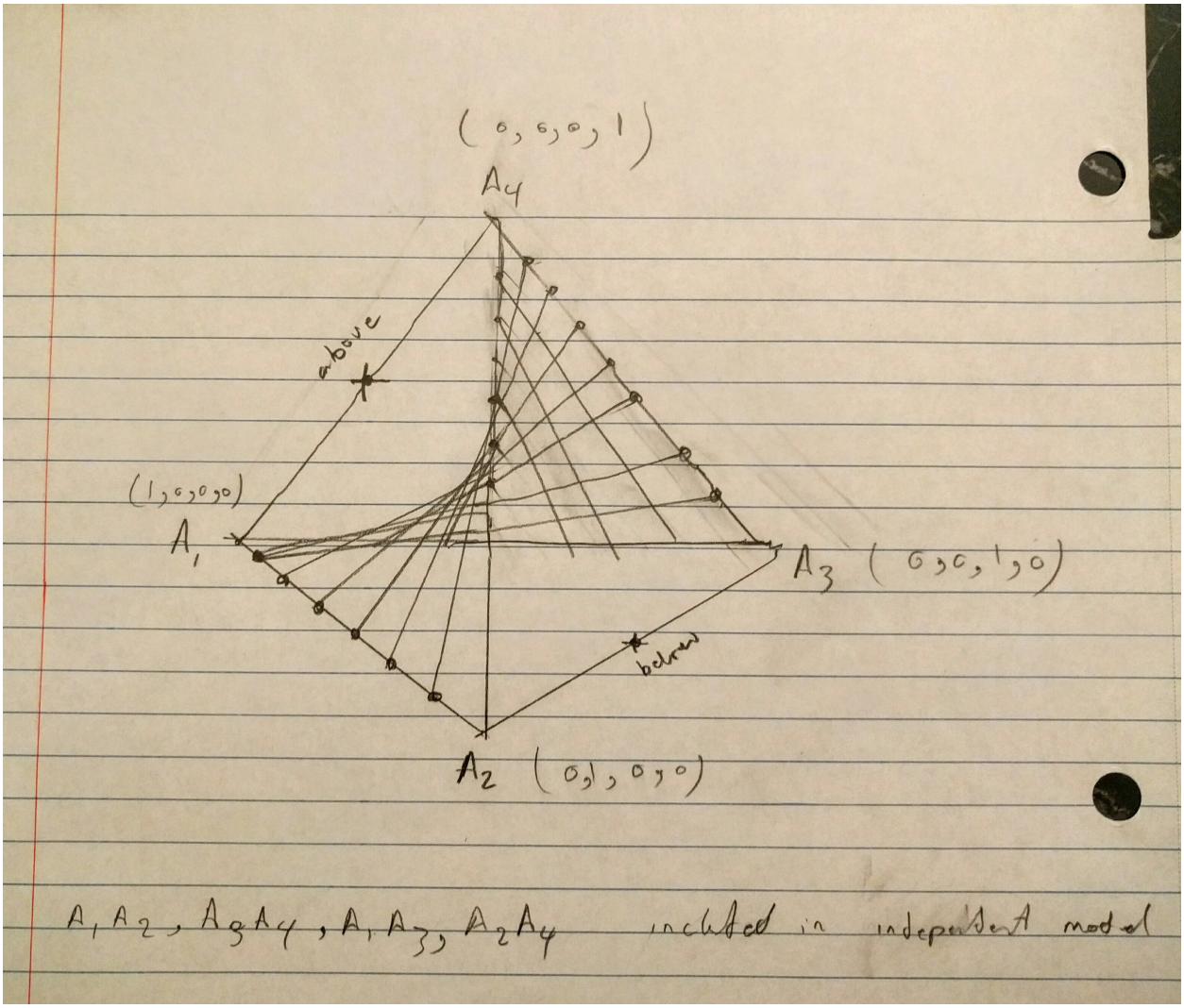
hw3

Mohammad Mottahedi

February 2, 2016

```
set.seed(1235)
```

1. (5 pts) The probability simplex for the four possible outcomes of two binary random variables is a tetrahedron. Draw the independence surface (the set of probability distributions corresponding to the independence model for a 2 ×2 table) inside the tetrahedron. What does it mean for the empirical distribution to be above or below this surface?



The independence surface is the hyperbolic paraboloid in the above picture. Points above independece surface indicates positive association between the two variables and points bellow it indicates and negative association.

2. (5 pts) True or False?

(a) In 2 × 2 tables, statistical independence is equivalent to a population odds ratios value of $\theta = 1$.

True.

(b) With rare events, when population proportions are very close to zero, difference of proportions is a better measure of association than the relative risk?

False.

3. (10 pts) Give an example of inferences in a two-way table for which multinomial, product multinomial, and Poisson sampling assumptions are appropriate.

Using the incidence of common cold involving french skiers example:

	Cold	No Cold	Total
Placebo	31	109	140
Vitamin C	17	122	139
Totals	48	231	279

Product multinomial is appropriate for inference about difference of proportion of getting cold given placebo or vitamin C.

Poisson sampling assumption is appropriate for infernece about proportion of skiers having cold and not having cold.

4. (30 pts) For adults who sailed on the Titanic on it fateful voyage, the odds ratio between gender (female, male) and survival (yes , no) was 11.4.

(a) What is wrong with the following interpretation: "The probability of survival for females was 11.4 times that for males"? Give the correct interpretation.

The odds of survival for women is 11.4 times the odds of survival for men but the probabilities and odds are not the same and the above statement is not correct. The correct intrepretations is that the odds of survival was 11.4 times given the person was a female.

(b) The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

$\pi_{s|f} = \frac{odd_f}{1+odd_f} = \frac{2.9}{3.9} = 0.74$

$\theta = \frac{odd_f}{odd_m} = 11.4$

$odd_m = 0.254$

$\pi_{s|m} = \frac{odd_m}{1+odd_m} = \frac{0.25}{1.25} = 0.2$

(c) Find the value of R in the interpretation, "The probability of survival for females was R times that for males".

$R = \frac{0.74}{0.2} = 3.7$

5. (20 pts) A handwriting expert claims to have ability to discern whether a note was written by a man or a woman. An experiment was performed in which five men and five women submitted handwriting samples. The samples were then presented to the judge (who was

Experts classification

True gender	Male	Female
Male	4	1
Female	1	4

told their were five men and five women), and he rendered his opinion on which samples were male and which were female.

Perform both approximate and exact tests. What is your conclusion regarding the expert's claim?

the exact test:

H_0: expert has no ability to decern odd ratio is equal to 1

H_A: expert has no ability to decern odd ratio is less or greater than 1

```
hw <- matrix(c(4,1,1,4), ncol = 2,
             dimnames = list(True = c("male", "female"),
                             expert_classification = c("male", "female")))

fisher.test(hw)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  hw
## p-value = 0.2063
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.45474 968.76176
## sample estimates:
## odds ratio
##    10.9072
```

there's not enough evidence to reject the null hypothesis.

approximate test:

```
hw <- matrix(c(4,1,1,4), ncol = 2,
             dimnames = list(True = c("male", "female"),
                             expert_classification = c("male", "female")))

result <- chisq.test(hw, correct = F)
```

```
## Warning in chisq.test(hw, correct = F): Chi-squared approximation may be
## incorrect
```

```
result
```

```
##
##  Pearson's Chi-squared test
##
## data:  hw
## X-squared = 3.6, df = 1, p-value = 0.05778
```

```
result$expected
```

```
##         expert_classification
## True     male female
##   male    2.5    2.5
##   female  2.5    2.5
```

```
result$residuals
```

```
##         expert_classification
## True           male     female
##   male    0.9486833 -0.9486833
##   female -0.9486833  0.9486833
```

```
LR <- 2*sum(hw*log(hw/result$expected))
```

p-value for χ_2 is greater than 0.05 and we can reject the null hypothesis which is contrary to exact test.

6. (30 pts) Sensitivity and specificity are measures often calculated for 2×2 tables. These measures are of particular interest when you want to determine the efficacy of a screen- ing test (e.g. for a disease, lie detection, etc.). To learn a bit about these measures, you should read the "AccuracyHandout.pdf" (see online Lesson 3.1.7: Difference in proportions https://onlinecourses.science.psu.edu/stat504/node/ 77). In Agresti(2013) you can find some information in Sec. 2.1.2. Of course, you can always search the web, or use other textbooks.

(a) (10 pts) a) How do sensitivity and specificity relate to conditional probabilities, rel- ative risk and odds ratios?

sensitivity is the conditional probability of a true positive test and specificity is the conditional probability of a true negative test.

odds ratio can be described as the ratio of sensitivity to specificity.

(b) (20 pts) PET scans can be used in diagnosing a certain type of cancer, such as lung cancer. Consider following hypothetical data on 150 adults in a mining town:

Table 1:default

Test result

Cancer	Positive	Negative
yes	65	5
no	3	77

Are the test results and lung cancer status independent? Report the results of chi- squared test of independence, and interpret.

```
test <- matrix(c(65,3,5,77), ncol = 2,
               dimnames = list(cancer = c("yes", "no"),
```

```
                                result = c("positive","negative")))

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

```
##
##  Pearson's Chi-squared test
##
## data:  test
## X-squared = 119.61, df = 1, p-value < 2.2e-16
```

```
#result$expected
#result$residuals
#result$observed
LR <- 2*sum(test*log(test/result$expected))
LR
```

```
## [1] 145.0244
```

```
1 - pchisq(LR, df=1)
```

```
## [1] 0
```

the χ^2 value is very large so the independence model is not correct.

Describe the association of the test results and cancer using any of the measures of associations you deem appropriate (e.g., difference in proportion, the relative risk and the odds ratio) and interpret your findings. Also report the sensitivity, specificity, false positive and false negative rates?

```
require(vcd)
```

```
## Loading required package: vcd
## Loading required package: grid
```

```
test <- matrix(c(65,3,5,77), ncol = 2,
               dimnames = list(cancer = c("yes", "no"),
                               result = c("positive","negative")))


RowSums=rowSums(test)
ColSums=colSums(test)

#COLUMN 1 RISK ESTIMATE
risk1_col1=test[1,1]/RowSums[1]
risk2_col1=test[2,1]/RowSums[2]
rho1=risk1_col1/risk2_col1
total1=ColSums[1]/sum(RowSums)


#COLUMN 2 RISK ESTIMATE
risk1_col2=test[1,2]/RowSums[1]
```

```r
risk2_col2=test[2,2]/RowSums[2]
total2=ColSums[2]/sum(RowSums)



#relative risk
rho1=risk1_col1/risk2_col1
rho2 = risk2_col2/risk1_col2
rbind(rho1,rho2)
```

```
##          yes
## rho1 24.7619
## rho2 13.4750
```

```r
# difference of proportion column two
diff1=risk2_col1-risk1_col1
diff2=risk2_col2-risk1_col2
rbind(risk1_col1,risk2_col1,diff1)
```

```
##                   yes
## risk1_col1  0.9285714
## risk2_col1  0.0375000
## diff1      -0.8910714
```

```r
rbind(risk1_col2,risk2_col2,diff2)
```

```
##                   yes
## risk1_col2 0.07142857
## risk2_col2 0.96250000
## diff2      0.89107143
```

```r
SE_diff2=sqrt(risk1_col2*(1-risk1_col2)/RowSums[1]+risk2_col2*(1-risk2_col2)/RowSums[2])
CI_diff2=cbind(diff2-qnorm(0.975)*SE_diff2,diff2+qnorm(0.975)*SE_diff2)
SE_diff2
```

```
##        yes
## 0.03739911
```

```r
CI_diff2
```

```
##         [,1]      [,2]
## no 0.8177705 0.9643723
```

```r
#odds Ratio
oddsratio(test, log=F)
```

```
##  odds ratios for cancer and result
##
## [1] 333.6667
```

```r
exp(confint(oddsratio(test)))
```

```
##                               2.5 %   97.5 %
## yes:no/positive:negative 76.80031 1449.648
```

```r
#sensitibity
65/(65+3)
```

```
## [1] 0.9558824
```

```r
#specificity
77/(77+5)
```

```
## [1] 0.9390244
```