

STAT 220: Basic Statistics for Quantitative Students

Spring 2006

Assignment due Mar. 31

6.12 a. Relative risk = Risk in drug group / Risk in placebo group = $\frac{10/100}{5/100} = 2$.

b. $\frac{15/100}{5/100} = 3$.

c. $\frac{6/100}{4/100} = 1.5$.

6.13 a. Odds ratio = $\frac{10/90}{5/95} = 2.11$.

b. $\frac{15/85}{5/95} = 3.35$.

c. $\frac{6/94}{4/96} = 1.53$.

6.27 a. In each instance calculate $Rate = Tax / Income$. Answers are:

Adjusted Gross Income	1974	1978
Under \$5,000	.054	.035
\$5,000 to \$9,999	.093	.072
\$10,000 to \$14,999	.111	.100
\$15,000 to \$99,999	.160	.159
\$100,000 or more	.384	.383

b. For 1974, the rate is .141. For 1978, the rate is .152.

c. Decreased.

d. Increased.

e. Overall the tax rate was higher in 1978, but the rate was higher within each income bracket in 1974. This occurred because a relatively larger portion of total income was in the higher income brackets (with higher tax rates) in 1978 than in 1974.

B. There are 75 people on the list, which may be downloaded into R using this line:

```
>pres=read.csv("http://www.stat.psu.edu/~dhunter/220/assignments/pres
s.csv")
> names(pres)
[1] "Presidents" "V.or.P" "Month" "Day" "Year"
[6] "Day.of.Week"
```

Part 1: The observed and expected counts may be produced and then viewed as follows. Note that the “table” command puts the months in alphabetical order, so the correct vector containing the number of days per month is 30,31,31,28,31,31,30,31,31,30,31,30:

```
> obs=table(pres[, "Month"])
> exp=75*c(30, 31, 31, 28, 31, 31, 30, 31, 31, 30, 31, 30)/365
> rbind(obs, exp)
      April  August December February  January    July    June
March
obs 5.000000 6.000000 3.000000 7.000000 8.000000 8.000000 4.000000
8.000000
exp 6.164384 6.369863 6.369863 5.753425 6.369863 6.369863 6.164384
6.369863
      May November  October September
obs 5.000000 8.000000 11.000000 2.000000
exp 6.369863 6.164384 6.369863 6.164384
```

The chi-square test now operates as usual. There are $12-1=11$ degrees of freedom:

```
> sum((obs-exp)^2/exp) # This is the chi-square statistic
[1] 11.32577
> 1-pchisq(11.32577, 11)
[1] 0.4163887
```

The p-value is not close to zero, which means that we have no evidence that the number of births per month is not proportional to the number of days per month.

Part 2: Proceed in much the same way as part 1. Here, we don’t have to worry about the weekdays being in alphabetical order because the expected number of births under the null hypothesis is the same for each day:

```
> obs=table(pres[, "Day.of.Week"])
> exp=75*rep(1/7, 7)
> rbind(obs, exp)
      Friday  Monday Saturday  Sunday Thursday  Tuesday Wednesday
obs 8.00000 11.00000 11.00000 13.00000 9.00000 15.00000 8.00000
exp 10.71429 10.71429 10.71429 10.71429 10.71429 10.71429 10.71429
```

Now for the chi-square test:

```
> sum((obs-exp)^2/exp) # This is the chi-square statistic
[1] 3.866667
> 1-pchisq(3.866667, 6)
[1] 0.6947143
```

Again, this is a large p-value, which means no evidence of an unusual pattern.

(a) Here, sex is the explanatory variable because we are interested in how sex affects voting behavior.

(b) Below are the row percents:

```
>source("http://www.stat.psu.edu/~dhunter/220/files/rfiles/tab
les.r")
> obs = table(s[, "Sex"], s[, "Vote2004"])
> rowpercents(obs)

      I was too young to be eligible   No   Yes
Female                                21.9 19.4 58.7
Male                                  8.2 23.9 67.9
```

The most striking thing in the table is that there are many more females than males who were too young.

(c) The null hypothesis says that males and females have the same voting behavior as classified by the three-level variable above. The alternative is the opposite of the null (it is not specific about how sex affects voting behavior).

(d) Here is the chi-square test:

```
> chisq.test(s[, "Sex"], s[, "Vote2004"])

Pearson's Chi-squared test

data:  s[, "Sex"] and s[, "Vote2004"]
X-squared = 14.8064, df = 2, p-value = 0.0006093
```

Alternatively, we could accomplish this “by hand”:

```
> obs

      I was too young to be eligible   No   Yes
Female                                54  48 145
Male                                  15  44 125
> rowtotals = apply(obs,1,sum)
> coltotals = apply(obs,2,sum)
> grandtotal=sum(obs)
> exp=outer(rowtotals,coltotals)/grandtotal
> exp

      I was too young to be eligible      No      Yes
Female                                39.54292 52.7239 154.7332
Male                                  29.45708 39.2761 115.2668
> sum((obs-exp)^2/exp)
[1] 14.80641
> 1-pchisq(14.80641,2)
[1] 0.0006092968
```

We conclude that there is strong evidence ($p=0.0006$) that the patterns for males and females are different for the population represented here. An examination of the row percents in part (a) suggests that this might be due to the much larger percentage of females who were too young, which suggests

that this sample is not actually representative of any interesting real population: In other words, it may simply be the case that for some reason, the women who enrolled in STAT 100 were younger on average than the men, and unless we are interested only in the population of STAT 100 students (which is unlikely), it appears that this particular sample may have introduced some bias with respect to the question of interest (voting).

There is also the possibility that either the men or women are not answering truthfully, though this possibility seems remote due to the sampling method used here in which participants' responses are completely anonymous.