Statistics 641, Fall 2011
Take Home Final Exam
Solutions

1. Suppose that we observe the following follow-up times, y_i ("+" indicates censoring):

   Group 1: 0.38 0.56 1.92 1.04+ 1.73
   Group 2: 1.84 2.03+ 3.78 4.13+ 5.93

   For each failure time, t_j, let R_j be the number of subjects at risk. Assign each subject a score as follows:

$$C_i = \begin{cases} \displaystyle\sum_{t_j \le y_i} \frac{1}{R_j} & \text{if censored} \\ \displaystyle\sum_{t_j \le y_i} \frac{1}{R_j} - 1 & \text{if not censored} \end{cases}$$

   Let $T = \sum_{\text{group 2}} C_i$.

We begin by sorting by follow-up times, and we have the following

y_i	δ_i	z_i	R_j	C_i		
0.38	1	1	10	$1/10 - 1$	=	-0.900
0.56	1	1	9	$1/10 + 1/9 - 1$	=	-0.789
1.04	0	1	–	$1/10 + 1/9$	=	0.211
1.73	1	1	7	$1/10 + 1/9 + 1/7 - 1$	=	-0.646
1.84	1	2	6	$\mathbf{1/10 + 1/9 + 1/7 + 1/6 - 1}$	=	-0.479
1.92	1	1	5	$1/10 + 1/9 + 1/7 + 1/6 + 1/5 - 1$	=	-0.279
2.03	0	2	–	$\mathbf{1/10 + 1/9 + 1/7 + 1/6 + 1/5}$	=	0.721
3.78	1	2	3	$\mathbf{1/10 + 1/9 + 1/7 + 1/6 + 1/5 + 1/3 - 1}$	=	0.054
4.13	0	2	–	$\mathbf{1/10 + 1/9 + 1/7 + 1/6 + 1/5 + 1/3}$	=	1.054
5.93	1	2	1	$\mathbf{1/10 + 1/9 + 1/7 + 1/6 + 1/5 + 1/3 + 1/1 - 1}$	=	1.054

where δ_i is censoring indicator and z_i is treatment. Note that times 1.04, 2.03, 4.13 are censoring times, so they don't contribute a $1/R_j$. Among subjects in group 2, we have

$$T = 5 \times 1/10 + 5 \times 1/9 + 5 \times 1/7 + 5 \times 1/6 + 4 \times 1/5 + 3 \times 1/3 + 1 \times 1/1 - 3 = 2.4032$$

(Aside: note that overall, we have $\sum_i C_i = 10/10 + 9/9 + 7/7 + 6/6 + 5/5 + 3/3 + 1/1 - 7 = 0$ so as defined, the scores have mean zero.)

(a) Using the random allocation rule (randomly assigning 5 subjects to group 1, the remainder to group 2), find the randomization p-value for the observed T.

Sorting the table above by C_i we have the following table where the subject numbers have been assigned according the order of the C_i.

i	z_i	C_i
1	1	−0.900
2	1	−0.789
3	1	−0.646
4	2	−0.479
5	1	−0.279
6	2	0.054
7	1	0.211
8	2	0.721
9	2	1.054
10	2	1.054

We can make a table of allocations of subjects to treatments and compute the corresponding T^*, shifting the allocation of subjects to group 2 successively to the left until we fall below 2.403:

i:	1	2	3	4	5	6	7	8	9	10	T^*	$T^* \geq T$?
	1	1	1	1	1	2	2	2	2	2	3.094	Y
	1	1	1	1	2	1	2	2	2	2	2.760	Y
	1	1	1	1	2	2	1	2	2	2	2.603	Y
	1	1	1	1	2	2	2	1	2	2	2.094	N
	1	1	1	2	1	1	2	2	2	2	2.560	Y
	1	1	1	2	1	2	1	2	2	2	2.403	Y
	1	1	1	2	1	2	2	1	2	2	1.894	N
	1	1	1	2	2	1	1	2	2	2	2.070	N
	1	1	2	1	1	1	2	2	2	2	2.394	N

We find only 5 allocations T^* at least as extreme (including the observed allocation). Thus the (one-sided) randomization p-value is $5/252 = 0.0198$.

(b) Compare treatment groups using the log-rank test.

One could use standard software (say `survdiff` in `R`), but it is instructive to do it by hand. We have 7 distinct failure times.

t_j	d_{j1}	n_{j1}	d_{j2}	n_{j2}	R_j	$E[d_{j2}]$	$\mathrm{Var}(d_{j2})$	
0.38	1	5	0	5	10	$1 \times 5/10$	$1 \times 9 \times 5 \times 5/10^2 \times 9$	$= 0.25000$
0.56	1	4	0	5	9	$1 \times 5/9$	$1 \times 8 \times 4 \times 5/9^2 \times 8$	$= 0.24691$
1.73	1	2	0	5	7	$1 \times 5/7$	$1 \times 6 \times 2 \times 5/7^2 \times 6$	$= 0.20408$
1.84	0	1	1	5	6	$1 \times 5/6$	$1 \times 5 \times 1 \times 5/6^2 \times 5$	$= 0.13889$
1.92	1	1	0	4	5	$1 \times 4/5$	$1 \times 4 \times 1 \times 4/5^2 \times 4$	$= 0.16000$
3.78	0	0	1	3	3	$1 \times 3/3$	$1 \times 3 \times 0 \times 3/3^2 \times 2$	$= 0.0$
5.93	0	0	1	1	1	$1 \times 1/1$	$1 \times 1 \times 0 \times 1/1^2 \times 0$	$= 0.0$
Totals:			3			5.403		0.9998

The log-rank statistic is

$$\frac{(3 - 5.403)^2}{0.9998} = \frac{2.403^2}{0.9998} = 5.776$$

The (2-sided) p-value is 0.0162 (one-sided is 0.0081).

(c) Comment on the results of these two tests.

Except for the sign, the sum of the scores in group 2 is exactly equal to the observed minus expected ($3 - 5.403 = -2.403$) from the log-rank test. The sum of the $E[d_{k2}]$ column in the log-rank table is exactly the first 7 terms of T in part (a): $5/10 + 5/9 + 5/7 + 5/6 + 4/5 + 3/3 + 1/1$. Therefore, (at least when there are no ties) the log-rank test can be constructed as a test based on scores assigned to each subject in a way similar to the rank scores in the Wilcoxon rank-sum test. (For observations $x_1, x_2, \ldots$, with no ties, the rank for observation x_i can be written as $\sum_{xj \leq x_j} 1$, so the log-rank scores are computed in a similar, although more complex way, than the ranks used in the Wilcoxon test.)

The one-sided randomization p-value is also larger (0.0198) than the one based on the chi-square approximation (0.0081). This is because the randomization p-value is based on a restricted sample space, conditional on the observed data, and in particular exactly 7 failures. The chi-square test is based on the population model which is not conditional on the observed data, and in particular, does not condition on the total number of failures.

For the log-rank, we required the assumption that the contributions from each table were uncorrelated in order to add the variances. By assigning scores to subjects rather than considering each failure time separately, we can conduct valid inference without requiring this assumption, essentially validating the use of the log-rank test.

2. Suppose that we have conducted a trial with two interim analyses and a final analysis. Data from each analysis have been compiled into the data file `data-final.csv` which contains the following variables:

   - `z`: Treatment variable (0,1)
   - `time1`: Follow-up time at analysis 1
   - `dead1`: Death indicator at analysis 1 (0=alive, 1=dead)
   - `time2`: Follow-up time at analysis 2
   - `dead2`: Death indicator at analysis 2 (0=alive, 1=dead)
   - `time3`: Follow-up time at analysis 3
   - `dead3`: Death indicator at analysis 3 (0=alive, 1=dead)

   For subjects who were still to be enrolled at the time of each interim analysis, the follow-up time is zero. The primary analysis is the comparison of treatment groups using the log-rank test.

(a) For each interim analysis, compute the log-rank Z-statistic. Use the convention that Z is positive if the difference favors group $z = 1$.

```
> data <- read.csv("data-final.csv")
> survdiff(Surv(time1, dead1)~z, data=data)
Call:
survdiff(formula = Surv(time1, dead1) ~ z, data = data)

       N Observed Expected (O-E)^2/E (O-E)^2/V
z=0 1122      118      125     0.446     0.936
z=1 1122      122      115     0.488     0.936

 Chisq= 0.9  on 1 degrees of freedom, p= 0.333
> survdiff(Surv(time2, dead2)~z, data=data)
Call:
survdiff(formula = Surv(time2, dead2) ~ z, data = data)

       N Observed Expected (O-E)^2/E (O-E)^2/V
z=0 1122      429      457      1.68      3.48
z=1 1122      456      428      1.79      3.48

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0622
> survdiff(Surv(time3, dead3)~z, data=data)
Call:
survdiff(formula = Surv(time3, dead3) ~ z, data = data)

       N Observed Expected (O-E)^2/E (O-E)^2/V
z=0 1122      602      644      2.71      5.62
z=1 1122      646      604      2.89      5.62

 Chisq= 5.6  on 1 degrees of freedom, p= 0.0178
```

For each analysis, the Z score is the square root of the chi-square statistic, with the appropriate sign. Since the problem statement indicates that a positive Z corresponds to benefit for treatment 1, we use the observed and expected values from the above to determine the sign. In each case, the observed number of deaths in group 1 is larger than the expected, indicating that the observed treatment difference favors group 0. Hence the Z statistics are

Analysis time	Z
1	$-\sqrt{0.936} = -0.967$
2	$-\sqrt{3.48} = -1.87$
3	$-\sqrt{5.62} = -2.37$

(b) Calculate the information fraction at each interim analysis.

The information is proportional to the total number of deaths at each analysis time.

Analysis time	Deaths	Information Fraction
1	$118 + 122 = 240$	$240/1248 = 0.192$
2	$429 + 456 = 885$	$885/1248 = 0.709$
3	$602 + 646 = 1248$	1

(c) Find the information fractions based on the Wald test and compare to those in part (b).

To perform the Wald test, we fit the Cox proportional hazards model:

```
> coxph(Surv(time1, dead1)~z, data=data)
Call:
coxph(formula = Surv(time1, dead1) ~ z, data = data)


   coef exp(coef) se(coef)     z    p
z 0.125      1.13    0.129 0.965 0.33

Likelihood ratio test=0.93  on 1 df, p=0.335  n= 2244
> coxph(Surv(time2, dead2)~z, data=data)
Call:
coxph(formula = Surv(time2, dead2) ~ z, data = data)


   coef exp(coef) se(coef)    z     p
z 0.125      1.13   0.0673 1.86 0.062

Likelihood ratio test=3.48  on 1 df, p=0.0622  n= 2244
> coxph(Surv(time3, dead3)~z, data=data)
Call:
coxph(formula = Surv(time3, dead3) ~ z, data = data)


   coef exp(coef) se(coef)    z     p
z 0.134      1.14   0.0567 2.37 0.018

Likelihood ratio test=5.61  on 1 df, p=0.0178  n= 2244
```

Information is the reciprocal of the variance of $\hat{\beta}$ (coefficient for treatment) so we have

Analysis time	$1/\text{var}\hat{\beta}$	Information Fraction
1	$1/0.129^2 = 59.95$	$59.95/311.12 = 0.193$
2	$1/0.0673^2 = 220.73$	$220.73/311.12 = 0.709$
3	$1/0.0567^2 = 311.12$	1

These information fractions are nearly identical to those based on numbers of deaths.

(d) Compute the critical values for stopping boundaries at each of the two interim analysis plus the critical value at the final analysis using the power α-spending function with $\rho = 3/2$.

Assuming 2-sided $\alpha = .05$, we have:

```
> summary(bounds(c(.192, .709, 1), iuse=c(3,3), phi=c(3,3)/2,
+    alpha=c(.025,.025)))

Lan-DeMets bounds for a given spending function

n =  3
Overall alpha:  0.05

Type: Two-Sided Symmetric Bounds
Lower alpha:  0.025
Upper alpha:  0.025
Spending function:  Power Family: alpha * t^phi

Boundaries:
    Time    Lower   Upper   Exit pr.  Diff. pr.
1  0.192  -2.8622  2.8622  0.0042065  0.0042065
2  0.709  -2.2157  2.2157  0.0298497  0.0256431
3  1.000  -2.1166  2.1166  0.0500000  0.0201503
```

(e) What can you conclude?

The Z statistics at each of the interim analyses are within the monitoring boundaries, so the trial would not have stopped early on that basis. At the final analysis, the Z crosses the lower boundary, so we would have demonstrated that treatment 0 is superior to treatment 1.

(f) Calculate the conditional power at each interim analysis given the alternative hypothesis H_A: $\beta = .223$ where β is the log-hazard ratio (corresponding to a hazard ratio of 1.25 favoring group $z = 1$).

Conditional power is given by

$$1 - \Phi\left(\frac{b_K - B(t) - (1-t)\theta}{\sqrt{1-t}}\right)$$

where $K = 3$, $b_K = 2.1166$ (from the table in part (b)), $B(t) = \sqrt{t}Z(t)$ with $t = 0.192$ or 0.709 at analyses 1 and 2 respectively, and $Z(t)$ is the observed Z at analysis time t. We compute as $\theta = \beta\sqrt{I_K} = .223/0.0567 = 3.933$ (0.0567 is the SE from the Cox model at the final analysis).

Plugging into the formula we have at analysis 1:

$$1-\Phi\left(\frac{2.1166-\sqrt{.192}\times(-.967)-(1-0.192)\times 3.933}{\sqrt{1-0.192}}\right)=1-\Phi(-0.709)=0.761$$

At analysis 2:

$$1-\Phi\left(\frac{2.1166-\sqrt{.709}\times(-1.87)-(1-0.709)\times 3.933}{\sqrt{1-0.709}}\right)=1-\Phi(4.72)=0.000001$$

Based on conditional power, there is still a substantial likelihood of success (showing benefit for treatment 1) at the first analysis where conditional power is 76%. However, at analysis 2, the probability of success assuming that the true $\beta = .233$, is miniscule. The trial could be stopped on that basis.

3. Suppose that we are using minimization as a covariate adaptive allocation scheme and we wish to balance with respect to smoking status and sex. Using the notation from class let $G_t = |x^t_{11} - x^t_{12}| + |x^t_{21} - x^t_{22}|$.

   Suppose that the next subject is a non-smoking female and we have:

	Smoker		Sex		
Group	Y	N	M	F	Total
1	15	26	19	22	41
2	16	28	21	23	44

   To which treatment group should the next subject be allocated?

If the next subject is allocated to group 1, the table will be:

	Smoker		Sex		
Group	Y	N	M	F	Total
1	15	27	19	23	42
2	16	28	21	23	44

so $G_1 = |27-28| + |23-23| = 1$.

If the next subject is allocated to group 2, the table will be:

	Smoker		Sex		
Group	Y	N	M	F	Total
1	15	26	19	22	42
2	16	29	21	24	44

so $G_1 = |26-29| + |22-24| = 5$.

Therefore, the next subject should be allocated to group 1.

4. 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 and score test statistic for the difference between treatment groups.

The hazard ratio is

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

Under H_0, the common estimate of the hazard is $\hat{\lambda} = (241 + 219)/(835 + 876) = 0.2688$ The score test is (see page 215 of the textbook)

$$(241 - 835 \times 0.2688)^2 \left(\frac{1}{0.2688 \times 876} + \frac{1}{0.2688 \times 835} \right) = 2.384$$

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

(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

$$\begin{aligned} > 80\%: &\quad \frac{51/262}{52/180} = .674 \\ 50\%\text{-}80\%: &\quad \frac{148/463}{153/352} = .735 \\ \leq 50\%: &\quad \frac{20/151}{36/303} = 1.115 \end{aligned}$$

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 difference between groups.
- Adherence to assigned treatment 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.