boundaries from the exact procedure to those from the asymptotic (assuming normality) procedure

```
> tobs <- c(0.3160, 0.6493, 1)
> summary(bounds(tobs, alpha=.1, iuse=3, phi=2))
...
Boundaries:
     Time   Upper   Exit pr.  Diff. pr.
1  0.3160  2.3269  0.0099856  0.0099856
2  0.6493  1.7796  0.0421590  0.0321734
3  1.0000  1.3639  0.1000000  0.0578410
```

(Note that the observed full information is 3.7632, very close to expected of 3.7895, so I'm not worrying about making a correction for full versus expected information, it will make very little difference.)

The score function at stage k is

$$U_k(0) = x_k - m_k/2$$

and information is given by equation (1). At stage 1, we have information 1.197, $m_1/2 = 3.5$ and $Z_k = \dfrac{x_k - m_k/2}{\sqrt{\mathcal{I}_k}}$ from the table:

$y_1 = x_1$:	4	5	6	7
$U_1(0)$:	0.5	1.5	2.5	3.5
Z_1:	0.457	1.371	2.285	3.199

Hence we would reject H_0 if $y_1 \geq 7$.

At stage 2, we have

$$Z_2 = \frac{U_1(0) + U_2(0)}{\sqrt{\mathcal{I}_1 + \mathcal{I}_2}} = \frac{x_1 - 7/2 + x_2 - 12/2}{\sqrt{1.197 + 1.263}} = \frac{y_2 - 9.5}{\sqrt{2.461}}$$

and we have the table:

y_2:	11	12	13	14
$U_1(0)$:	1	2	3	5
Z_1:	0.956	1.594	2.231	2.869

we reject H_0 if $y_2 \geq 13$.

Similarly, for $k = 3$,

$$Z_3 = \frac{U_1(0) + U_2(0) + U_3(0)}{\sqrt{\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}3}} = \frac{x_1 - 7/2 + x_2 - 12/2 + x_3 - 9/2}{\sqrt{1.197 + 1.263 + 1.303}} = \frac{y_3 - 14.5}{\sqrt{3.763}}$$

and we have the table:

y_2:	16	17	18	19
$U_1(0)$:	2	3	4	5
Z_1:	1.031	1.546	2.062	2.577

we reject H_0 if $y_3 \geq 18$. In this case the stopping boundaries for both the exact and asymptotic procedures are the same at stages 1 and 2, but using the asymptotic procedure, we would reject H_0 if $y_3 \geq 17$ while using the exact procedure we would reject if $y_3 \geq 18$.

4. Suppose we conduct a trial with 3 dose groups (low, medium, high) and placebo (none), and a binary outcome (dead/alive). At the conclusion of the trial we observed the following table:

Treatment:	placebo ($i = 0$)	low ($i = 1$)	medium ($i = 2$)	high ($i = 3$)
Alive	10	18	22	25
Dead	90	82	78	75
	100	100	100	100

(a) Letting $\pi_i, i = 0, 1, 2, 3$ be the probability of death in group i (where i is defined in the table), the global null hypothesis is

$$H_{0123}\colon \pi_0 = \pi_1 = \pi_2 = \pi_3.$$

By reversing the roles of outcome and treatment, we may test this null hypothesis using the Wilcoxon rank-sum test (uncorrected). (This test is sensitive to the alternative hypothesis in which subjects who die are, for example, more likely to have received a lower dose than subjects who survive.)

Conduct this test of H_{0123} at level $\alpha = 0.05$.

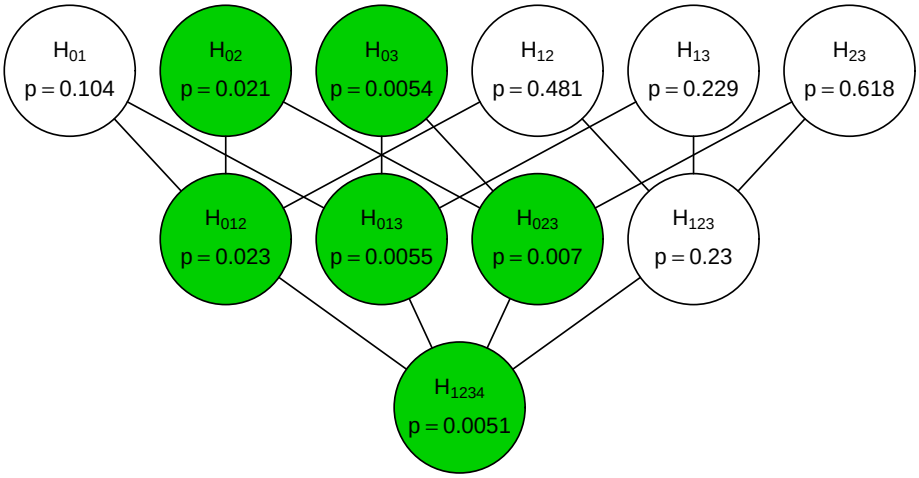
(For convenience, the dataset `dataF.csv` has one record per subject with a treatment variable, `z`, and death variable `dead`. You can perform the Wilcoxon test directly using this dataset)

```
> dataF <- read.csv("dataF.csv")
> wilcox.test(z~dead, data=dataF, correct=F)
Wilcoxon rank sum test
data:  z by dead
W = 14637.5, p-value = 0.005053
alternative hypothesis: true location shift is not equal to 0
```

(b) We can implement a closed testing procedure as follows. First test the global null hypothesis H_{0123} at level $\alpha = .05$. If we reject H_{0123}, then test each of $H_{012}, H_{013}, H_{023}$, and H_{123} where $H_{012}\colon \pi_0 = \pi_1 = \pi_2$, $H_{013}\colon \pi_0 = \pi_1 = \pi_3$, etc, at level $\alpha = .05$ using the same test restricted to the treatment groups $\{0, 1, 2\}$, $\{0, 1, 3\}$, etc. (Note that the function `Wilcox.test` takes a `subset` argument, e.g., `subset=z!=3` to use only groups 0,1,2.)

Depending upon which hypotheses H_{jkl} are rejected, we can proceed to test pairwise hypotheses $H_{jk}\colon \pi_j = \pi_k$ at level $\alpha = 0.05$. (Note that for 2 groups and 2 outcomes, the Wilcoxon test (uncorrected) and the Pearson chi-square test (uncorrected) give identical results, so either can be used.)

Draw the diagram (similar to the diagram in the Dec 11 lecture) showing the relationships between all 11 hypotheses and using this procedure, determine which of the pairwise hypotheses, H_{jk}, can be rejected at level $\alpha = 0.05$.



Tests with extraneous output deleted:

```
> > wilcox.test(z~dead, data=dataF, correct=F,subset=z!=3)
W = 7450, p-value = 0.02302
> wilcox.test(z~dead, data=dataF, correct=F,subset=z!=2)
W = 8045.5, p-value = 0.005496
> wilcox.test(z~dead, data=dataF, correct=F,subset=z!=1)
W = 8425.5, p-value = 0.006951
> wilcox.test(z~dead, data=dataF, correct=F,subset=z!=0)
W = 8337.5, p-value = 0.2303
```

```
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%0:1)
W = 2808, p-value = 0.1039
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%c(0,2))
W = 3288, p-value = 0.02096
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%c(0,3))
W = 3637.5, p-value = 0.005362
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%c(1,2))
W = 3400, p-value = 0.4806
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%c(1,3))
W = 3725.5, p-value = 0.2294
> wilcox.test(z~dead, data=dataF, correct=F,subset=z%in%c(2,3))
W = 3745.5, p-value = 0.6177
```

The hypotheses in green have p-values less than 0.05. For hypotheses H_{02} and H_{03}, all hypotheses below are also green, and therefore, the pairwise hypothese H_{02} and H_{03} can be rejected at 0.05. No other pairwise hypotheses can be rejected.

(c) Using the Bonferroni procedure for all 6 pairwise hypotheses, which can be rejected at overall level $\alpha = 0.05$?

Using the Bonferroni procedure for 6 hypothese, we reject any particular hypothesis if the corresponding p-value is less than 0.05/6=0.00833. The only one of the 6 pairwise hypotheses meeting this criterion is H_{03}.

(d) Using the Bonferroni procedure for the three pairwise hypotheses that compare an active dose to placebo, which can be rejected at overall level $\alpha = 0.05$?

Using the Bonferroni procedure for 3 hypothese, we reject any particular hypothesis if the corresponding p-value is less than 0.05/3=0.01667. The only one of the 3 pairwise hypotheses meeting this criterion is H_{03}.

(e) Using the Holm procedure for all 6 pairwise hypotheses, which can be rejected at overall level $\alpha = 0.05$?

Using Holm, we order the 6 pairwise hypotheses and compare them to 0.05/6=0.00833, 0.05/5=0.01, 0.05/4=0.0125, 0.05/3=0.0167, 0.05/2=0.025, and 0.05 respectively. The ordered p-values are 0.00536, 0.0210, 0.1039, 0.229, 0.481, and 0.618. The first (for H_{03}) is less then its critical value, 0.00833, however, the second, (for H_{02}) is not less than its critical value, 0.01, so we cannot reject it, and testing stops.

(f) Using the Holm procedure for the three pairwise hypotheses that compare an active dose to placebo, which can be rejected at overall level $\alpha = 0.05$?

Using Holm, we order the 3 pairwise hypotheses and compare them to 0.05/3=0.0167, 0.05/2=0.025, and 0.05 respectively. The ordered p-values are 0.00536, 0.0210, and 0.1039. The first (for H_{03}) is less then its critical value, 0.0167, however, the second, (for H_{02}) is also less than its critical value, 0.025, however $0.104 > 0.05$, so we can reject H_{03} and H_{02} but not H_{01}.