Statistics 641, Fall 2014
Homework #5
Solutions

1. Suppose all subjects in a randomized trial are followed for 1 year, and at the end of that time they either survive disease free (DFS), survive but experience a recurrence of disease, or die. Note that these responses are naturally ordered: DFS is good, death is bad, and recurrence is in between. Subjects are randomly assigned either treatment 1 (control) or treatment 2 (experimental). We observe the following:

	DFS	Recurrence	Dead
Treatment 1	33	17	41
Treatment 2	44	18	25

   (a) Compute "by hand" the test statistic for the Wilcoxon rank-sum test assuming that responses are ordered as shown.

   The ranks for the DFS subjects range from 1 to 77, for Recurrence from 78 to 112, and the Dead subjects from 113 to 178. Thus the mean ranks in the three groups are (1+77)/2=39, (78+112)/2=95, and (113+178)/2=145.5. The rank sum for Treatment 2 is $T_2 = 44 \times 39 + 18 \times 95 + 25 \times 145.5 = 7063.5$. The expected value of T_2 is $87 \times (178 + 1)/2 = 7785.5$, so $U = 7063.5 - 7785.5 = -723$. The variance is

$$\frac{87 \times 91}{178 - 1}\left(\frac{1}{178}(77 \times 39^2 + 35 * 975^2 + 66 \times 145.5^2 - \frac{(1 + 178)^2}{4}\right) = 101620.8$$

   The test statistic is $723^2/101620.8 = 5.14$ ($p = 0.023$). This statistic has (asymptotically) a chi-square distribution with 1 df.

   (b) What do these results suggest regarding the effect of treatment.

   Lower average rank in group 2 suggests that subjects in treatment group 2 tend to have better outcomes than those in treatment group 1, so treatment 2 is superior to treatment 1.

2. Suppose that we have 8 subjects in each of two groups. We observe the following responses:
   Control: 0.2 0.8 1.9 2.2 2.6 3.9 8.2 21.8
   Experimental: 2.8 5.1 7.1 7.7 12.3 18.8 27.1 39.7

   (a) Calculate "by hand" the Wilcoxon rank sum test statistic for the comparison of the two groups.

   Use "C" and "E" to denote the two groups. The observations are ordered as follows:

	0.2	0.8	1.9	2.2	2.6	2.8	3.9	5.1	7.1	7.7	8.2	12.3	18.8	21.8	27.1	39.7
group:	C	C	C	C	C	E	C	E	E	E	C	E	E	C	E	E
rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

   The sum of the ranks in group "E" is $T = 6 + 8 + 9 + 10 + 12 + 13 + 15 + 16 = 89$. The expected rank in group E is $8 \times 17/2 = 68$, so the $T - E[T] = 89 - 68 = 21$.

Because there are no ties, the variance is $8 \times 8 \times 17/12 = 90.667$. The chi-square test statistic is

$$\frac{21^2}{90.667} = 4.86$$

This statistic has (asymptotically) a chi-square distribution with 1 df. The p-value is 0.0274. (Note that by default the `wilcox.test` function in R uses the exact distribution, so the p-value is slightly different. Will discuss exact p-value later.)

(b) Calculate "by hand" the Mann-Whitney U-statistic for the comparison of the two groups.

First, there are no ties, so there will be no 1/2's. Second, in this table, for each observation in group "E", M_j is the number of "C" subjects that are smaller:

Experimental:	2.8	5.1	7.1	7.7	12.3	18.8	27.1	39.7
M_j	5	6	6	6	7	7	8	8

The sum of the values in the second row is $5 + 6 + 6 + 6 + 7 + 7 + 8 + 8 = 53$. The expected value is $8 \times 8/2 = 32$. Hence $U = 53 - 32 = 21$, the same as the Wilcoxon rank sum. The variance is the same as above so the rest of the computation is identical.

(c) Using your software of choice (or by hand if you wish), perform the t-test for the comparisons between the two groups.

Using the `t.test` function in *R*,

```
> t.test(c(0.2,0.8,1.9,2.2,2.6,3.9,8.2,21.8),
+  c(2.8,5.1,7.1,7.7,12.3,18.8,27.1,39.7))

Welch Two Sample t-test

data:  c(0.2, 0.8, 1.9, 2.2, 2.6, 3.9, 8.2, 21.8) and
      c(2.8, 5.1, 7.1, 7.7, 12.3, 18.8, 27.1, 39.7)
t = -1.9093, df = 10.993, p-value = 0.08265
alternative hypothesis: true difference in means is not equal to 0
```

The p-value from the t-test is 0.0827.

Note

- the data do not appear to be close to normal: there are many small values (values less than, say, 5 or 10) but several relatively large values (greater than 20). The data are skewed significantly to the right.
- The p-value using the Wilcoxon/Mann-Whitney test is much smaller than the p-value from the t-test.
- For non-normal data, the Wilcoxon/Mann-Whitney test can have greater power than the t-test, which is sensitive to large values. Because the Wilcoxon/Mann-Whitney test is based on ranks, it is insensitive to major deviations from normality.

3. The data file "data5.csv" contains columns

   - `x0`: baseline value of response variable
   - `x1`: value of response variable at first follow-up time
   - `x2`: value of response variable at second follow-up time
   - `z`: treatment variable (0,1)

   Assume that the responses are normally distributed.

   (a) For the first follow-up response (`x1`) test the null hypothesis that there is no difference by treatment by

      i. ignoring baseline

      Fit model for `x1`with just `z`:

```
> summary(lm(x1~z, data=D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   60.911      1.809  33.663   <2e-16 ***
z              3.903      2.516   1.551    0.126
```

      The mean difference is 3.90 with SE 2.52, and t-statistic 1.55.

      ii. using change from baseline (`x1-x0`)

```
> summary(lm(x1-x0~z, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2714     0.9915  -0.274   0.7853
z             3.5914     1.3786   2.605   0.0117 *
```

      The mean difference is 3.59 with SE 1.38 and t-statistic 2.61.

      iii. fitting regression model $x_1 = \alpha_0 + \alpha_1 x_0 + \beta z + \epsilon$.

```
> summary(lm(x1~z+x0, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7195     3.6147   3.796 0.000369 ***
z             3.6626     1.2251   2.990 0.004172 **
x0            0.7713     0.0573  13.461  < 2e-16 ***
```

      The mean difference is 3.66 with SE 1.23 and t-statistic 2.99.

Why do the conclusions differ from these three analysis?

In (c) the coefficient for `x0` is .77, suggesting (assuming equal variances for `x0` and `x1`) that the correlation is greater than .5, so the change from baseline should have smaller variance than the follow-up value alone. This is borne out in the differences between (a) and (b). Since this coefficient is not too close to one, we expect that the regression model in (c) should have smaller variance than either (a) or (b), and again this is borne out in the results.

(b) Repeat each of (i), (ii), and (iii) above for the response at the second follow-up time (`x2`).

```
> summary(lm(x2~z, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.618      2.063  29.868   <2e-16 ***
z              6.872      2.868   2.396     0.02 *

> summary(lm(x2-x0~z, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4357     2.6931   0.162   0.8721
z             6.5610     3.7446   1.752   0.0852 .
> summary(lm(x2~z+x0, data=D))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.7430     8.4511   6.241 6.55e-08 ***
z             6.8270     2.8644   2.383   0.0206 *
x0            0.1451     0.1340   1.083   0.2836
```

In the third analysis, the coefficient for `x0` is small, suggesting that there is much less correlation between `x2` and `x0` than between `x1` and `x0`. Therefore, we expect that the change from baseline will be much less efficient than ignoring baseline altogether. Again this is borne out in the results. In this case the third analysis gives essentially the same result as the first.

(c) Comment on the differences between (a) and (b).

Change from baseline beats observed follow-up value alone when correlation between baseline and follow-up is high, and loses when correlation is low. In either case, the regression model is at least as good as the others and should always be preferred.

4. Suppose that we have a phase II, single arm trial using a two stage design. The hypotheses of interest are H_0: $p \leq 0.15$ versus H_1: $p \geq 0.4$ where p is the true success rate. Let y_k be the total number of successes through stage k, $k = 1, 2$. Note: you can use the functions `dbinom` and `pbinom` in R to calculate binomial probabilities.

   (a) We enroll 16 subjects in stage 1 and stop and accept H_0 if we observe $a_1 = 3$ or fewer responses ($y_1 \leq 3$), otherwise we continue to stage 2. Find the stopping probabilities under both H_0 and H_1.

   Under H_0, the stopping probability is $\Pr\{y_1 \leq 3\} = .7899$, and under H_1 it is .0651.

   In R:

```
> pbinom(3, 16, .15)
[1] 0.7898907
> pbinom(3, 16, .40)
[1] 0.06514674
```

   (b) At stage 2 we enroll an additional 16 subjects and reject H_0 if $y_2 > 8$. Compute the overall probabilities of rejection under both H_0 and H_1 for the two-stage trial.

   We accept H_0 if $y_1 \leq 3$ and $y_2 \leq 8$. Under H_0 this probability is

$$\Pr\{y_1 \leq 3\} \quad + \quad \sum_{i=4}^{8} \Pr\{y_1 = i\} \Pr\{x_2 \leq 8 - i\} = .9659$$

   In R:

```
> pbinom(3, 16, .15) + sum(dbinom(4:8,16,.15)*pbinom(4:0, 16, .15))
[1] 0.9658661
```

   Under H_1, this probability is 0.0969.

```
> pbinom(3, 16, .40) + sum(dbinom(4:8,16,.40)*pbinom(4:0, 16, .40))
[1] 0.09691022
```

   Therefore, the rejection probabilities are 1-0.9659=0.0341 and 1-0.0969=0.9031 under H_0 and H_1 respectively.

   Alternatively,

```
> 1-pbinom(8,16,.15) - sum(dbinom(4:8,16,.15)*(1-pbinom(4:0, 16, .15)))
[1] 0.03413386
> 1-pbinom(8,16,.40) - sum(dbinom(4:8,16,.40)*(1-pbinom(4:0, 16, .40)))
[1] 0.9030898
### or, equivalently, (note that pbinom(x,...) is zero if x < 0)
> sum(dbinom(4:16,16,.15)*pbinom(8-4:16, 16, .15,lower=F))
[1] 0.03413386
> sum(dbinom(4:16,16,.40)*pbinom(8-4:16, 16, .40,lower=F))
[1] 0.9030898
```

(c) Compute the expected sample sizes for $p = 0.15$ and $p = 0.4$.

N is either 16 or 32, depending on whether we stop at stage 1. Under H_0,
$E[N] = 16 \times 0.7899 + 32 \times (1 - 0.7899) = 19.36$ and under H_1,
$E[N] = 16 \times 0.0651 + 32 \times (1 - 0.0651) = 30.96$.

(d) Suppose, instead, we perform a single stage trial with $N = 32$ subjects and we reject H_0 if we observe more than 8 successes. Find the type I and type II error rates. What is the advantage of the two-stage trial?

Under H_0, probability of rejection (type I error) is $\Pr\{y > 8\} = 1 - 0.9587 = .0413$, and under H_1 the acceptance probability (type II error) is $\Pr\{y \leq 8\} = 0.0575$. The type I error rate is slightly smaller for the two-stage trial, but the type II error rate is larger. The advantage of the 2 stage trial is that we can have the potential to stop earlier and discard ineffective treatments sooner.
