Discussion 5

I. Review

1. Contigent Table

   (a) Difference in proportions
   A 95% confidence interval for $p_1 - p_2$ is

   $$\hat{p}_1 - \hat{p}_2 - 1.96SE(\hat{p}_1 - \hat{p}_2) < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + 1.96SE(\hat{p}_1 - \hat{p}_2)$$

   where $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

   (b) Odds Ratio

   i. odds: $\frac{p}{1-p}$, odds ratio(OR): $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

   ii. If $OR = 1$, the two group has the same odds, $p_1 = p_2$; if $OR > 1$, then the first group has higher odds than the second, so $p_1 > p_2$; if OR¡1, then the opposite.

   iii. 95% confidence interval for odds ratio is

   $$exp\left(log(\hat{OR}) - 1.96SE(log(OR))\right) < \frac{p_1/(1-p_1)}{p_2/(1-p_2)} < exp\left(log(\hat{OR}) - 1.96SE(log(OR))\right)$$

   where $SE(log(OR)) = \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}}$

2. χ^2 test for contigency table:

   (a) Test the variables are independent in the contigency table

   (b) χ^2 test statistics

   $$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})}{E_{ij}}$$

   where r is the number of row, c is the number of collumn. $E_{ij} = \frac{\text{(row total)(column total)}}{\text{grand total}}$. O_{ij} is the observed count in row i and column j

   (c) If the two variables are independent, then X^2 has a χ^2 distribution with degree of freedom $(r-1)(c-1)$

   (d) For a 2×2 table the degree of freedom is 1.

3. G - test for contigency table:

   (a) G-test is based on the likelihood ratio test.

   (b) G test statistics:

   $$G = 2\left(\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij}log(\frac{O_{ij}}{E_{ij}})\right)$$

   (c) if the two variables are independent (under H_0), then G has a χ^2 distribution with d.f. of (r-1)(c-1)

II.Practice Problems

1. A study of randomly selected 15,513 births in the U.S. found a total of 5171 babies were born with a finger defect, either syndactyly (fused fingers), polydactyl(extra figers) or adactyly(few than five fingers). Among these babies, whether the mom smoked was recorded. Off these babies with finger defects, 4366 has monthers that did not smoke while pregnant, and the rest had mothers that did smoke while pregnant. Off those baby with normal fingers, 9062 of their mothers did not smoke while pregant, while the remaining 1280 did smoke while pregnant.

   (a) Examine a mosaic plot that compare the estimated conditional probability of finger defects given mother smoking or not.

   (b) Find a point estimate and a 95% confidence interval for the difference in the finger defects probabilities of mother smoking or not? Does smoking increase the baby finger defects?

   (c) Find a point estimate and a 95% confidence interval for the odds ratio of smoking or not. Interpret your result.

   (d) Perform the χ^2 test for the independence both by R and hand.

   (e) Perform the G-test for the independence by hand.

   (f) Relate the results of these hypothesis tests to what you saw in the mosaic plot. What features apparent in the plot do you think had the largest influence on the results of the tests?

   (g) The test statistics for the two tests involve summing over all cells in the table. Which cells contributed the most to these sums?

2. a sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently dieting on the other. We hypothesize, for example, that the proportion of dieting individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant. The data might look like this:

	men	women	Total
dieting	1	9	10
not dieting	11	3	14
Total	12	12	24

   a. Is it suitable for analysis by a chi-squared test?

   b. Perform Fisher's exact test on the observed data.

III Solutions of the Practice problems

1. (a) The observation table is

	defect	no defect	Total
smoke	805	1280	2085
no smoke	4366	9062	13428
Total	5171	10342	15513

```
> source('mosaic.R')
> finger = matrix(c(805,4366,1280,9062),nrow = 2, ncol = 2)
> rownames(finger) = c('smoke','no smoke')
> colnames(finger) = c('defect','no defect')
> finger
         defect no defect
smoke       805      1280
no smoke   4366      9062
> mosaic(t(finger))
```

(b) $\hat{p}_1 = 805/2085 = 0.3861, \hat{p}_2 = 4366/13428 = 0.3251, \hat{p}_1 - \hat{p}_2 = 0.061$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.3861 * (1 - 0.3861)}{2085} + \frac{0.3251 * (1 - 0.3251)}{13428}} = 0.0114$$

The 95% confidence interval of $p_1 - p_2$ is $0.061 - 1.96 * 0.0114 < p_1 - p_2 < 0.061 + 1.96 * 0.0114$, that is 0.0386 to 0.0833. Since 0 is not in the 95% confidence interval, with the 95% confidence we can say that smoking increases the baby finger defects.

(c) $\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = 1.3056$ and $log(\hat{OR}) = 0.2667$

$$SE(log(\hat{OR})) = \sqrt{1/805 + 1/1280 + 1/4366 + 1/9062} = 0.0486$$

The 95% C.I. for OR is $exp(0.2667 - 1.96 * 0.0486) < OR < exp(0.2667 + 1.96 * 0.0486)$, that is 1.1870 to 1.4361. The C.I. is above 1, so the smoking mother has higer odds to have a finger defect baby

(d) Using R

```
> chisq.test(finger)

        Pearson's Chi-squared test with Yates' continuity correction

data:  finger
X-squared = 29.8964, df = 1, p-value = 4.558e-08
```

By hand

	Expected			$\frac{(\text{Obs.-Exp.})^2}{Exp.}$	
	defect	no defect		defect	no defect
smoke	695	1390	smoke	17.41	8.71
no smoke	4476	8952	no smoke	2.70	1.35

$X^2 = 17.41 + 8.71 + 2.7 + 1.35 = 30.17$. the d.f = 1, the $p - value = P(\chi_1^2 > 30.17) = 3.957845e - 08$. P-value is very small, we reject the null hypothesis that smoking and baby finger defects are independent.

```
> 1-pchisq(30.17,1)
[1] 3.957845e-08
```

Obs.$log(\frac{Obs.}{Exp.})$		
	defect	no defect
smoke	118.28	-105.53
no smoke	-108.64	110.67

$G = 2 * (118.28 - 105.53 - 108.64 + 110.67) = 29.57$ the d.f =1, the $p - value = P(\chi_1^2 > 29.57) = 5.393326e-08$. P-value is very small, we reject the null hypothesis that smoking and baby finger defects are independent. You will see that the X^2 and G are very similar results.

(f) From the mosaic plot, we see that the smoke group has higher defects rates. The result is consistent which the mosaic plot.

(g) It is easy to see that the smoke and defect cell has the most important effects on the the test statistics. This is the main cell violate the independent hypothesis, because the smoke will increase the probability of defects

2. a, These data would not be suitable for analysis by a chi-squared test, because the expected values in the table are all below 10.

   b, The question we ask about these data is: knowing that 10 of these 24 teenagers are dieters, and that 12 of the 24 are female, what is the probability that these 10 dieters would be so unevenly distributed between the women and the men? If we were to choose 10 of the teenagers at random, what is the probability that 9 of them or even more would be among the 12 women?

   In another word, if we fix the marginal totals, what kinds of configuration will give us a dieting proportion of more than 9/12 (the observed proportion) among the women? The following two are the only cases, including the observed one.

	men	women	Total
dieting	1	9	10
not dieting	11	3	14
Total	12	12	24

	men	women	Total
dieting	0	10	10
not dieting	12	2	14
Total	12	12	24

   Hence, the p-value equals to

$$\frac{\binom{12}{9}\binom{12}{1}}{\binom{24}{10}} + \frac{\binom{12}{10}\binom{12}{0}}{\binom{24}{10}} = 0.001379728.$$

   You will find the following R command useful for calculating the above probabilities.

```
> sum(dhyper(9:10,12,12,10))
[1] 0.001379728
```

   Another way to run Fisher exact test is to use the build-in function of *fisher.test* in R.

```
> x=matrix(c(1,11,9,3),ncol=2)
> x
     [,1] [,2]
[1,]    1    9
[2,]   11    3
> fisher.test(x,alternative="less")
```

```
        Fisher's Exact Test for Count Data

data:  x
p-value = 0.001380
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.3260026
sample estimates:
odds ratio
0.03723312
```