

Statistics with R

Categorical Inferences

Scott Hetzel

University of Wisconsin – Madison

Summer Institute for Training in Biostatistics (2009)

Derived from: “Introductory Statistics with R” by: Peter Dalgaard

and from previous notes by Deepayan Sarkar, Ph.D

What we Discussed Last Time

- Importing data sets with `read.table()`
- Random sampling with `sample()`
- The Binomial and Normal Distributions
- Generating tables with `table()`
- Creating bar plots to graphically represent tables with `barplot()`

Basic Categorical Inferences

Dr. Gangnon will talk about basic inferences for continuous variables and how to conduct these tests in **R** later this week. Most likely including functions like: `t.test()`, and `wilcox.test()`, see chapter 4 of the text if you would like to take an early peek.

Today we will talk about basic inferences for categorical variables, which is most of chapter 7 in the text.

Categorical Inferences: Binomial Test

Remember the example that stated that 9.8% of young adults between the ages of 18 and 24 are left handed? Let's say we wanted to confirm or refute that statement. A possible way of doing this is to collect a random sample of 18 to 24 year olds and count the number of left handers in that sample and compare the proportion to 9.8%.

There is a function in **R** that will test the hypothesis of a population probability of success in a binomial distribution: `binom.test(x, n, p)`

Can you think of a way we could get a random sample of say.. I don't know, 20, 18 to 24 year olds??? Ok so lets count the number of left handers among us. Does this sample refute or fail to refute that 9.8% of 18 to 24 year olds are left handed? Possible outcomes could be:

```
> binom.test(3,20,0.098)$p.value
```

```
[1] 0.438776
```

```
> binom.test(0,20,0.098)$p.value
```

```
[1] 0.2530
```

```
> binom.test(5,20,0.098)$p.value
```

```
[1] 0.04006
```

Two Independent Proportions

`prop.test` can be used to do a very similar test to that of `binom.test` however a continuity correction is used in `prop.test` and `binom.test` is an exact probability test therefore `binom.test` is preferable in the single proportion testing. However, `prop.test` can make a test of multiple proportions that `binom.test` cannot.

`prop.test` can return a list of items that you can isolate and look at individually with the `$` operator. Such as: `prop.test()$p.value`, `prop.test()$conf`, and `prop.test()$estimate`. See `str(prop.test())` for full list.

As an example, let's say that sample A had 450 subjects and 120 of them had disease X. Sample B had 500 subjects and 175 of them had disease X. Is there a difference in the proportion of people with disease X by sample?

```
> prop.test(x=c(120,175), n=c(450,500))$p.value
```

```
[1] 0.006904
```

```
> prop.test(x=c(120,175), n=c(450,500))$conf
```

```
[1] -0.14390160 -0.02276507 # Confidence interval around the difference estimate
```

Two Independent Proportions

As with the situation of one proportion `prop.test()` uses a continuity correction which makes the confidence intervals a little wider so you are more likely to fail to reject. If you want at least be sure that the p-value is correct, you can use Fisher's exact test. The relevant function is `fisher.test`, which requires the data to be given in matrix form.

For the example from the last slide, to use the Fisher Exact test we would do the following:

```
> mat <- matrix(c(120, 175, 330, 325), nrow=2)
> rownames(mat) <- c("A", "B")
> colnames(mat) <- c("S", "F")
> mat
      S    F
A  120  330
B  175  325
> fisher.test(mat)$p.value
[1] 0.006155627
```

Note: The inference is based on the odds ratio and not the difference in proportions. The null is that the odds ratio is equal to 1.

Chi-Square Test

For the analysis of tables with more than two levels on at least one side, we use `chisq.test`. There are three ways a table may arise like this:

- The total in each row might be fixed in advance and you would be interested in testing whether the distribution over columns is the same for each row.
- The opposite, column totals might be fixed and we would be interested in if the distribution over the rows is the same for each column.
- Only the total number is chosen and the individuals are grouped randomly according to the row and column criteria. This case we would be interested in testing the hypothesis of statistical independence.

The convenient thing is that no matter the case, the analysis is the same.

Chi-Square Test (Cont.)

If there is no relation between rows and columns, then you would expect to have the following cell values:

$$E_{ij} = \frac{n_{i.} * n_{.j}}{n_{..}}$$

The test statistic:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

has a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom. The sum is over the entire table. O denotes the observed values and E the expected values in each field of the table.

Chi-Square Test (Cont.)

Consider the following data:

```
> caff.material <- matrix(c(652,1537,598,242,36,46,38,21,218,327,106,67),
+ nrow=3,byrow=T)
> colnames(caff.material) <- c("0","1-150","151-300",>300")
> rownames(caff.material) <- c("Married", "Prev.Married", "Single")
> caff.material
```

	0	1-150	151-300	>300
Married	652	1537	598	242
Prev.Married	36	46	38	21
Single	218	327	106	67

```
> chisq.test(caff.material)$p.value
[1] 2.187e-09
```

The p-value is highly significant so we can say that the null hypothesis of independence is rejected and there is a dependence between caffeine intake and marital status. However, we are not sure where or the nature of the deviations.

Chi-Square Test (Cont.)

```
> round(chisq.test(caff.material)$expected, 2)
      0      1-150    151-300    >300
Married    705.83  1488.01    578.07  257.09
Prev.Married  32.86   69.27    26.91  11.97
Single    167.31   352.72   137.03   60.94
> chisq.test(caff.material)$observed
      0      1-150    151-300    >300
Married    652    1537     598     242
Prev.Married  36     46     38     21
Single    218    327    106     67
```

A table of contributions to the test statistic can be helpful to see which deviations are contributing to the significant test statistics. This is not readily available but can be computed.

```
> E <- chisq.test(caff.material)$expected
> O <- chisq.test(caff.material)$observed
> (O-E)^2/E
      0      1-150    151-300    >300
Married    4.1055981  1.612783  0.6874502  0.8858331
Prev.Married  0.3007537  7.815444  4.5713926  6.8171090
Single    15.3563704  1.875645  7.0249243  0.6023355
```

There are some large contributions, particularly from too many non-caffeine drinking singles, and the distribution among previously married is shifted in the direction of a larger intake.

Exercises in Using R

There is a 3-dimensional array of data in R called `HairEyeColor`. We will use this data to practice some of the things we just learned and some things from the past.

1. First of all, give `HairEyeColor` a shorter name so it is easier to type. How about HEC.
2. Using indexing, how many Hazel eyed, Brown haired females are there? And how many Green eyed, Blond men are there? Hint: Look at `str(HEC)` to get index order.
3. How many males are there? Females?
4. Randomly selecting one person from the sample, what is the probability that they will have Blue eyes?
5. What is the probability that they will have Brown or Blue eyes and is a man?
6. Given that the person does NOT have Red hair, what is the probability that the person is female with Brown or Hazel eyes?

Exercises in Using R Answers

1.

```
> HEC <- HairEyeColor
```
2.

```
> HEC["Brown", "Hazel", "Female"]  
[1] 29  
> HEC["Blond", "Green", "Male"]  
[1] 8
```
3.

```
> sum(HEC[, , "Male"])  
[1] 279  
> sum(HEC[, , "Female"])  
[1] 313
```
4.

```
> sum(HEC[, "Blue", ])/sum(HEC)  
[1] 0.3631757
```
5.

```
> sum(HEC[, c("Brown", "Blue"), "Male"])/sum(HEC)  
[1] 0.3361486
```
6.

```
> sum(HEC[-3, c("Brown", "Hazel"), "Female"])/sum(HEC[-3, , ])  
[1] 0.2783109
```

Exercises in Using R

1. It is hypothesized that the proportion of Black haired people is about 23% of the population. Does this sample uphold or refute that statement?
2. Test if there is a difference in the proportion of men with Blond hair compared to the proportion of women with Blond hair.
3. For Blue eyed people is there an association between hair color and gender? In other words is the distributions of hair colors different for blue eyed women compared to blue eyed men?

Exercises in Using R Answers

- ```
> sum(HEC["Black",,])/sum(HEC)
[1] 0.1824324
> binom.test(sum(HEC["Black",,]), sum(HEC), 0.20)$p.value
[1] 0.3042261
```
- ```
> NofB <- c(sum(HEC["Blond",,"Female"]), sum(HEC["Blond",,"Male"]))
> NofB
[1] 81 46
> NTot <- c(313, 279)
> prop.test(NofB, NTot)$p.value
[1] 0.007399568
```
- ```
> HEC[, "Blue",]
 Male Female
Black 11 9
Brown 50 34
Red 10 7
Blond 30 64
> chisq.test(HEC[, "Blue",])$p.value
[1] 0.001544
```

# Exercises in Using R

---

Use the cancer data in the `survival` package to do the following:

1. Turn the variable `sex` into a factor with levels "M"=1 and "F"=2
2. Turn the variable `status` into a factor with levels "Censored"=1, "Dead"=2
3. Is there an association between death and sex? Report using the odds ratio.

# Exercises in Using R Answers

---

```
1. > sex.f <- factor(cancer$sex, levels=c(1,2), labels=c("M", "F"))

2. > status.f <- factor(cancer$status, levels=c(1,2),
+ labels=c("Censored", "Dead"))

3. > tabl <- table(sex.f, status.f)
> tabl
 Censored Dead
M 26 112
F 37 53
> fisher.test(tabl)$p.value
[1] 0.0004349142
> fisher.test(tabl)$estimate
odds ratio
0.3342709
```