Midterm 1

Mohammad Mottahedi

March 4, 2016

Contents

1. (20 pts)

For the following statements, answer true (T) or false (F).

(a) **F**, In 2×2 tables, statistical independence is equivalent to an odds ratio of $\theta = 0$.

(b) **T**, If both row and column of a two-way contingency table are ordinal, Pearson's chi-squared test of independence and the Mantel-Haenszel statistic have the same power.

(c) **F**, For testing independence with random samples, Pearson's X2 statistic and the likelihood-ratio G2 statistic both have chi-squared distributions for any sample size, as long as the sample was randomly selected.

(d) **F**, In a 5 × 2 contingency table that compares 5 groups on a binary response variable, the G2 chi-squared statistic with df = 4 for testing independence can be exactly partitioned into 4 separate independent chi-squared statistics that each have df = 1 by comparing row 1 to row 5, row 2 to row 5, row 3 to row 5, and row 4 to row 5.

(e) **T**, When the bad outcome in a 2 × 2 contingency table is quite rare, relative risk is close to the odds ratio.

(f) **F**, When adjusting for overdispersion in regression models for discrete response, both the estimates of regression coefficients β s and their standard errors will become larger.

(g) **T**, If X and Y have at most five outcomes, and Z has K categories, one can test conditional independence of X and Y, controlling for Z, using the Cochran-MantelHaenszel test.

(h) **F** ,The difference in proportions, relative risk, and odds ratio are valid measures for summarizing 2 x 2 tables for either prospective or retrospective (case-control) studies.

(i) **F**, Suppose that income (high, low) and gender are conditionally independent, given type of job (secretarial, construction, service, etc). Then, income and gender are also independent in the 2 × 2 marginal table, ignoring type of job.

(j) **F**, In a GLM with three predictors X1, X2, X3 and one predicted variable Y , the term b0 + b1X1 + b2X2 + b3X3 is the expected value of Y .

2. (13 pts)

Considering the following data from a hypothetical retrospective health study in which patients with a disease were matched against similar individuals without.

Table 1: Risk factor and disease

		Disease	
		Yes	No
Risk Factor	Present	46	22
	Absent	50	55

(a)

(5pts) Construct a 95% confidence interval for the population odds ratio.

$\hat{\theta} = \frac{P(present|Yes)/P(Absent|yes)}{P(present|no)/P(Absent|no)} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 2.3$

$\hat{V}(log\hat{\theta}) = 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22} = 0.1099209$

log confidence interval:

$log\theta \pm 1.96\sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}} = (0.1830843, 1.482734)$

confidence interval:

$(1.2009156, 4.4049722)$

(b)

(5 pts) Based on the answer to part (a), does it seem plausible that the variables are independent? Explain, and if not discuss the effect.

The odds ratio is 2.3 which is greater than 1 and the 95% confidence interval does not contain 1 which implies that the two variables are dependent to each other.

(c)

(3 pts) Can we make a statement about relative risk?

Since this is a prospective study it's not possible to directly calculate the relative risk. But if the conditional properties of the disease is small ($\pi_{1|1}$ and $\pi_{1|2}$) we can approximate the relative risk using calculated odds ratio.

$\frac{1-\pi_{1|1}}{1-\pi_{1|2}} = 0.6794118$

Since probability of disease regardless of risk factor is large (0.55) the calculating relative risk using odds ratio is not appropriate.

3.

(15 pts) Table below shows data from the 2002 General Social Survey cross classifying a person's perceived happiness with their family income. The table displays the observed and expected cell counts and the standardized residuals for testing independence.

(a)

Show how to obtain the estimated expected cell count of 79.7 for the second cell in the first column.

first we have to find marginal total: $n_{2+} = 646$ and $n_{+1} = 168$

$\hat{n_{21}} = \frac{n_{2+}n_{+1}}{n} = 79.6828194$

(b)

For testing independence, X2 = 73.4. report the df value and the p-value, and state your conclusion.

$df = (I-1)(J-1) = (3-1)(3-1) = 4$

p-value = $4.3298698 \times 10^{-15}$

The p-value is very close to zero and we can reject the independence null hypothesis and say the two variables are associated.

(c)

Interpret the standardized residuals in the corner cells having counts 110 and 94. That is, say something about the magnitude and direction, and what if anything they say about the lack of fit.

The number of very happy people in the above average category and the not so happy people in the below average income level is much higher from what is expected based on independence model. These two cells contribute to lack of fit because of the standardized residuals that are greater than 3. Specially the not too happy peoples with below average income which has 7.37 standard residual.

4.

(14pts) Consider a three-way contingency tables with variables X, Y, and Z, where each has 3 possible outcomes (so we can have X = 1, X = 2, or X = 3 and similarly for Y and Z).

(a)

(5 pts) List the possible loglinear models for the joint distribution (use loglinear model notation; for example one of the models is (XY, Y Z)).

Model	Log Linear notation	df
saturated	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda ijk^{XYZ}$	26
Homogeneous association	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	18
conditional independence (XY,XZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	14
conditional independence (XY, YZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	14
conditional independence (XZ, YZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	14
joint independence (XY, Z)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	10
joint independence (XZ, Y)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	10
joint independence (YZ, X)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	10
complete independence (X,Y,Z)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	6

(b)

(5 pts) Give the model degrees of freedom for each (hint, this is the number of parameters needed to specify a point on the model).

Model	Log Linear notation	df
saturated	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda ijk^{XYZ}$	26
Homogeneous association	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	18
conditional independence (XY,XZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	14
conditional independence (XY, YZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	14
conditional independence (XZ, YZ)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	14
joint independence (XY, Z)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	10
joint independence (XZ, Y)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	10
joint independence (YZ, X)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	10
complete independence (X,Y,Z)	$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	6

(c)

(4 pts) Give the degrees of freedom for the $\Delta G2$ test comparing (XY, Y Z) to the homogeneous association model.

$df = (18 - 14) = 4$

5.

(18 points) For data from Florida on Y= whether someone convicted of multiple murders receives the death penalty (1=yes, 0=no), the prediction equation is logit($\hat{\pi}$) = -2.06 +0.87d - 2.4v , where d and v are defendant's race and victims' race (1=black, 0=white). The following are true-false questions based on the prediction equations.

(a) **T**, The estimated probability of the death penalty is lowest when the defendant is white and victim is black.

(b) **F**, Controlling for victims' race, the estimated log odds of the death penalty for white defendants equals 0.87 times the estimated log odds for black defendants. If we instead let d = 1 for white defendants and 0 for black defendants, the estimated coefficient of d would be -0.87.

(c) **T** If the estimated odds ratio between the death penalty outcome and defendant's race is different for different categories of victims' race, an interaction term is needed.

(d) **T** The intercept term exp(-2.06)/(1+e-2.06) is the estimated probability of the death penalty when the defendant and victims were white (i.e. d = v = 0).b

(e) **T** If there were 500 cases with white victims and defendants, then the model fitted count (i.e. estimated expected frequency) for the number who receive the death penalty equals 500exp(-2.06)/(1+e-2.06)

(f) **F**, This model assumes that d and y are conditionally independent given v.

6.

a (5pts)

```
result <- chisq.test(table, correct = FALSE)
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 13.874, df = 4, p-value = 0.007708
```

the χ^2 test statistic is large with p-value 0.0078 (<0.05 with 95% confidence interval) so we can reject the null hypothesis of independence and say the husband and wife height are not independent.

b (5 pts)

tall-medium vs short

tall-medium	short
174	103
95	35

```r
table <- matrix(c(174,95,103,35), nrow = 2)
result <- chisq.test(table, correct = FALSE)
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 4.1569, df = 1, p-value = 0.04147
```

χ^2 value is 4.15 with p-value equal to 0.04 so we can reject the null hypothesis and say the two variables are not independent.

tall vs midium-short

tall	Medium-short
32	83
56	236

```r
table <- matrix(c(32,56,83,236), nrow = 2)
result <- chisq.test(table, correct = FALSE)
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 3.6411, df = 1, p-value = 0.05637
```

χ^2 value is 3.6 with p-value > 0.05, so we cannot reject the null hypothesis which suggest that the two variables are independent.

(c) (5 pts)

```r
table <- matrix(c(32,25,31,37,80,64,46,57,35), nrow = 3, dimnames = list(husband = c("tall", "medium",
pears.cor(table, c(3,2,1), c(3,2,1))
```

```
## [1] -0.05108454  1.05950985
```

$r = -0.05108454$, $M^2 = 1.05950985$, df =1

p-value = 0.303327

The Mantel-Haenszel statistic is small and the p-value shows that we cannot reject the null hypothesis under the independence model (r=0, M2=0).

(d)

Analysis in part b showed that we can weakly reject the null hypothesis of independence in the case of short vs medium-tall and the short-medium vs tall are independent of each other. The Pearson correlation and MH statistics showed that there is small negative correlation between the height of husband and wife (tall man tend to have shorter wives) but MH statistic showed that this relation is not significant (p-value = 0.3) and we cannot reject the null hypothesis. So we can assume that there's not a significant relationship between height of the husband and wife.