Statistics 641, Spring 2018
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	Total
Treatment 1	33	17	41	91
Treatment 2	44	18	25	87
	77	35	66	178

(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, suggesting that 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 and corresponding p-value 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.

(d) Comment why these three are or are not different.

- The Wilcoxon rank-sum and Mann-Whitney tests are algebraically identical, so they do not differ.
- The p-value using the Wilcoxon/Mann-Whitney test is much smaller than the p-value from the t-test.

- 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 t-test valid for non-normal data provided that the sample sizes are sufficiently large.
- 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 and is always a valid test of its null hypothesis ($P_1 + P_2/2 = 1/2$ where P_1 and P_2 are defined in the notes.)

3. The dataset `data5.csv` contains the following columns

- `y`: Continuous response variable measured at a fixed follow-up time for which larger values correspond to better outcomes. This variable is missing for dead subjects.
- `dead`: Indicator of death before the response `y` could be measured (1=dead, 0=alive).
- `z`: Randomly assigned treatment (0/1).

(a) Assess the effect of treatment on mortality.

```
> data <- read.csv("../hw/data5.csv")
> table(data$z, data$dead)
      0  1
  0 88 12
  1 66 34
> chisq.test(data$dead, data$z, correct=F)
Pearson's Chi-squared test
data:  data$dead and data$z
X-squared = 13.6646, df = 1, p-value = 0.0002185
```

The observed mortality rate in group `z=0` is 12%, while the in group `z=1` is 34%. The p-value for this difference based on the Pearson chi-square test is 0.0002, so this is strong evidence that treatment 1 increases mortality relative to treatment 0.

(b) Assess the difference between treatment groups for the non-missing responses `y` (survivors only). Does this represent a causal effect of treatment?

```
> wilcox.test(y˜z, data=data, correct=F)
Wilcoxon rank sum test
data:  y by z
W = 3161.5, p-value = 0.3471
alternative hypothesis: true location shift is not equal to 0
```

Based on the Wilcoxon rank sum test, there is no evidence of a difference among survivors in the response `y`. However, because we have implicitly conditioned on a post-randomization condition (alive), this analysis cannot tell reliably us anything about the effect of treatment on `y`. In particular, we cannot infer that there is no causal effect of treatment on `y`.

(c) By assigning a low score to death (lower than any observed y), assess the difference between treatment groups for the composite outcome of death and y. Does this represent a causal effect of treatment?

Create a variable called yDead which takes the observed value of y for survivors, and a value of 0 for deaths. Note that the smallest observed value of y is 3, so the value of 0 is smaller than all observed values.

```
> min(data$y,na.rm=T)
[1] 3
> data$yDead <- ifelse(data$dead==1, 0, data$y)
> wilcox.test(yDead~z, data=data, correct=F)
Wilcoxon rank sum test
data:  yDead by z
W = 6357.5, p-value = 0.0008456
alternative hypothesis: true location shift is not equal to 0
> aggregate(yDead~z, median, data=data)
  z yDead
1 0   9.5
2 1   8.1
```

There is a highly significantly difference between treatment groups ($p = 0.00085$). Note that the median response in group z=0 is 9.5 and the median response in group z=1 is 8.1, so we may conclude that treatment z=1 has a adverse causal effect relative to treatment z=1 on the composite outcome of death and y.

(d) Create the following summary measures for the effect of treatment on the composite outcome:

i. Difference in mean scores,

```
> aggregate(yDead~z, mean, data=data)
  z yDead
1 0 9.319
2 1 6.623
> diff(aggregate(yDead~z, mean, data=data)[,2])
[1] -2.696
```

The difference in means is -2.696 using the convention that deaths receive a score of zero. If a different numeric value is chosen for death, the mean difference will change.

ii. Difference in median scores,

```
> diff(aggregate(yDead~z, median, data=data)[,2])
[1] -1.4
```

The difference in medians is -1.4 using the convention that deaths receive a score of zero.

iii. Hodges-Lehmann estimate. Note that this can be computed using trial and error by choosing the value of `r` in a function call similar to the following that makes the one-sided *p*-value equal to .5:

```
> wilcox.test(y0,y1-r, alternative="greater")
```

where `y0` represents the score for group `z=0` and `y1` represents the score for group `z=1`. (There are other ways of formulating this and the estimate can also be found without trial and error using, for example, the function `uniroot`.)

In the following (trial and error), I've deleted unnecessary output. I try a series of values of `r` in the example above, starting with 0,1,2,3, and once we realize that the estimate is between 2 and 3, we can narrow it down relatively quickly.

```
##try r=0
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1],correct=F,
+    alternative="greater")
data:  data$yDead[data$z == 0] and data$yDead[data$z == 1]
##try r=-1
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+1,correct=F,
+    alternative="greater")
W = 5657, p-value = 0.05375
##try r=-2
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+2,correct=F,
+     alternative="greater")
W = 5169, p-value = 0.3394
##try r=3-
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+3,correct=F,
+    alternative="greater")
W = 4692, p-value = 0.7748
##try r=-2.3
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+2.3,correct=F,
+    alternative="greater")
W = 5022.5, p-value = 0.478
##try r=-2.4
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+2.4,correct=F,
+    alternative="greater")
W = 4963.5, p-value = 0.5356
##try r=-2.35
> wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]+2.35,correct=F,
+    alternative="greater")
W = 5000, p-value = 0.5
## Note that the solution is not unique: any value of  r with
##    2.3 < r < 2.4 works.
##
## Use uniroot to find solution.  Again, it is not unique, so uniroot
## grabs the first value that it finds, based on its search algorithm (after
##   only 2 tries!)
> uniroot(function(r) wilcox.test(data$yDead[data$z==0],data$yDead[data$z==1]-r,
+    correct=F, alternative="greater")$p.value-.5, c(-3,0))
$root
[1] -2.353134

$f.root
[1] 0

$iter
[1] 2

$estim.prec
[1] 0.6468656
## We can also get the answer directly from wilcox.test by setting the
## argument ‘conf.int=T’.
> wilcox.test(yDead~z, data=data, correct=F,conf.int=T)
Wilcoxon rank sum test
data:  yDead by z
W = 6357.5, p-value = 0.0008456
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.7000537 4.1999804
sample estimates:
difference in location
              2.332577
```

The Hodges-Lehmann estimate is -2.35, but as noted above this value is not unique.

How reliably does each of these estimates reflect the effect of treatment on the outcome?

First, note that the difference in means depends on the choice of the value that we assign to deaths. E.g., if we assign a value of -1000 to deaths, rather than 0, we have

```
> data$yDead1000 <- ifelse(data$dead==1, -1000, data$y)
> diff(aggregate(yDead1000~z, mean, data=data)[,2])
[1] -222.696
> diff(aggregate(yDead1000~z, median, data=data)[,2])
[1] -1.4
## Hodges-Lehmann directly using wilcox.test:
> wilcox.test(yDead1000~z, data=data, correct=F,conf.int=T)
...
difference in location
              2.399834
```

so the difference of means differs dramatically from the previous result and is highly dependent on the value that we choose for deaths. On the other hand, since both the difference in medians and the Hodges-Lehmann estimates depend only on the ranks, the imputed value has no impact (although for HL, we need to ensure that the imputed value is lower than the lowest value even after the location shift is applied. In this case, it is.) In the above, the HL estimate differs from the previous, but it is still within the interval (2.3,2.4).

Whether either the difference in medians or the HL estimate reflect an effect on y or simply an effect on mortality is unclear. As with composite failure time outcomes, the independent effect of treatment on a non-fatal outcome in the presence of mortality is not identifiable. Either of these non-parametric estimates is preferred over the mean, but it is unclear how meaningful either one is.

(e) Comment on the ability of randomized trials to assess the independent effect of treatment on a non-fatal outcome measure when study subjects die before they can be assessed.

We know two things: 1) that treatment 0 affords better survival than treatment 1, and 2) that there is a net beneficial effect on the composite of death and y. However, as noted above, it is not possible to independently assess a direct effect on y. For example, maybe treatment has no effect at all on y (or the underlying biological process that affects y) and mortality is independent of y. Or, maybe z=1 positively affects the biological process that leads to increased values of y, but also tends to kill people who would have had high values. The net effect is that, among survivors, there is no difference in observed responses, however, (in an alternate universe) had no one died, we would have seen better responses for the group z=1. Many more scenarios are also consistent with the observed data, so any conclusions regarding the effect on y should be viewed cautiously.

4. Suppose that we have a phase II, single arm trial using a two stage design. The hypotheses of interest are $H_0\colon\ p \leq 0.15$ versus $H_1\colon\ p \geq 0.4$ 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.