October 9, 2017

STAT 545: Solutions to Homework 1

Dennis D. Cox
Department of Statistics
Rice University

October 9, 2017

1 Exercise 3.7

The data table appears below.

		Made Second?		
		Yes	No	Total
Made First?	Yes	152	33	185
	No	37	8	45
	Total	189	41	230

Performing the χ^2 test in R:

```
> kb = matrix(c(152,37,33,8),nrow=2)
> kb
     [,1] [,2]
[1,]  152   33
[2,]   37    8
starting httpd help server ... done
> kbtest = chisq.test(kb,correct = F)
> kbtest$p.value
[1] 0.9924673
```

We definitely cannot reject independence here. In fact, the χ^2 value is significant on the small side (1 - the p.value = 0.0075). That doesn't seem to mean anything here, but in a scientific study it could mean that someone is faking

the data and not putting enough variability in it. There have been examples of scientific misconduct that have been discovered by statisticians this way. But in this case, we conclude that there is no reason to doubt independence of the free throws.

2 Exercise 3.16

(a) We entered the data into R and did a chi^2 test of independence:

```
> tab3.16 = matrix(c(9,44,13,10,11,52,23,22,9,41,12,27),nrow=3,byrow=T)
> rownames(tab3.16) = c("low","mid","high")
> colnames(tab3.16) = c("HS","HSGrad","Col","ColGrad")
> tab3.16
     HS HSGrad Col ColGrad
low   9     44  13      10
mid  11     52  23      22
high  9     41  12      27
> test3.16a = chisq.test(tab3.16)
> test3.16a$p.value
[1] 0.1809674
```

The p-value is too large to reject the null hypothesis of independence.

(b) One more command gives the standardized residuals:

```
> test3.16a$stdres
             HS     HSGrad        Col    ColGrad
low   0.4061328  1.5828205 -0.1286367 -2.1078423
mid  -0.1898118 -0.5440627  1.3041565 -0.4031584
high -0.1903291 -0.9459053 -1.2374420  2.4360173
```

None of the standardized residuals are real big except the two in the last column which are larger than 2 in magnitude. The last column shows an increasing trend from lower family income to higher, so there is some evidence that for students from higher incomes having higher aspiration of finishing college. One also notes some tendency for the residuals in the first row to decrease, which is consistent with students from lower income families having lower educational aspiration.

(c) Since both variables (family income, educational aspiration) are ordinal, we expect that a test of ordinal association may be more appropriate.

The easiest one is to use Pearson's correlation with linear scores. I did a brief search for an R-package that does this and didn't find anything right away. The R manual that is linked to on the web page for the book just mentions "unpacking" the table, so that is what I did. It seems to be easiest to write a script of function to do this.

```
tab2xy = function(tab){
# function to unpack a 2-way table to a matrix of xy-values
# INPUT: tab: a I by J table of counts stored as a matrix
# OUTPUT: xy: a n by 2 matrix where n is the sum of the counts in tab
n = sum(tab)
xy  = matrix(NA,nrow=n,ncol=2)
I = nrow(tab)
J = ncol(tab)
k = 0
for(i in 1:I){
for(j in 1:J){
xy[(k+1):(k+tab[i,j]),] = cbind(rep(i,tab[i,j]),rep(j,tab[i,j]))
k = k+tab[i,j]
}
}
return(xy)
}
```

Then, we checked that it worked:

```
> source("/Users/dcox/MyStuff/Instruct/Stat545.17/Hw03solns/tab2xy.Rd")
> xy3.16 = tab2xy(tab3.16)
> table(xy3.16[,1],xy3.16[,2])

     1  2  3  4
  1  9 44 13 10
  2 11 52 23 22
  3  9 41 12 27
```

Now, we can do the test of correlation:

```
> test3.16b=cor.test(xy3.16[,1],xy3.16[,2])
> test3.16b$p.value
[1] 0.02905296
```

Now, we do get a statistically significant result at the 0.05 level. Of course, we would probably have expected higher family income is associated with higher aspiration, in which case a one sided test is appropriate, rather than the default two sided test:

```
> test3.16c=cor.test(xy3.16[,1],xy3.16[,2],alternative="greater")
> test3.16c$p.value
[1] 0.01452648
```

Of course, the p-value is cut in half. We could also look at the test based on a standardized estimate of γ:

```
> test3.16d = GKgamma(tab3.16)
> test3.16d$gamma/test3.16d$sigma
[1] 2.044439
> pnorm(-2.044439)
[1] 0.0204551
```

This shows positive association (the estimated gamma is positive) and is significantly positive at about the same level as we got with the test based on correlation of the linear scores.

Thus, we do find a significant positive association between family income and educational aspiration in this sample of high school students. The test using ordinal association is more powerful than the chi-squared test of independence for alternatives of positive association, so it is reasonable to use this type of test in this setting where we expect positive association.

3 Exercise 3.19

The data are shown in the table below.

	no lead	lead	Total
normal	18	7	25
malformed	7	7	14
Total	25	14	39

We entered the data into R in order to do the analysis.

```
> tab3.19 = matrix(c(18,7,7,7),nrow=2)
> rownames(tab3.19)=c("norm","mal");colnames(tab3.19)=c("noPb","Pb")
> tab3.19
     noPb Pb
norm   18  7
mal     7  7
> test3.19 = fisher.test(tab3.19,alternative="greater")
> test3.19$p.value
[1] 0.152573
```

At the usual 0.05 level of significance, we cannot reject the null hopothesis of no association or negative association vs. the alternative of positive association between the presence of lead (chemical symbol Pb) and malformations. The p-value is small, but not small enough. Assuming there is a biological reason to believe the alternative is true, I would suggest to the researcher to get a larger sample size. One could estimate a sample size that's "big enough" based on this prelimary data (e.g., use the table to estimate the cell probabilities, simulate data from the multinomial distribution with different sample sizes, perform the test, and estimate the sample size where H_0 is rejected, say, 80% of the time, 80% being a commonly used power level for estimating sample sizes).

4 Exercise 3.31

(a) We have the marginal probabilities

$$\begin{aligned} \pi_{1\cdot} &= \pi_{11} + \pi_{12} \\ &= \theta^2 + \theta(1-\theta) \\ &= \theta, \\ \pi_{\cdot 1} &= \pi_{11} + \pi_{21} \\ &= \theta. \end{aligned}$$

Clearly then $\pi_{11} = \pi_{1\cdot}\pi_{\cdot 1}$, and we have independence of the events $[X = 1]$ and $[Y = 1]$. For binary random variables, it suffices to check that any one of the four possible joint events (corresponding to the cells in the 2 by 2 table) are independent. Also, we have $\pi_{1\cdot} = \pi_{\cdot 1}$ so the marginal distributions are the same. Thus, X and Y are independent with the identical distribution.

(b) The multinomial likelihood (ignoring factors that do not depend on the parameter $(\pi_{11}, \pi_{12}, \pi_{21})$) is

$$\ell(\pi_{11}, \pi_{12}, \pi_{21}) \;=\; \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} (1 - \pi_{11} - \pi_{12} - \pi_{21})^{n_{22}}.$$

Under H_0, each π_{ij} is a function of θ, so we plug in the formulae for each of the cell probabilities in terms of θ and get a likelihood in terms of θ:

$$\begin{aligned} \ell(\theta) &= \theta^{2n_{11}} [\theta(1-\theta)]^{n_{12}} [\theta(1-\theta)]^{n_{21}} (1-\theta)^{2n_{22}} \\ &= \theta^{2n_{11}+n_{12}+n_{21}} (1-\theta)^{n_{12}+n_{21}+2n_{22}}. \end{aligned}$$

We take logarithms ($L = \log \ell$) and differentiate and set to 0:

$$\begin{aligned} \frac{d}{d\theta} L(\theta) &= \frac{2n_{11} + n_{12} + n_{21}}{\theta} - \frac{n_{12} + n_{21} + 2n_{22}}{1-\theta} \\ &= 0, \end{aligned}$$

This gives

$$\begin{aligned} (1-\theta)(2n_{11} + n_{12} + n_{21}) &= \theta(n_{12} + n_{21} + 2n_{22}) \\ (2n_{11} + n_{12} + n_{21}) &= 2\theta(n_{11} + n_{12} + n_{21} + n_{22}) \\ \hat{\theta} &= \frac{2n_{11} + n_{12} + n_{21}}{2(n_{11} + n_{12} + n_{21} + n_{22})} \\ &= \frac{n_{1\cdot} + n_{\cdot 1}}{2n} \\ &= \frac{p_{1\cdot} + p_{\cdot 1}}{2}. \end{aligned}$$

The second to last line follows from the definitions of $n_{1\cdot}$ and $n_{\cdot 1}$ and the fact that all the entries in the table add up to n. Of course, we need to check the sign of the second derivative of the log likelihood to make sure we have a maximum:

$$\begin{aligned} \frac{d^2 L}{d\theta} &= -\frac{2n_{11} + n_{12} + n_{21}}{\theta^2} - \frac{n_{12} + n_{21} + 2n_{22}}{(1-\theta)^2} \\ &< 0. \end{aligned}$$

This shows that the log likelihood is strictly concave, hence any stationary point is the unique maximizer of the log likelihood.

Now that was a lot of work, and I wonder if there isn't an easier way to do this problem. Looking back at the likelihood, it is essentially a Bernoulli trial likelihood

$$\ell(\theta) \;=\; \theta^y (1-\theta)^{m-y}$$

where we observe y successes in m trials and

$$\begin{aligned} y &= 2n_{11} + n_{12} + n_{21}, \\ m - y &= n_{12} + n_{21} + 2n_{21}, \\ m &= 2n_{11} + 2n_{12} + 2n_{21} + 2n_{21} \\ &= 2n. \end{aligned}$$

Now we know the MLE for such a model is $\hat{\theta} = y/m$, and that gives the result with a lot less effort.

(c) We use the chi-squared test statistic with

$$\begin{aligned} \hat{\mu}_{11} &= n\hat{\theta}^2 \\ \hat{\mu}_{12} &= n\hat{\theta}(1-\hat{\theta}) \\ \hat{\mu}_{21} &= \hat{\mu}_{12} \\ \hat{\mu}_{22} &= n(1-\hat{\theta})^2. \end{aligned}$$

The degrees of freedom are the dimension of the full parameter space (which is 3) minus the dimension of the null parameter space (which is 1), thus we have 2 degrees of freedom. The other way of computing it is to look at the number of independent constraints, which is 2, namely

$$\begin{aligned} \pi_{11} &= \pi_{1\cdot}\pi_{\cdot 1} \\ \pi_{1\cdot} &= \pi_{\cdot 1}. \end{aligned}$$

(d) The easiest way I can think of to do this is to have the R function compute the chi-squared value with the probability vector given by the estimated values under the null hypothesis, and then get the p-value by hand. The p-value that will be returned by the chisq.test function will be based on 3 degrees of freedom, so we need to do it with just 2 d.f.

```
> kb = matrix(c(152,37,33,8),nrow=2)
> thetahat = (2*152+33+37)/(2*230)
> test3.31d = chisq.test(as.vector(kb),p=c(thetahat^2,
```

```
+  thetahat*(1-thetahat),thetahat*(1-thetahat),(1-thetahat)^2))
> test3.31d$statistic
X-squared
0.2291154
> # not significant, but to get a p-value:
> pchisq(test3.31d$statistic,2,lower.tail=FALSE)
X-squared
0.8917605
```

The p-value of 0.8917605 is large - there is NO EVIDENCE against the null hypothesis that Kobe Bryant's pairs of free throws are anything other than independent with the same probability of success for each attempt.

5 Exercise 3.34

(a) To show this, it suffices to produce a table and then show the Pearson residuals are different. Let's try this:

```
> tab3.34 = matrix(c(20,15,10,30),nrow=2)
> test3.34 = chisq.test(tab3.34)
> test3.34$residuals
          [,1]      [,2]
[1,]  1.603567 -1.500000
[2,] -1.309307  1.224745
> test3.34$stdres
          [,1]      [,2]
[1,]  2.834734 -2.834734
[2,] -2.834734  2.834734
```

We see that the Pearson residuals (given with the command in the third line) are all different in absolute value, but the standardized residuals are all the same.

(b) This problem involves a lot of elementary algebra, which makes it rather tedious, but anyway let's get it over with. After a little playing around, the student should figure out that the square-rooted quantity in the denominator for each standardized residual is in fact the same. Using the facts that $\hat{\mu}_{ij} = np_{i\cdot}p_{\cdot j}$, and $1 - p_{1\cdot} = p_{2\cdot}$, etc., we have

$$\begin{aligned} \hat{\mu}_{11}(1-p_{1\cdot})(1-p_{\cdot 1}) &= np_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2} \\ \hat{\mu}_{12}(1-p_{1\cdot})(1-p_{\cdot 2}) &= np_{1\cdot}p_{\cdot 2}p_{2\cdot}p_{\cdot 1}, \end{aligned}$$

and so on. So it's really only the numerators where things can be different.

Writing $n_{ij} = np_{ij}$ and plugging in for the $p_{i\cdot}$ and $p_{\cdot j}$ and cranking through the algebra, we get

$$\begin{aligned} n_{11} - \hat{\mu}_{11} &= n\left[p_{11} - (p_{11} + p_{12})(p_{11} + p_{21})\right] \\ &= n\left[p_{11} - p_{11}^2 - p_{11}p_{21} - p_{12}p_{11} - p_{12}p_{21}\right] \\ &= n\left[p_{11}(1 - p_{11}) - p_{11}p_{21} - p_{12}p_{11} - p_{12}p_{21}\right] \\ &= n\left[p_{11}(p_{12} + p_{21} + p_{22}) - p_{11}p_{21} - p_{12}p_{11} - p_{12}p_{21}\right] \\ &= n\left[p_{11}p_{22} - p_{12}p_{21}\right]. \end{aligned}$$

One can do this for all four cells and get the desired answer.

It would be a lot of work to solve the problem by repeating the algebraic computation above three more times. I wonder if there is an easier way to solve this problem. We in fact know (or can easily see) that the row and column totals for the observed n_{ij} and "expected" $\hat{\mu}_{ij}$ are in fact the same: for example

$$\begin{aligned} \hat{\mu}_{1\cdot} &= \hat{\mu}_{11} + \hat{\mu}_{12} \\ &= \frac{n_{1\cdot}n_{\cdot 1}}{n} + \frac{n_{1\cdot}n_{\cdot 2}}{n} \\ &= \frac{n_{1\cdot}(n_{\cdot 1} + n_{\cdot 2})}{n} \\ &= \frac{n_{1\cdot}n}{n} \\ &= n_{1\cdot}, \end{aligned}$$

and similarly, $\hat{\mu}_{\cdot 2} = n_{\cdot 2}$, etc. Therefore

$$n_{11} - \hat{\mu}_{11} + n_{12} - \hat{\mu}_{12} \;=\; n_{11} + n_{12} - (\hat{\mu}_{11} + \hat{\mu}_{12}) \;=\; n_{1\cdot} - \hat{\mu}_{1\cdot} \;=\; 0.$$

Thus, the standardized residual $r_{21} = -r_{11}$. Similarly, $r_{12} = -r_{11}$ and $r_{22} = r_{11}$. This is the pattern we saw in the numerical example in part (a).

(c) Oh goodness, do I have to go through a lot of algebra? We observed above that they all have the same denominator, and the numerators have the same absolute value. Of course, X^2 is just the sum of squares of the Pearson residuals. Using the notations of equations (3.13) and (3.14), pp. 80-81, we

have

$$\begin{aligned} X^2 &= \sum_{i,j} e_{ij}^2 \\ &= \sum_{i,j} r_{ij}^2 (1 - p_{i+})(1 - p_{+j}) \\ &= r_{11}^2 \left(p_{2+}p_{+2} + p_{1+}p_{+2} + p_{2+}p_{+1} + p_{1+}p_{+1}\right) \\ &= r_{11}^2. \end{aligned}$$

By way of explanation, the first equation is just the definition of X^2 and the Pearson residual e_{ij}, the second equation follows from the definition of the standardized residuals r_{ij}, the third from the fact that the r_{ij} have equal absolute values and the relations between the row and column proportions noted before (e.g., $p_{+1} + p_{+2} = 1$), and the last equation follows easily. Note that $p_{i+}p_{+j}$ is the estimate of π_{ij} under independence, and this estimate is a probability vector.

6 Exercise 3.39

(a) Note that "combinations" refers to subsets of 3 individuals from among the 6 individuals listed. We know the number of such subsets is $\binom{6}{3} = 20$. Listing them out in a simple ordering scheme:

$$(F1, F2, F3), (F1, F2, M1), (F1, F2, M2), (F1, F2, M3), (F1, F3, M1),$$

$$(F1, F3, M2), (F1, F3, M3), (F2, F3, M1), (F2, F3, M2), (F2, F3, M3),$$

$$(F1, M1, M2), (F1, M1, M3), (F1, M2, M3), (F2, M1, M2), (F2, M1, M3),$$

$$(F2, M2, M3), (F3, M1, M2), (F3, M1, M3), (F3, M2, M3), (M1, M2, M3).$$

The contingency table: the sampling unit is a job applicant, and the two variables of interest are gender and success at being hired. The outcome $(F2, M1, M3)$ gives the table

	Hired	Not Hired
Female	1	2
Male	2	1

(b) I guess we are using $p_1 - p_2$ as the test statistic. The observed value is $2/3 - 1/3 = 1/3$, and we reject the null hypothesis $H_0 : \pi_1 \leq \pi_2$ for large

values of the test statistic. So the p-value can be computed as the *chance* of getting a sample with a test statistic value $\geq$ our observed value of 1/3. In this instance, $p_1 - p_2 \geq 1/3$ simply means that the males were in the majority of the set of candidates hired. Clearly half the samples have majority males and half have majority females, since the numbers of males and females in the population is the same and the sample size is odd (so no samples can have equal numbers of males and females). You can count which of the 10 of the 20 samples listed above have majority males.

I put the word *chance* in italics, since it refers to a probability. Our sample space in this case consists of subsets of size 3 selected from a population of 6 people. This isn't really a multinomial model for the data, but, as discussed in the book, may be thought of as basically the conditional distribution of the cell counts given the marginal sums. It can also be considered as a randomization distribution under the null hypothesis as discussed in class.

There is the question of whether this is a version of Fisher's test. For 2×2 tables, there are multiple test statistics that may be used. In the example 3.5.2, pp. 91-92, the author uses n_{11} as the test statistic. He is looking at a one sided alternative $\pi_1 > \pi_2$ exactly as in this problem. In the problem at hand, the author proposes we use $p_1 - p_2$ as the test statistic. Are they equivalent? Well, we can write

$$\begin{aligned} p_1 - p_2 &= \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} \\ &= \frac{n_{11}}{n_{1+}} - \frac{n_{+1} - n_{11}}{n_{2+}} \\ &= \left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right) n_{11} - \frac{n_{+1}}{n_{2+}}. \end{aligned}$$

We see that the above is strictly increasing as a function of n_{11}. The expression also depends on marginal totals (e.g., n_{1+}), but the randomization that gives rise to Fisher's null distribution preserves the marginal totals, i.e. Fisher's test is conditional on the marginal totals. Therefore, every table generated under Fisher's randomization distribution will have a value of $p_1 - p_2$ $\geq$ the observed value if and only if it has a value of $n_{11} \geq$ the observed value. So the two different test statistics give the same test in terms of rejecting or accepting the null hypothesis.