Statistics 641, Fall 2014
Homework #7
Solutions

1. Suppose that we have two treatments with a 1:1 permuted block randomization with blocks of size 6 (i.e., within each block of 6 we randomly allocate 3 to each treatment). We enroll 12 subjects and in the two blocks we observe the following summary tables:

Group	D	A	Total
1	3	0	3
2	0	3	3
	3	3	6

Group	D	A	Total
1	2	1	3
2	0	3	3
	2	4	6

(a) Calculate the size of the reference set (all possible allocations of treatments to subjects).

We have two blocks of size 6. Within each block we allocate 3 to group 1 and the remaining 3 to group 2. There are $\binom{6}{3} = 20$ ways of doing this within each block. Thus the reference set has $20 \times 20 = 400$ allocations.

(b) If x_j is the number of deaths in group 1 for block j, find the sample space for $U(0) = \sum_j x_j - E[x_j]$ and corresponding sampling probabilities under the randomization distribution. (*Hint: x_j has a hypergeometric distribution*).

In x_j be the entry in the upper left corner of block j, $j = 1, 2$.

x_1 takes 4 possible values: 0,1,2,3. $E[x_1] = 3 \times 3/6 = 1.5$
x_2 takes 3 possible values: 0,1,2. $E[x_1] = 2 \times 3/6 = 1$

$U(0)$ takes values according to the following table:

$x_2 \backslash x_1$:	0	1	2	3
0	-2.5	-1.5	-0.5	0.5
1	-1.5	-0.5	0.5	1.5
2	-0.5	0.5	1.5	2.5

Each x_j has a hypergeometric distribution, so if m_j is the total number of deaths in block j (3 or 2), then the x_j has probability

$$\Pr\{x_j = x\} = \frac{\binom{3}{x}\binom{3}{m_j - x}}{\binom{6}{m_j}}$$

$$\Pr\{x_1=0\}=\Pr\{x_1=3\}=\frac{\binom{3}{0}\binom{3}{3}}{\binom{6}{3}}=\tfrac{1}{20},$$
$$\Pr\{x_1=1\}=\Pr\{x_1=2\}=\frac{\binom{3}{1}\binom{3}{2}}{\binom{6}{3}}=\tfrac{9}{20}$$

Similarly,

$$\Pr\{x_2=0\}=\Pr\{x_2=2\}=\frac{\binom{3}{0}\binom{3}{2}}{\binom{6}{2}}=\tfrac{1}{5}, \qquad \Pr\{x_1=1\}=\frac{\binom{3}{1}\binom{3}{1}}{\binom{6}{2}}=\tfrac{3}{5}$$

Thus,

$$\begin{aligned}
\Pr\{U(0)=-2.5\} &= \Pr\{x_1=x_2=0\}=\frac{1}{20}\times\frac{1}{5}=\frac{1}{100}\\
\Pr\{U(0)=-1.5\} &= \Pr\{x_1=1,x_2=0\}+\Pr\{x_1=0,x_2=1\}=\frac{9}{20}\times\frac{1}{5}+\frac{1}{20}\times\frac{3}{5}=\frac{12}{100}\\
\Pr\{U(0)=-0.5\} &= \Pr\{x_1=2,x_2=0\}+\Pr\{x_1=1,x_2=1\}+\Pr\{x_1=2,x_2=0\}\\
&= \frac{9}{20}\times\frac{1}{5}+\frac{9}{20}\times\frac{3}{5}+\frac{1}{20}\times\frac{1}{5}=\frac{37}{100}
\end{aligned}$$

By symmetry,

$$\Pr\{U(0)=2.5\}=\frac{1}{100} \qquad \Pr\{U(0)=1.5\}=\frac{12}{100} \qquad \Pr\{U(0)=0.5\}=\frac{37}{100}$$

Hence $U(0)$ takes values $\{-2.5,-1.5,-0.5,0.5,1.5,2.5\}$ with probabilities $\{1/100,12/100,37/100,37/100,12/100,1/100\}$.

(c) Calculate the one-sided randomization p-value for the observed data.

The observed value of $U(0)$ is $(3-1.5)+(2-1)=2.5$. The one-sided randomization p-value is $\Pr\{U(0)\geq 2.5\}=0.01$ from the distribution in part (b).

(Note that the stratified chi-square statistic (Mantel-Haenszel) is $2.5^2/.708=8.824$ *which corresponds to a large-sample p-value of 0.003.)*

2. The dataset `data7.csv` contains 60 observations with the following variables:

- `w`: Categorical baseline variable (levels 1,2,3,4)
- `y`: Response (outcome)
- `z`: Treatment

The dataset is sorted by `w` with the following frequencies:

`w`:	1	2	3	4
	8	12	20	20

Treatment (`z`) was assigned using permuted blocks of size 4 within each stratum defined by levels of `w`. For simplicity, there are no incomplete blocks (the size of each stratum is a multiple of 4). Let β be the difference between treatment groups.

(a) Find the estimates of β unadjusted and adjusted (for `w`) and associated standard errors using ordinary least squares (`lm`). Comment on the differences between these two models.

```
> summary(lm(y~z,data=data7))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  119.600      6.043  19.790   <2e-16 ***
z             16.307      8.547   1.908   0.0614 .
...
Residual standard error: 33.1 on 58 degrees of freedom
...
> summary(lm(y~z+w,data=data7))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   84.784      9.811   8.642 7.94e-12 ***
z             16.307      6.730   2.423  0.01871 *
w2            14.096     11.897   1.185  0.24119
w3            37.523     10.904   3.441  0.00111 **
w4            58.468     10.904   5.362 1.68e-06 ***
...
Residual standard error: 26.07 on 55 degrees of freedom
...
```

The unadjusted and adjusted estimates of β are identical. This is a consequence of the permuted block randomization that guarantees that the distribution of `w` is identical for the two treatment groups.

The standard errors for $\widehat{\beta}$ are different—the SE of the adjusted $\widehat{\beta}$ is smaller—because the adjusted model accounts for the variability explained by `w`, whereas the unadjusted model does not. The additional variability accounted for by the adjusted model is reflected in the smaller residual standard error (26.07 versus 33.1).

(b) Find the size of the reference set for

i. complete randomization (independent unbiased coin flips),

ii. random allocation rule (30 subjects assigned `z=0`, the rest `z=1`, ignoring `w`) and

iii. permuted block randomization (block size 4) stratified by `w`.

i. $2^{60} = 1.153 \times 10^{18}$

ii. $\binom{60}{30} = 1.18 \times 10^{17}$

iii. $6^{15} = 4.70 \times 10^{11}$ (15 blocks, each with 6 possible allocations.)

Each of these is too large enumerate so the exact randomization distribution cannot be easily obtained.

(c) Since the reference sets are too large to enumerate, by randomly sampling from each of the randomization distributions in the part (b), (10,000 should only take a few seconds), calculate the

- variance and standard error of $\widehat{\beta}$.
- two-sided randomization p-value.

Compare these to the models in part (a).

Using 10,000 replications, I get the following (because this is based on random samples, your answers will vary a little from these.)

The randomization distribution for the unadjusted $\widehat{\beta}$ yields:

```
> betaC <- replicate(10000, {data7$z <- sample(0:1,60, repl=T)
+   lm(y~z,data=data7)$coef[2]})
> var(betaC)
[1] 79.36542
> sd(betaC)
[1] 8.908727
> betaR <- replicate(10000, {data7$z <- sample(rep(0:1,30), 60, repl=F)
+   lm(y~z,data=data7)$coef[2]})
> var(betaR)
[1] 75.98217
> sd(betaR)
[1] 8.716775
> betaB <- replicate(10000, {data7$z <- c(replicate(15,sample(rep(1:2,2),4)))
+   lm(y~z,data=data7)$coef[2]})
> var(betaB)
[1] 50.01049
> sd(betaB)
[1] 7.071809
```

The SEs for complete randomization and the random allocation rule are comparable to one another and to the variance from the unadjusted model in part (a).

The SE for permuted block randomization is smaller than the other two and comparable to the SE from the adjusted model in part (a).

The randomization distribution for the adjusted $\widehat{\beta}$ yields:

```
> betaC <- replicate(10000, {data7$z <- sample(0:1,60, repl=T)
```

```
+   lm(y~z+w,data=data7)$coef[2]})
> var(betaC)
[1] 53.32874
> sd(betaC)
[1] 7.302653
>  betaR <- replicate(10000, {data7$z <- sample(rep(0:1,30), 60, repl=F)
+   lm(y~z+w,data=data7)$coef[2]})
>  var(betaR)
[1] 51.99803
>  sd(betaR)
[1] 7.210966
```

The SEs for the adjusted $\widehat{\beta}$ are comparable to one another and the SE for permuted block randomization and the adjusted model in part (a).

In particular, if we use the adjusted model, it makes little difference which randomization scheme we had used.

NOTE: because I did not specify which $\widehat{\beta}$ to calculate the randomization for, I give credit for either one.

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 $G_t = |x_{11}^t - x_{12}^t| + |x_{21}^t - x_{22}^t|$.

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.