STATISTICS 571 Discussion #5

I. Review

1. Bayes Theorem:
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$
   (a) A false negative is defined as a person who tests out as negative but who is actually positive
   (b) A false positive is defined as a person who tests out as positive but who is actually negative

2. 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, and $p_1 = p_2$; 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 $SE(log(OR)) = \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}}$

3. χ^2 test for contigency table:
   (a) Test the variables are independent in the contigent table
   (b) χ^2 test statistics
   $$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
   where r is the number of row, c is the number of collumn. $E_{ij} = \frac{\text{(ith row sum)(jth column sum)}}{(totalsum)}$. 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.

4. 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. Down syndrome (DS) is a chromosomal condition that occurs in about one in 1000 pregnancies. One test for DS is called the triple test, which screens for levels of three hormomes in maternal blood at around 16 weeks of pregnancy. The triple test is not perfect, however. It does not always correctly identifies a fetus with DS(an error called a false negative), and sometimes it incorrectly identifies a fetus with a normal set of chromosomes as DS (an error called a false positive). Under normal conditions, the detection rate of the triple test (i.e., the probability that a fetus with DS will be correctly scored as having DS) is 0.60. The false positive rate (i.e., the probability that a test would say incorrectly that a normal fetus has DS) is 0.05. Mary did a triple test, the result is positive, What is the probability that Mary has a real DS.

2. 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. Of these babies with finger defects, 4366 has mothers that did not smoke while pregnant, and the rest had mothers that did smoke while pregnant. Of 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?

III Solutions of the Practice problems

1. The question is to find P(DS|positive result). Before we do calculation, let us first to figure out what we have known. P(DS)=1/1000 = 0.001. P(positive result|DS) = 0.6, P(positive result|no DS) = 0.05. From these we can calculate(Law of total probability)
P(positive result) = P(positve result|DS)P(DS) + P(positive result|no DS)P(no DS) = 0.6*0.001+0.05*(1-0.001)=0.05055
Then we can use Bayes theorem

$$P(DS|\text{positive result}) = \frac{P(DS)P(\text{positve result}|DS)}{P(\text{positive result})} = \frac{0.60 * 0.001}{0.05055} = 0.012$$

2. (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