Statistics 641, Fall 2013
Homework #4
Solutions

1. A randomized, two-arm trial is conducted comparing a control treatment (A) to a experimental treatment (B). The primary outcome is all-cause mortality.

   (a) After completion of the trial (between 1.5 and 3 years follow-up), we observe the following table:

	Subjects	Deaths	Person-years follow-up
A	500	241	835
B	500	219	876

   Assume exponential survival and compute the hazard ratio, the Wald test statistic for the log-hazard ratio and a 95% confidence interval for the hazard ratio.

The hazard ratio is

$$\frac{\widehat{\lambda}_B}{\widehat{\lambda}_A} = \frac{219/876}{241/835} = 0.866$$

From page 5 of the notes from September 26 (lecture 8), the variance of $\widehat{\beta} = \log \widehat{\lambda}_B/\widehat{\lambda}_A$ is

$$\mathrm{Var}\left(\widehat{\beta}\right) = \frac{1}{241} + \frac{1}{291} = 0.008715$$

and the Wald statistic is

$$\frac{\widehat{\beta}^2}{\mathrm{Var}\left(\widehat{\beta}\right)} = \frac{(\log 0.866)^2}{0.00875} 2.37$$

or

$$\frac{\widehat{\beta}}{\sqrt{\mathrm{Var}\left(\widehat{\beta}\right)}} = 1.54 \sim N(0,1)$$

(This corresponds to a p-value of 0.12, so it does not reach conventional levels of statistical significance.)

You could also calculate a Wald test on the hazard ratio scale, rather than the log-hazard ratio scale. By the delta-method

$$\mathrm{Var}\left(e^{\widehat{\beta}}\right) = e^{2\widehat{\beta}}\mathrm{Var}\left(\widehat{\beta}\right) = 0.866^2 \times 0.00875 = 0.00656$$

and the test is

$$\frac{(1-0.866)^2}{0.00656} = 2.73$$

or

$$\frac{1-0.866}{\sqrt{0.00656}} = 1.65$$

The 95% CI for β is $\log 0.866 \pm 1.961\sqrt{0.008715} = (-0.327, 0.039)$, and the 95% CI for $\lambda_B/\lambda_A = (0.721, 1.040)$

(b) It is noted that many subjects do not adhere to their assigned treatment and when subjects are classified by their level of adherence, we observe the following table:

	Adherence	Subjects	Deaths	Person-years follow-up
A	$> 80\%$	110	52	180
	50%–80%	240	153	352
	$\leq 50\%$	150	36	303
B	$> 80\%$	145	51	262
	50%–80%	280	148	463
	$\leq 50\%$	75	20	151

Compute hazard ratios for subjects within each stratum based on adherence ($> 80\%$, 50%–80%, $\leq 50\%$), and note the differences with the overall comparison in part (a). Which result is more credible as an assessment of the effect of treatment and why?

The within-stratum hazard ratios are

$$> 80\%: \quad \frac{51/262}{52/180} = .674$$

$$50\%\text{-}80\%: \quad \frac{148/463}{153/352} = .735$$

$$\leq 50\%: \quad \frac{20/151}{36/303} = 1.115$$

The hazard ratios for the "better compliers" ($> 80\%$ and 50%-80% strata) are much smaller than the overall comparison (the *intention-to-treat (ITT) analysis*). It might be tempting to conclude that this analysis provides evidence of treatment benefit that is not evident in the intention-to-treat analysis, however, the within-stratum analysis is not credible for several reasons.

- The randomization ensures that the treatment assignments are independent of outcomes (thereby ensuring that the groups are comparable), and therefore, the ITT analysis is a valid test of the null hypothesis that there is no net causal effect of treatment on outcomes.
- Adherence to assigned treatment clearly depends on both treatment and outcome.
  - 110 out of 500 (22%) of group A subjects and 145 out of 500 (29%) of group B subjects were in the $> 80\%$ groups, so these are almost certainly non-comparable subsets. Similarly, the group A and group B subjects within the other two strata are not comparable.
  - The event rates clearly differ by adherence:

    52/180=.289, 153/352 =.435, 36/303=.119 in group A
    51/262 = 0.195, 148/463 = 0.320, 20/151= 0.132 in group B

    so adherence and outcomes are certainly related.

  Almost certainly, the apparent benefit of B relative to A in the stratified analysis is a result of confounding and not true benefit of treatment.

  Note that even had we not been able to identify differences in adherence between groups, we cannot rely on the validity assumptions required for the analysis by adherence category, and therefore, such analyses must be viewed skeptically.

2. (a) Suppose that we have a population composed of four sub-populations A, B, C and D (note that in general we can't observe A, B, C, D so we don't know which groups individuals are in, or even that there are such groups), with probability of death varying according to the following table:

	A	B	C	D
Proportion of population:	0.6	0.2	0.1	0.1
Probability of Death:	0.2	0.4	0.6	0.6

Let $Y_t(u)$ and $Y_c(u)$ be the potential outcomes (1=dead, 0=alive) for subject u, and assume that we have two treatments, t (experimental) and c (control) but that neither has any effect on mortality (i.e., the null hypothesis of no treatment effect is true). Also let $T(u)$ be the treatment *received* by subject u.

i. Calculate $E[Y_c(u)], E[Y_t(u)]$ and $\mu = E[Y_t(u)] - E[Y_c(u)]$.

$$\begin{aligned} E[Y_c(u)] = E[Y_t(u)] &= \sum_{g=A}^{D} \Pr\{Y_c(u) = 1 | u \in g\} \Pr\{u \in g\} \\ &= 0.2 \times 0.6 + 0.4 \times 0.2 + 0.6 \times 0.1 + 0.6 \times 0.1 = 0.32. \end{aligned}$$

Since these expectations are equal, $\mu = 0$.

ii. Suppose that all subjects are assigned c, however, those in groups A and B receive c but subjects in groups C and D do not. Calculate $E[Y_c(u)|T(u) = c]$.

We have the following table

Stratum	$T(u)$	proportion	$E[Y(u)]$
A	c	0.6	0.2
B	c	0.2	0.4
C	–	0.1	0.6
D	–	0.1	0.6

$$\begin{aligned} E[Y_c(u)|T(u) = c] &= \frac{\Pr\{Y_c(u) = 1|u \in A\} \Pr\{u \in A\} + \Pr\{Y_c(u) = 1|u \in B\} \Pr\{u \in B\}}{\Pr\{u \in A\} + \Pr\{u \in B\}} \\ &= \frac{0.2 \times 0.6 + 0.4 \times 0.2}{0.6 + 0.2} \\ &= 0.25 \end{aligned}$$

iii. Suppose that all subjects are assigned t, however, those in group A actually receive t but subjects in groups B, C and D do not. Calculate $E[Y_t(u)|T(u) = t]$.

We have the following table

Stratum	$T(u)$	proportion	$E[Y(u)]$
A	t	0.6	0.2
B	–	0.2	0.4
C	–	0.1	0.6
D	–	0.1	0.6

$$\begin{aligned} E[Y_t(u)|T(u) = t] &= \Pr\{Y_t(u) = 1|u \in A\} \\ &= 0.2 \end{aligned}$$

iv. Compare the results of 2(a)ii and 2(a)iii. What does this tell you about a *Per-Protocol* analysis when the null hypothesis is true?

The *Per-Protocol* analysis shows that there is a difference in mortality rates between subjects assigned t who adhere to their assigned treatment and subjects assigned c who adhere to their assigned treatment. While this is in fact *true*, it is does *not* represent a causal effect of treatment, but rather a bias due to the selection of subjects that receive their respective treatments (there is *no* effect of treatment on probability of death).

Specifically, treatment received is *not* independent of outcomes. Note that for subjects assigned c the probability of having received c given that a subject is dead is

$$\begin{aligned}\Pr\{T(u) = c|\text{Dead}\} &= \frac{\Pr\{\text{Dead}, T(u) = c\}}{\Pr\{\text{Dead}\}} \\ &= \frac{0.6 \times 0.2 + 0.2 \times 0.4}{0.6 \times 0.2 + 0.2 \times 0.4 + 0.1 \times 0.6 + 0.1 \times 0.6} \\ &= \frac{.2}{.32} = 0.625\end{aligned}$$

whereas the probability of having received c given that a subject is alive is

$$\begin{aligned}\Pr\{T(u) = c|\text{Alive}\} &= \frac{\Pr\{\text{Alive}, T(u) = c\}}{\Pr\{\text{Alive}\}} \\ &= \frac{0.6 \times 0.8 + 0.2 \times 0.6}{0.6 \times 0.8 + 0.2 \times 0.6 + 0.1 \times 0.4 + 0.1 \times 0.4} \\ &= \frac{.6}{.68} = 0.8824.\end{aligned}$$

Similarly, for subjects assigned t

$$\Pr\{T(u) = t|\text{Dead}\} = 0.375$$

and

$$\Pr\{T(u) = t|\text{Dead}\} = 0.7059.$$

(b) Using the conditions of part 2(a), except that subjects in group D *always* receive treatment t, and subjects in group B *always* receive treatment c regardless of assigned treatment ("crossovers"). Note that subjects in group C receive *neither* treatment. Suppose further that subjects are randomly *assigned* either t or c with equal probability.

i. Calculate $E[Y_c(u)|T(u) = c]$ and $E[Y_t(u)|T(u) = t]$.

In the following table "proportion" refers to the proportion of the study population in a given stratum (A,B,C,D) *and* an assigned treatment group (one half of the proportions in the table in 1(a))

	Assigned c			Assigned t		
Stratum	$T(u)$	proportion	$E[Y(u)]$	$T(u)$	proportion	$E[Y(u)]$
A	c	0.3	0.2	t	0.3	0.2
B	c	0.1	0.4	c	0.1	0.4
C	–	0.05	0.6	–	0.05	0.6
D	t	0.05	0.6	t	0.05	0.6

Subjects who *receive* c are either those in group A and assigned c, or those in group B.

$$
\begin{aligned}
&E[Y_c(u)|T(u)=c] \\
&= \frac{\Pr\{Y_c(u)=1|u\in A, \text{assigned } c\}\Pr\{u\in A, \text{assigned } c\} + \Pr\{Y_c(u)=1|u\in B\}\Pr\{u\in B\}}{\Pr\{u\in A, \text{assigned } c\} + \Pr\{u\in B\}} \\
&= \frac{0.2\times 0.3 + 0.4\times 0.2}{0.3+0.2} \\
&= 0.28
\end{aligned}
$$

Subjects who *receive* t are either those in group A and assigned t, or those in group D.

$$
\begin{aligned}
&E[Y_t(u)|T(u)=t] \\
&= \frac{\Pr\{Y_t(u)=1|u\in A, \text{assigned } t\}\Pr\{u\in A, \text{assigned } t\} + \Pr\{Y_t(u)=1|u\in D\}\Pr\{u\in D\}}{\Pr\{u\in D, \text{assigned } t\} + \Pr\{u\in D\}} \\
&= \frac{0.2\times 0.3 + 0.6\times 0.1}{0.3+0.1} \\
&= 0.30
\end{aligned}
$$

ii. Compare the results in 2(b)i. What does this tell you about an *As-Treated* analysis when the null hypothesis is true?

Similar to 2(a), the *As-Treated* analysis indicates that there is a difference in mortality rates between subjects *receiving* t and subjects *receiving* c. Again, however, it does *not* due represent a causal effect of treatment, but rather a bias in the selection of subjects receiving their respective treatments. Again, treatment received is *not* independent of outcomes.

(c) Now suppose that treatment t is effective and that if t is received, the probabilities of death in groups A, B C and D are

	A	B	C	D
Probability of Death:	0.15	0.3	0.4	0.5

Also assume that treatment c has no effect (same as no treatment).

i. Calculate $E[Y_t(u)]$ and $\mu = E[Y_t(u)] - E[Y_c(u)]$.

$$E[Y_t(u)] = 0.15\times 0.6 + 0.3\times 0.2 + 0.4\times 0.1 + 0.5\times 0.1 = 0.24$$

Since $E[Y_c(u)]$ is the same as in 2(a), we have

$$\mu = 0.24 - 0.32 = -0.08$$

ii. Now suppose subjects are randomized 1:1 to t or c. As in 2(a)ii, when assigned treatment c, subjects in groups A and B receive it, while those in C and D receive no treatment. As in 2(a)iii, when assigned t, subjects in group A receive t, while those in groups B, C and D receive no treatment.

Let $\tilde{Y}_\tau(u)$ be the potential outcome for subject u if *assigned* treatment $\tilde{T}(u) = \tau$. and calculate $E[\tilde{Y}_t(u)|\tilde{T}(u) = t]$ and $\tilde{\mu} = E[\tilde{Y}_t(u)|\tilde{T}(u) = t] - E[\tilde{Y}_c(u)|\tilde{T}(u) = c]$.

Because adherence makes no difference to subjects assigned c, $E[\tilde{Y}_c(u)|\tilde{T}(u) = c] = 0.32$ as in part 2(a).

For subjects assigned t, those in group A receive t and therefore $\tilde{Y}_t(u) = Y_t(u)$ and while those in groups B, C and D receive no treatment and therefore $\tilde{Y}_t(u) = Y_c(u)$ (because no treatment and c yield the same response).

$$E[\tilde{Y}_c(u)|\tilde{T}(u) = t] = 0.15 \times 0.6 + 0.3 \times 0.2 + 0.6 \times 0.1 + 0.6 \times 0.1 = 0.29.$$

Therefore $\mu = 0.29 - 0.32 = -0.03$, which is smaller in absolute value than $\mu = -.08$ (which imposes full-adherence).

---

iii. Assuming the conditions of 2(c)ii, calculate $E[Y_t(u)|T(u) = t]$ and $\mu_A = E[Y_t(u)|T(u) = t] - E[Y_c(u)|T(u) = c]$.

Similar to 2(a)ii, $E[Y_c(u)|T(u) = c] = 0.25$ and similar to 2(a)iii, $E[Y_t(u)|T(u) = t] = 0.15$. Therefore $\mu_A = 0.15 - 0.25 = -0.10$.

---

iv. Compare the results in 2(c)i, 2(c)ii and 2(c)iii. What does this tell you about the *Intention to Treat Analysis* and the *Per-Protocol* analysis when the null hypothesis is false?

The ITT analysis provides a valid causal analysis of *assignment to treatment*. The *per-protocol* analysis is biased because, as above, adherence is not independent of outcome.

The ITT estimand is $\tilde{\mu} = -.03$, and while smaller in absolute value than the full-adherence estimand, μ, it is impossible to recover μ from observed data with incomplete adherence. Given that subjects in groups B, C and D were never observed having received t, it is impossible to know their outcomes. The *per-protocol* analysis does *not* recover $\widehat{\mu}$ as it ignores subjects in groups C and D entirely, and includes group B subjects only for treatment c. It can't possibly yield a sensible answer.

---

3. 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.

4. 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.

5. The data file "data1.csv" (in csv format, comma delimited, same dataset as used in homework 1) 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))
[snip]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   60.914      1.809  33.677   <2e-16 ***
z              3.886      2.515   1.545    0.128
[snip]
```

      The mean difference is 3.886 with SE 2.515, and t-statistic 1.545.

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

```
> summary(lm(x1-x0~z, data=D))
[snip]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2679     0.9919  -0.270   0.7881
z             3.5745     1.3792   2.592   0.0122 *
[snip]
```

      The mean difference is 3.5745 with SE 1.3892 and t-statistic 2.592.

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

```
> summary(lm(x1~z+x0, data=D))
[snip]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7465     3.6149   3.803 0.000361 ***
z             3.6458     1.2252   2.976 0.004337 **
x0            0.7709     0.0573  13.454  < 2e-16 ***
[snip]
```

      The mean difference is 3.6458 with SE 1.2252 and t-statistic 2.976.

Why do the conclusions differ from these three analysis?

In (c) the coefficient for `x0` is .771, 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))
[snip]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.618      2.063  29.868   <2e-16 ***
z              6.872      2.868   2.396   0.0200 *
[snip]

> summary(lm(x2-x0~z, data=D))
[snip]
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 .
[snip]
> summary(lm(x2~z+x0, data=D))
[snip]
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
[snip]
```

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.