**BMI 713: Computational Statistics for Biomedical Sciences**

# Assignment 6

October 14, 2010 (due October 21)

## Contingency Table

1. (Taken from [1]) A statistical analysis that combines information from several studies is called a meta-analysis. A meta-analysis compared aspirin with placebo on incidence of heart attack and of stroke, seperately for men and from women (*J. Am. Med. Assoc.*, **295**: 306-313, 2006). For the Women's Health Study, heart attacks were reported for 198 of 19,934 taking aspirin and for 193 of 19,942 taking placebo. We are interested in whether aspirin was helpful for reducing the risk of heart attack.

   (a) State the null hypothesis and the alternative hypothesis.

   **Sol'n.** *Let $p_A$ be the true rate of heart attack among women who take aspirin and $p_P$ be the true rate of heart attack among women who are given placebo. Then $H_0 : p_A = p_P$ and $H_A : p_A \neq p_P$.*

   (b) Construct the $2 \times 2$ table that cross classifies the treatment (aspirin, placebo) with whether a heart attack was reported (yes, no).

   **Sol'n.** *Following is the table of observed values with marginal sums.*

| | *Aspirin* | *Placebo* | |
|---|---|---|---|
| *Heart attack* | *198* | *193* | *391* |
| *No heart attack* | *19,934* | *19,942* | *39,876* |
| | *20,132* | *20,135* | *40,267* |

   *The matrix can be directly constructed in `R`. We will use this later.*

```
> o <- matrix(c(198, 19934, 193, 19942), ncol=2)
> o
      [,1]  [,2]
[1,]   198   193
[2,] 19934 19942
```

   (c) Perform the chi-square test. Report the test statistic, the degree of the freedom and the P-value. What conclusion can you draw from this test?

   **Sol'n.** *To compute the table of expected values, we must first determine the proportion of heart attacks in the combined population:*

$$\hat{p} = \frac{391}{40,267} \approx 0.00971.$$

   *Using $\hat{p}$ and the number of patients in each of the Aspirin ($n_A$) and placebo groups ($n_P$), we can generate the expected value for each cell in the previous table:*

$$E = \begin{pmatrix} \hat{p} \\ 1-\hat{p} \end{pmatrix} \cdot (n_A, n_P) = \begin{bmatrix} \hat{p} \cdot n_A & \hat{p} \cdot n_P \\ (1-\hat{p}) \cdot n_A & (1-\hat{p}) \cdot n_P \end{bmatrix} \approx \begin{bmatrix} 195.49 & 195.51 \\ 19,936.51 & 19,939.49 \end{bmatrix}.$$

   *The matrix product $\binom{\hat{p}}{1-\hat{p}} \cdot (n_A, n_P)$ is sometimes called the* outer product*. The table of expected values above can be easily computed in `R` by using the `outer` function:*

```
> p.hat <- 391/40267
> n.A <- 20132
> n.P <- 20135
> e <- outer(c(p.hat, 1 - p.hat), c(n.A, n.P))
```

```
> e
           [,1]       [,2]
[1,]   195.4854   195.5146
[2,] 19936.5146 19939.4854
```

*Now we may compute the $\chi^2$ statistic by summing $\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$, where we let the indices $i$ and $j$ run over each cell in the table. That is,*

$$\begin{aligned}\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} &\approx \frac{(198-195.49)^2}{195.49} + \frac{(19,934-19,936.51)^2}{19,936.51} \\ &\quad + \frac{(193-195.51)^2}{195.51} + \frac{(19,942-19,939.49)^2}{19,939.49} \\ &= 0.0651.\end{aligned}$$

*Using the values `o` and `e` we computed in `R`, this sum can be computed in a compact `R` command:*

```
> sum((o - e)^2/e)
[1] 0.06532009
```

*The minor disagreement in value is due to the fact that the values computed by hand were rounded to two decimal places. In a $2 \times 2$ table with fixed margins, there is always exactly one degree of freedom. We can now compute the p-value:*

```
> 1 - pchisq(sum((o - e)^2/e), df=1)
[1] 0.7982767
```

*The p-value is quite high, so we cannot reject the null hypothesis $H_0$.*
*To perform the test automatically in `R`, use*

```
> chisq.test(o, correct=FALSE)

Pearson's Chi-squared test

data:  o
X-squared = 0.0653, df = 1, p-value = 0.7983
```

2. Sir Ronald Fisher, a statistician and geneticist, described a tea tasting experiment in his book *The design of Experiments* to illustrate his test – now known as Fisher's exact test. A colleague of Fisher claimed that she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment in which she tasted eight cups of tea. Four cups had milk added first, and the other four had tea added first. She was told there were four cups of each type and she should try to select the four that had milk added first. The cups were presented to here in random order. Table 1 shows a possible result of this experiment.

Table 1: Fiser's Tea Tasting Experiment

| | Guess Added First | | |
|---|---|---|---|
| Added First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | |

(a) State the null hypothesis and the alternative hypothesis.

**Sol'n.** *Let $p_M$ be Fisher's colleague's success rate of identifying milk-first cups and let $p_T$ be the colleague's* failure *rate of identifying tea-first cups. Then $H_0 : p_M = p_T$ and $H_A : p_M \neq p_T$.*

(b) Perform the Fisher's exact test. What is the P-value.

**Sol'n.** *Since the table has fixed margins, there are exactly 5 possible tables:*

$$\begin{vmatrix}0 & 4\\ 4 & 0\end{vmatrix}, \quad \begin{vmatrix}1 & 3\\ 3 & 1\end{vmatrix}, \quad \begin{vmatrix}2 & 2\\ 2 & 2\end{vmatrix}, \quad \begin{vmatrix}3 & 1\\ 1 & 3\end{vmatrix}, \text{ and } \begin{vmatrix}4 & 0\\ 0 & 4\end{vmatrix}.$$

*Each of these tables is not equally likely to occur. The hypergeometric distribution can be used to determine the p-value of each table via:*

$$P\left(\begin{vmatrix}a & b\\ c & d\end{vmatrix}\right) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

*The probabilities are, in order,*

$$0.01428571 \quad 0.22857143 \quad 0.51428571 \quad 0.22857143 \quad 0.01428571.$$

*Summing the probability of each table with equal or lesser p-value than the observed table gives:*

$$p = 0.01428571 + 0.22857143 + 0.22857143 + 0.01428571 \approx 0.486.$$

*With a p-value of 0.486, we will not reject the null hypothesis.*
*To perform the test in* `R`*, use:*

```
> x
     [,1] [,2]
[1,]    3    1
[2,]    1    3
> fisher.test(x)

Fisher's Exact Test for Count Data

data:  x
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

What conclusion can you get?

3. (Simpson's paradox). The result that a marginal association can have different direction from the conditional associations is called Simpson's paradox. This result applies to quantitative as well as categorical variables. To illustrate Simpson's paradox, here we use an example in medical study [2, 3] comparing the success rates of two treatments for kidney stones. The two treatments are open surgery (treatment A) and percutaneous nephrolithotomy (treatment B). The patients can be classified into two groups according to the kidney stone size, small stone group and large stone group. Table 2 shows the surgery results of 700 patients under the two treatments.

(a) What is the overall success rates of treatment A and treatment B? Based on this result, which treatment is better? Perform a proper test. What is the P-value?

**Sol'n.** *For the combined group, the success rate of each treatment is*

$$p_A = \frac{2730}{2730+770} = 0.78, \quad p_B = \frac{2890}{2890+610} \approx 0.826.$$

*In this case, treatment B has a higher success rate than treatment A. We can perform Fisher's exact test to determine if the difference is significant.*

Table 2: Success rates for different groups of stone size.

| | | Treatment | |
|---|---|---|---|
| Group | Treatment result | A | B |
| Small stone | Success | 810 | 2340 |
| | Failure | 60 | 360 |
| Large stone | Success | 1920 | 550 |
| | Failure | 710 | 250 |
| Both | Success | 2730 | 2890 |
| | Failure | 770 | 610 |

```
> m
     [,1] [,2]
[1,] 2730 2890
[2,]  770  610
> fisher.test(m, alternative="l")

        Fisher's Exact Test for Count Data

data:  m
p-value = 8.723e-07
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.8279782
sample estimates:
odds ratio
 0.7483852
```

*The p-value is very small, so we may conclude that treatment B is superior to treatment A.*

(b) For the small stone group, calculate the success rates of treatment A and treatment B. Which treatment if better? Perform a proper statistic test and report the P-value.

**Sol'n.** *In the small stone group, the success rates of the two treatments are:*

$$p_A = \frac{810}{810+60} \approx 0.931, \quad p_B = \frac{2340}{2340+360} \approx 0.867.$$

*In contrast to the combined group in part (a), treatment A seems more potent than treatment B. Furthermore, we see that this difference is significant according to Fisher's exact test:*

```
> s
     [,1] [,2]
[1,]  810 2340
[2,]   60  360
> fisher.test(s, alt="g")

        Fisher's Exact Test for Count Data

data:  s
p-value = 5.321e-08
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.625226      Inf
```

```
sample estimates:
odds ratio
  2.076559
```

(c) Repeat the above analysis for the large stone group.

**Sol'n.** *The success rates for treatments A and B are*

$$p_A = \frac{1920}{1920+710} \approx 0.730, \quad \frac{550}{550+250} \approx 0.688.$$

*Again, we see the opposite conclusion from that of part (a): in the large stone group, we find that treatment A is more successful than treatment B. Again we may apply Fisher's exact test to determine if A's success rate is significantly greater than B's:*

```
> fisher.test(l, alt="g")

        Fisher's Exact Test for Count Data

data:  l
p-value = 0.01108
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.058908      Inf
sample estimates:
odds ratio
  1.229086
```

*The p-value* $\sim 0.01$ *is significant at the* $\alpha = 0.05$ *significance level.*

(d) Are the conclusions from (a), (b) and (c) consistent? If not, can you explain why?

**Sol'n.** *The conclusions seem to be inconsistent. In the combined group, there is compelling evidence that treatment B is superior to treatment A; however, when examining each of the two groups individually, treatment A is shown to be superior to treatment B.*

*One explanation is that large stone patients are inherently harder to treat than small stone patients. In the small stone group, many patients chose the inferior treatment B and were successful due to the fact that their condition was simply easier to treat. On the other hand, many difficult, large stone cases opted for the superior treatment A but were unsuccessful.*

*So long as these two groups are kept separate, A's superiority is evident; however, once they are combined, the differences in the difficulty of treatment are lost and treatment B appears (falsely) to be the superior treatment.*

4. **For extra credit** (Conservativeness of Fisher's exact test). For small samples, because of the discreteness of the exact distribution used in Fisher's exact test, Fisher's exact test tends to be conservative, i.e. the real type I error rate is smaller than the nominal significance level. Here we use simulation to study this phenomenon.

(a) Generate two random numbers `n11`, `n21` from the Binomial distribution Binom$(10, 0.5)$.

```
> n11 <- rbinom(1, 10, 1/2)
> n21 <- rbinom(1, 10, 1/2)
> n11
[1] 6
> n21
[1] 4
```

(b) Use the following command to construct a matrix `A`, `A = matrix(c(n11,10-n11,n21,10-n21),nrow=2)`

```
> A <- matrix(c(n11, 10-n11, n21, 10-n21), nrow=2)
> A
     [,1] [,2]
[1,]    6    4
[2,]    4    6
```

(c) Use the matrix `A` as the input table and perform Fisher's exact test. What is the P-value? At the significance level 0.05, do you reject the null hypothesis?

**Sol'n.** *Performing Fisher's exact test:*

```
> fisher.test(A)

        Fisher's Exact Test for Count Data

data:  A
p-value = 0.6563
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.2773893 19.1425577
sample estimates:
odds ratio
  2.158166
```

*The p-value is 0.6563. At a significance level of* $\alpha = 0.05$*, we cannot reject the null hypothesis.*

(d) Repeat the above steps 1000 times. How many times do you reject the null hypothesis at the significance level 0.05? How many times do you expect to reject the null? Are these two numbers close?

**Sol'n.** *Use the following loop to run 1,000 simulations:*

```
> num <- 0
> for (i in 1:1000) {
+   n11 <- rbinom(1, 10, 1/2)
+   n21 <- rbinom(1, 10, 1/2)
+   A <- matrix(c(n11, 10-n11, n21, 10-n21), ncol=2)
+   if (fisher.test(A)$p.value < 0.05)
+     num <- num + 1
+ }
> num
[1] 19
```

*Thus, 19 out of 1,000 simulations were incorrectly rejected the null hypothesis—a type I error. The rate was 19/1000=0.019. The rate we expect to see is* $\alpha = 0.05$*, so the rate of rejection was lower than we expected.*

***Advanced note:*** *One way to perform this simulation in* `R` *is to use the* `apply` *family of functions. Simply create a function which returns 1 if a type I error occurs and 0 otherwise, then sum the value of this function applied 1000 times. Here is an example:*

```
> f
# f's argument i is just a dummy variable: we will never
# make use of it.  It is required by sapply
function(i) {
  n11 <- rbinom(1, 10, 1/2)
  n21 <- rbinom(1, 10, 1/2)
  A <- matrix(c(n11, 10-n11, n21, 10-n21), ncol=2)
  if (fisher.test(A)$p.value < 0.05)
    return(1)
```

```
  else
    return(0)
}
# Apply f to the values in 1:1000 (these will be the argument
# 'i', which we will ignore).
> sum(sapply(1:1000, f))
[1] 15
# Another simulation.
> sum(sapply(1:1000, f))
[1] 18
```

# References


[1] Agresti, Alan (2007) An Introduction to Categorical Data Analysis. Wiley

[2] Charig, C. R. D. R. Webb, S. R. Payne, and O. E. Wickham (1986). Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy. In *British Medical Journal*. 292:897-882.

[3] Julious, S. A., and Mullee, M. A. (1994). Confounding and Simpson's paradox In *British Medical Journal*. 309:1480-1481