Homework 04
Joseph Blubaugh
jblubau1@tamu.edu
STAT 659-700

2.29

At the $\alpha = .05$ level we conclude that the results of using prednisolone are significantly better than the control.

	Normalization	No.Effect
Prednisolone	7	8
Control	0	15

```
    Fisher's Exact Test for Count Data

data:  dta
p-value = 0.006322
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.978391      Inf
sample estimates:
odds ratio
       Inf
```

2.30

At the $\alpha = .05$ level we conclude that the odds ratio between radiation therapy and surgery are non significantly different.

	Cancer.Controlled	Cancer.Not.Controleed
Surgery	21	2
Radiation Therapy	15	3

```
    Fisher's Exact Test for Count Data

data:  dta
p-value = 0.3808
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.2864828       Inf
sample estimates:
odds ratio
  2.061731
```

2.31

a) The two sided pvalue is not significant at the $\alpha = .05$ level.

```
library(epitools)
ormidp.test(21, 2, 15, 3)
```

```
  one.sided two.sided
1 0.2430911 0.4861822
```

b) The midpoint pvalue lowers the pvalue but not significantly enough to make a difference. The advantages the midpvalue has over the regular one is that is helps to correct for discrete distributions where the exact pvalue cannot be calculated and it is typically used where there are few observations.

2.33

a)

```
3 Way Table with adder

$white
             Death.Penalty No.Death.Penalty
killed white          19.5            132.5
killed black           0.5              9.5

$black
             Death.Penalty No.Death.Penalty
killed white          11.5             52.5
killed black           6.5             97.5
```

b) At the $\alpha = .05$ level the odds ratio between death penalty and victims race are not significant for white defendents, but the odds ratio is significant for black defendents.

```
Partial Table for White Defendents


    Fisher's Exact Test for Count Data

data:  dta$white
p-value = 0.6135
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3108227       Inf
sample estimates:
odds ratio
       Inf
```

```
Partial Table for Black Defendents


    Fisher's Exact Test for Count Data

data:  dta$black
p-value = 0.0107
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.214554 12.874620
sample estimates:
odds ratio
  3.737623
```

c) Simpsons Paradox does appear to be in play here because the marginal odds ratio's differ from the results when looking at the odds ratis separately by defendent.

```
Marginal Table holding defendent constant


             Death.Penalty No.Death.Penalty
killed white            30              184
killed black             6              106



    Fisher's Exact Test for Count Data

data:  dta
p-value = 0.02423
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.127473 8.719466
sample estimates:
odds ratio
   2.87246
```

2.35

This occurence would be known as simpsons paradox where the sub levels contradict the overall level. This could happen because the sample sizes for the different sub levels are lopsided and so the proportions might be larger for each level in south carolina but the overall death rate may be larger for maine.

2.37

a) Conditional on gender non-white males are 5 times more likely to be victims of murder than white males and non-white females are 3 times more likely to be victims of murder than white females

	White	NonWhite
Male	0.0049	0.0263
Female	0.0023	0.0072

```
(Male = ((.0263)/(1 - .0263)) / ((.0049)/(1 - .0049)))
```

```
[1] 5.485311
```

```
(Female = ((.0072)/(1 - .0072)) / ((.0023)/(1 - .0023)))
```

```
[1] 3.145885
```

b)

```
marginal.white = mean(c(.0049, .0023))
marginal.non.white = mean(c(.0263, .0072))

(marginal.non.white / (1 - marginal.non.white)) / (marginal.white / (1 - marginal.white))
```

```
[1] 4.715004
```

3.2

a) `.0304 - .0003 = .0301`

b) Yes this is clearly an outlier, not only judging by the graphic, but also that the actual value is almost 4 times that of the predicted value.
$\hat{\pi} = -.0003 + .0304(.0079) = .00205$
$\frac{\pi_i}{\hat{\pi}} = .0774/.00205 = 3.85$
$\pi_i - \hat{\pi}_i = .0774 - .00205 = .07535$

3.3

a) The prediction equation is: $.00255 + .00109x$. Intercept is the estimated probability of a child having a birth defect if the mother consumes no alcohol. The Alcohol parameter .001 is the expected increase in birth defect probability for a one unit increase in alcohol consumption.

b) For 0: $.00255 + .00109 * 0 = .0026$, For 7: $.00255 + .00109 * 7 = .0102$ The relative risk between 0 and 7 drinks is $.0102/.0026 = 3.9$

3.4

a) The prediction equation is: $.0026 + .0007x$ The intercept is essentially the same but the estimated increase in probability for an additional unit of alcohol is now .0007 as opposed to .0102 so the algorithm was sensitive to the small sample size. The relative risk is now $.0075/.0026 = 2.9$

```
proc genmod data=defects;
    model defects/count = drinks / dist=bin link=identity;
```

Table 4: Output from SAS

Parameter	DF	Estimate	Std.Error	Wald.lwr	Wald.upr	Wald.Chi.sq	Pvalue
Intercept	1	.0026	.0003	.002	.0033	58.14	<.001
Drinks	1	.0007	.0007	-.0008	.0021	.8	.3699

b) The prediction equation is: $.0027 + .0003x$. The intercept is essentially the same, but the increase in probability for additional unit of alcohol is now .0003. The relative risk between 0 and 4 alcoholic drinks is now $.0039/.0027 = 1.4$

Table 5: Output from SAS

Parameter	DF	Estimate	Std.Error	Wald.lwr	Wald.upr	Wald.Chi.sq	Pvalue
Intercept	1	.0027	.0004	.0019	.0034	58.58	<.001
Drinks	1	.0003	.0005	-.0006	.0013	.43	.5107

c) The prediction equation is now: $-5.92 + .1759x$ but the output is interpreted as odds. Intercept probability of birth defects when consuming no alcohol: $exp(-5.92)/(1 + exp(-5.92)) = .0027$ The increased probability of birth defects from an additional unit of alcohol: $exp(.1759)/(1 + exp(.1759)) = .543$ which is much higher than the other models.

```
proc genmod data=defects;
    model defects/count = drinks / dist=bin link=logit;
```

Table 6: Output from SAS

Parameter	DF	Estimate	Std.Error	Wald.lwr	Wald.upr	Wald.Chi.sq	Pvalue
Intercept	1	-5.92	.1187	-6.153	-5.687	2487	<.001
Drinks	1	.1759	.1705	-.158	.51	1.06	.302

3.5

The linear relationship is preserved even when changing the values of x. The algorithm is measuring the effects on y based on changes in x so it really doesn't matter what the x values are as long as the relationship between changes in x and y are preserved.

Table 7: Partial SAS Output (0,2,4,6)

Parameter	DF	Estimate
Intercept	1	.0176
Drinks	1	.0181

Table 8: Partial SAS Output (0,1,2,3)

Parameter	DF	Estimate
Intercept	1	.0176
Drinks	1	.0362

Table 9: Partial SAS Output (1,2,3,4)

Parameter	DF	Estimate
Intercept	1	.0176
Drinks	1	.0362

Additional A)

Department	Male.Yes	Male.No	Female.Yes	Female.No
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317

a)

```
dta$OR = with(dta, (Male.Yes * Female.No) / (Male.No * Female.Yes))
dta$se = with(dta, (sqrt(1/Male.Yes + 1/Female.No + 1/Male.No + 1/Female.Yes)))
dta$Conf.lwr = with(dta, OR - 1.96 * se)
dta$Conf.Upr = with(dta, OR + 1.96 * se)
```

b) The marginal odds ratio shows that there is a significant overall difference between the entrace of male and female applicants, however 2 of the 6 sublevels have confidence intervals that cross 1 and so are not significantly different at the the $\alpha = .05$ level. Since most of the sub levels odds ratios agree with the marginal odds ratio, simpsons paradox is not really present.

```
marginal = colSums(dta[, 2:5])
OR = (marginal[1] * marginal[4]) / (marginal[2] * marginal[3])
se = sqrt(1/marginal[1] + 1/marginal[4] + 1/marginal[2] + 1/marginal[3])

## Confidence Interval for the Marginal OR
OR + c(-1, 1) * 1.96 * se
```

```
[1] 1.71585 1.96631
```

c)

```
library(DescTools)
library(lawstat)

dta = xtabs(freq ~ .,
            cbind(expand.grid(Gender = c("Male", "Female"),
                              Entrace = c("Yes", "No"),
                              Department = c("1", "2", "3", "4", "5", "6")),
                  freq =  c(512, 89, 313, 19,
                            353, 17, 207, 8,
                            120, 202, 205, 391,
                            138, 131, 279, 244,
                            53, 94, 138, 299,
                              22, 24, 351, 317)))

## Ho: OR = 1, Ha: OR > 1
BreslowDayTest(dta, OR = 1)
```

```

    Breslow-Day test on Homogeneity of Odds Ratios

data:  dta
X-squared = 19.938, df = 5, p-value = 0.001283
```

```
## Ho: OR_1 = OR_2 = OR_3 = OR_4 = OR_5 = OR_6, Ha: At least one set of OR are not equal
cmh.test(dta)
```

```

    Cochran-Mantel-Haenszel Chi-square Test

data:  dta
CMH statistic = 1.52460, df = 1.00000, p-value = 0.21692, MH
Estimate = 0.90470, Pooled Odd Ratio = 1.84110, Odd Ratio of level
1 = 0.34921, Odd Ratio of level 2 = 0.80250, Odd Ratio of level 3
= 1.13310, Odd Ratio of level 4 = 0.92128, Odd Ratio of level 5 =
1.22160, Odd Ratio of level 6 = 0.82787
```

d) Since the common odds ratio includes 1 and the pvalue = .72 we can conclude that a single common odds ratio can be used instead of using the odds ratio for each department.

```

    Cochran-Mantel-Haenszel Chi-square Test

data:  dta
CMH statistic = 0.12498, df = 1.00000, p-value = 0.72369, MH
Estimate = 1.03100, Pooled Odd Ratio = 1.56390, Odd Ratio of level
1 = 0.80250, Odd Ratio of level 2 = 1.13310, Odd Ratio of level 3
= 0.92128, Odd Ratio of level 4 = 1.22160, Odd Ratio of level 5 =
0.82787
```

```
## 90% Confidence for the Common Odds Ratio
1.031 + c(-1, 1) * 1.644 * .0723
```

```
[1] 0.9121388 1.1498612
```

Additional B)

District	Blacks.Yes	Blacks.No	Whites.Yes	Whites.No
NC	24	9	47	12
NE	10	3	45	8
NW	5	4	57	9
SE	16	7	54	10
SW	7	4	59	12

a) At the $\alpha = .1$ level, all districts show odds ratios that cross 1.

```
## Conditional on Distric
dta$OR = with(dta, (Blacks.Yes * Whites.No)/(Blacks.No * Whites.Yes))
dta$se = with(dta, sqrt(1/Blacks.Yes + 1/Whites.No + 1/Blacks.No + 1/Whites.Yes))
dta$Conf.Lwr = with(dta, OR - 1.644 * se)
dta$Conf.Upr = with(dta, OR + 1.644 * se)

dta
```

```
  District Blacks.Yes Blacks.No Whites.Yes Whites.No        OR        se
1       NC         24         9         47        12 0.6808511 0.5073339
2       NE         10         3         45         8 0.5925926 0.7619420
3       NW          5         4         57         9 0.1973684 0.7606937
4       SE         16         7         54        10 0.4232804 0.5691007
5       SW          7         4         59        12 0.3559322 0.7022390
    Conf.Lwr Conf.Upr
1 -0.1532059 1.514908
2 -0.6600400 1.845225
3 -1.0532121 1.447949
4 -0.5123212 1.358882
5 -0.7985487 1.510413
```

b) Simpsons paradox is in play in this case because the marginal odds ratio shows that there is a clear difference between merit pay based on race, but when reviewed at the district level the 90% confidence interval shows that there is no difference between merit pay and race.

```
## Marginal OR
marginal = colSums(dta[, 2:5])
OR = (marginal[1] * marginal[3]) / (marginal[2] * marginal[4])

se = sqrt(1/marginal[1] + 1/marginal[2] + 1/marginal[3] + 1/marginal[4])

## 90% Cofidence
OR + c(-1, 1) * 1.644 * se
```

```
[1] 11.34168 12.25164
```

c) We conclude from the Breslow Day test that the odds ratios differ between race and merit because the pvalue is small. We also conclude from the CMH test that the odds ratios are not equal and so should be viewed separately.

```
    Breslow-Day test on Homogeneity of Odds Ratios

data:  dta
X-squared = 10.96, df = 4, p-value = 0.02702


    Cochran-Mantel-Haenszel Chi-square Test

data:  dta
CMH statistic = 7.8149000, df = 1.0000000, p-value = 0.0051817, MH
Estimate = 0.4617300, Pooled Odd Ratio = 0.4469900, Odd Ratio of
level 1 = 0.6808500, Odd Ratio of level 2 = 0.5925900, Odd Ratio
of level 3 = 0.1973700, Odd Ratio of level 4 = 0.4232800, Odd
Ratio of level 5 = 0.3559300
```

Additional C)

```
dta = data.frame(
  Class = c(1, 2),
  Male.Yes = c(100, 101),
  Male.No = c(99, 98),
  Female.Yes = c(99, 100),
  Female.No = c(98, 97)
)

## Odds Ratio conditional on Class
OR.class = with(dta, (Male.Yes * Female.No) / (Female.Yes * Male.No))

## Odds Ratio conditional on Gender
OR.gender = c( (100*98)/(101*99) , (99*97)/(100*98) )

OR.class; OR.gender
```

```
[1] 0.9998980 0.9996939


[1] 0.980098 0.979898
```