Statistics 641, Fall 2013
Homework #2
Solutions

1. The data file "data2.csv" (in csv format, comma delimited) contains columns

   - `z`: treatment variable (0,1).
   - `w`: categorical baseline covariate with four levels, 1–4.
   - `x`: value of response variable

   (note that you can actually read the data directly into `R` from the web via
   `data<-read.csv(url("http://www.biostat.wisc.edu/~cook/641.hw/data2.csv"))`)

   Assume that the responses, X, are normally distributed.

   (a) Let β be the true treatment difference and find the overall (ignoring W) estimate of β, and the corresponding t-test of $H_0 \colon \beta = 0$.

```
> data2 <- read.csv(url("http://www.biostat.wisc.edu/~cook/641.hw/data2.csv"))
> summary(lm(x~z,data=data2))
   .
   .
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7826     0.2960  22.914   <2e-16 ***
z             0.5984     0.4186   1.429    0.156
   .
   .
```

   The difference between groups does not reach statistical significance.

   (b) Find the estimate of β using a linear model adjusted for W as a categorical variable. *(Note: in the dataset `w` is a numeric variable, but you can convert to a categorical variable using `as.factor()`.)*

```
> summary(lm(x~z+as.factor(w),data=data2))
   .
   .
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.4793     0.2372  23.096  < 2e-16 ***
z               0.7831     0.2109   3.713 0.000346 ***
as.factor(w)2  -0.3908     0.2983  -1.310 0.193337
as.factor(w)3   0.7784     0.3023   2.575 0.011577 *
as.factor(w)4   4.0343     0.2839  14.210  < 2e-16 ***
   .
   .
```

```
Residual standard error: 1.048 on 95 degrees of freedom
```

Stratified by W, the difference between groups does reach statistical significance (unless you choose a very small type I error rate).

(c) Calculate the score functions $U_\beta(0)$ and Fisher information, $\mathcal{I}_\beta(0)$, within each stratum defined by W. You may use the estimate of σ^2 from the model you fit in part (b).

Within each stratum, the score functions have form (see page 9 of lecture 4 from Sept 9)

$$U_\beta(0) = \frac{n_0 n_1}{(n_0 + n_1)\sigma^2} (\bar{x_1} - \bar{x_0})$$

which depend on means and counts by group/stratum. Use `tapply` and `table` to get these:

```
> Means <- tapply(data2$x, data2[c("z","w")], mean)
> Means
   w
z          1        2        3         4
  0 5.737273 5.020833 6.186667  9.435333
  1 6.085000 5.945455 7.126000 10.386923
> Ns <- table(data2$z, data2$w)
> Ns


     1  2  3  4
  0 11 12 12 15
  1 16 11 10 13
> Means[2,]-Means[1,]
        1         2         3         4
0.3477273 0.9246212 0.9393333 0.9515897   ## differences in means
```

Using the Residual standard error from the model above (1.048), we have

```
> I <- Ns[2,]*Ns[1,]/(Ns[2,]+Ns[1,])/1.048^2
> U <- I*(Means[2,]-Means[1,])
> U
       1        2        3        4
2.063788 4.831560 4.665044 6.033979
> I
       1        2        3        4
5.935077 5.225448 4.966335 6.340945
```

(d) Combine the score functions from (c) into a test of $H_0\colon \beta = 0$. Compare with part (b).

The t-statistic is

```
> sum(U)/sqrt(sum(I))
[1] 3.711876
```

This is the same as the 3.713 (up to round-off error in σ) as in part (b). Note that the F-statistic is

```
> sum(U)^2/sum(I)
[1] 13.77802
```

2. Suppose in a mortality trial of treatment A versus treatment B we observe the following table by baseline strata:

	Stratum 1		Stratum 2		Stratum 3		Total	
	A	B	A	B	A	B	A	B
Alive	18	21	8	11	1	4	27	36
Dead	5	1	5	1	12	6	22	8
Total	23	22	13	12	13	10	49	44

(a) Calculate the score test (Pearson chi-square) statistic (uncorrected) for the difference in overall mortality, ignoring strata.

The Pearson chi-square statistic is

$$\frac{(22 - 30 \times 49/93)^2}{63 \times 30 \times 49 \times 44/93^3} = \frac{6.194^2}{5.066} = 7.572$$

Note that if `All` is a table containing the overall cell counts, we can use the `chisq.test` function in `R` to do the test:

```
> All
     [,1] [,2]
[1,]   27   36
[2,]   22    8
> chisq.test(All, correct=F)

Pearson's Chi-squared test

data:  All
X-squared = 7.5721, df = 1, p-value = 0.005928
```

(b) Calculate the stratified score test for the adjusted difference in overall mortality.

Letting x_k represent the number dead in group A and stratum k, we have

Stratum	$U(0)$	$I(0)$
A	$5 - 6 \times 23/45 = 1.933$	$23 \times 22 \times 39 \times 6/45^3 = 1.299$
B	$5 - 19 \times 13/25 = 1.880$	$13 \times 12 \times 19 \times 6/25^3 = 1.138$
C	$12 - 18 \times 13/23 = 1.827$	$13 \times 10 \times 5 \times 18/23^3 = 0.962$
Total	5.639	3.399

The stratified score chi-square statistic is

$$\frac{5.639^2}{3.399} = 9.356$$

Note that the numerator is slightly smaller than the unadjusted (5.639 versus 6.194), however, the denominator is quite a bit smaller (3.399 versus 5.066), resulting in a larger test statistic. Stratification increases efficiency by reducing variance.

If `Grouped` is a $2 \times 2 \times 3$ array containing the stratified cell counts, we can use the function `mantelhaen.test` to perform a similar, but not identical, test. This function also gives an estimate of the common odds-ratio.

```
> Grouped
, , 1
     [,1] [,2]
[1,]   18   21
[2,]    5    1
, , 2
     [,1] [,2]
[1,]    8   11
[2,]    5    1
, , 3
     [,1] [,2]
[1,]    1    4
[2,]   12    6
> mantelhaen.test(Grouped, correct=F)
Mantel-Haenszel chi-squared test without continuity correction
data:  Grouped
Mantel-Haenszel X-squared = 9.0354, df = 1, p-value = 0.002648
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.03864222 0.56807600
sample estimates:
common odds ratio
        0.1481611
```

The difference between theses tests (9.356 versus 9.035) is that `mantelhaen.test` uses the conditional *hypergeometric variance* rather then unconditional variance. (See page 6 of lecture 3)

3. Suppose that we observe two random variables X_1, X_2, each with a Poisson distribution and mean parameters λ_1 and λ_2 respectively. For $i = 1, 2$, let

$$\theta_i = \log \lambda_i = \alpha + \beta z_i$$

where $z_1 = 0$ and $z_2 = 1$. Suppose that we observe $X_1 = x_1$ and $X_2 = x_2$ ($x_i > 0$).

(a) Compute the maximum likelihood estimate of α under the null hypothesis $H_0\colon \beta = 0$.

The density function for a Poisson random variable is $f(x;\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, so with two RVs, the joint log-likelihood is

$$\ell(\alpha,\beta) = x_1\alpha - e^{\alpha} + x_2(\alpha+\beta) - e^{\alpha+\beta} + \text{stuff}$$

So

$$\begin{aligned} U_\alpha(\alpha,\beta) &= x_1 - e^{\alpha} + x_2 - e^{\alpha+\beta} \\ U_\beta(\alpha,\beta) &= x_2 - e^{\alpha+\beta} \end{aligned}$$

Under $H_0\colon \beta = 0$, we solve

$$\begin{aligned} 0 &= U_\alpha(\alpha,0) \\ &= x_1 - e^{\alpha} + x_2 - e^{\alpha} \end{aligned}$$

so $\hat\alpha_0 = \log \dfrac{x_1 + x_2}{2}$.

(b) Derive the score test for $H_0\colon \beta = 0$.

$$U_\beta(\hat\alpha_0, 0) = x_2 - e^{\hat\alpha_0} = x_2 - \frac{x_1+x_2}{2} = \frac{x_2 - x_1}{2}$$

The Fisher information is

$$I_\beta(\hat\alpha_0,0) = \mathrm{Var}\big(U_\beta(\hat\alpha_0,0)\big) = \frac{\mathrm{Var}\big(x_1\big) + \mathrm{Var}\big(x_2\big)}{4}$$

Under H_0, $\mathrm{Var}\big(x_1\big) = \mathrm{Var}\big(x_2\big) = \lambda$, where λ is the common Poisson rate, e^{α}. Plugging in $\hat\alpha_0$,

$$I_\beta(\hat\alpha_0,0) = \frac{x_1+x_2}{4}$$

Hence the score test is

$$\frac{(x_2-x_1)^2}{x_2+x_1}.$$

(c) If $p = \lambda_2/(\lambda_1 + \lambda_2)$, the conditional distribution for X_2 given the sum $m = X_1 + X_2$ is binomial with size m and probability p.

$$f(x_2; p) = \binom{m}{x_2} p^{x_2}(1-p)^{m-x_2}.$$

Derive the score test for $H_0\colon \beta = 0$ using the conditional distribution.

Under $H_0\colon \beta = 0$, $p_0 = 1/2$. Using the results derived in class, the score test is

$$\frac{(x_2 - mp_0)^2}{mp_0(1-p_0)} = \frac{(x_2 - (x_1 + x_2)/2)^2}{(x_1 + x_2)/4} = \frac{(x_2 - x_1)^2}{x_2 + x_1}$$

exactly the same as the unconditional test.

(d) Suppose that in a two-armed randomized trial with equal numbers of subjects per arm we observe 65 and 49 deaths in groups A and B respectively. Assuming that the numbers of deaths follow a Poisson distribution, perform the score test for the null hypothesis that the death rates are the same in the two arms.

Using the above, the score chi-square statistic is

$$\frac{(65 - 49)^2}{65 + 49} = \frac{16^2}{114} = 2.25$$

At $\alpha = 0.05$, the critical value is 3.84, so we would not reject H_0

(e) Suppose further that the population in the previous part has two strata, one high risk and one low risk. If the numbers of deaths are 51 and 14 in the high and low risk strata respectively in arm A and 37 and 12 in the high and low risk strata respectively in arm B, perform the stratified score test for the null hypothesis that the death rates are the same in the two arms.

In the high risk stratum, we have

$$U_\beta(\hat\alpha, 0) = (51 - 37)/2 = 14/2, \qquad I_\beta(\hat\alpha, 0) = (51 + 37)/4 = 88/4.$$

and in the high risk stratum, we have

$$U_\beta(\hat\alpha, 0) = (14 - 12)/2 = 2/2, \qquad I_\beta(\hat\alpha, 0) = (14 + 12)/4 = 26/4.$$

The stratified test statistic is

$$\frac{(14 + 2)^2/4}{(88 + 26)/4} = \frac{16^2}{114} = 2.25$$

exactly the same as the unstratified test. For Poisson data, stratification does not change the test.

(Note that this procedure gives a "quick and dirty" way of assessing the statistical significance of event rates between two equal sized groups.)