BIO5312: R Session 8
Categorical data analysis

Yujin Chung

October 18th, 2016

Fall, 2016

Today's R session

- `chisq.test()`
- `fisher.test()`
- `table()`
- `dhyper()`, `phyper()`

table(): contingency table

```
> x1 = rbinom(100, 1, prob=c(.1,.9))
> table(x1)
x1
 0  1
49 51

> x2 = rbinom(100, 5, prob=c(.1,.9))
> table(x1,x2)
    x2
x1    0  1  2  3  4  5
  0 25 15  5  0  4  0
  1  2  2  1  3 16 27
```

table()

```
> x3 = rbinom(100, 2, prob=c(.1,.9))
> table(x1,x2,x3)
, , x3 = 0

   x2
x1   0  1  2  3  4  5
  0 18 14  3  0  0  0
  1  2  2  1  0  1  0

, , x3 = 1

   x2
x1   0  1  2  3  4  5
  0  7  1  2  0  2  0
  1  0  0  0  2  2  8
```

Goodness-of-fit test

```
> phenotype = read.table("phenotype.txt")
> obs = table(phenotype)
> obs
phenotype
  G   Y
140 416

## H0: Y:G = 3:1
> res = chisq.test(obs, p= c(1/4,3/4))
> res
Chi-squared test for given probabilities
X-squared = 0.0095923, df = 1, p-value = 0.922
```

Goodness-of-fit test

```
> res$expected
  G   Y
139 417
> ev = res$expected
> sum((obs-ev)^2/obs) # test statistic
[1] 0.009546703
```

Test of homogeneity

```
> lungcancer = read.table("LungCancer.txt",header=T)
> table(lungcancer)
        cancer
smoking    lc    nc
     cs    50  9950
     ns    10 39990
     xs    25 49975
> obs = table(lungcancer)
> res=chisq.test(obs)
> res
Pearson's Chi-squared test

data:  obs
X-squared = 226.96, df = 2, p-value < 2.2e-16
```

Test of independence

```
> lungcancer = read.table("LungCancer_drinkingHabit.txt"
      ,header=T)
> obs = table(lungcancer)
> obs
        cancer
drinking   lc   nc
      hd    9  891
      nd   30 2070
> res=chisq.test(obs)
> res

Pearson's Chi-squared test with Yates' continuity correction

data:  obs
X-squared = 0.59875, df = 1, p-value = 0.4391
```

Test of independence

```
> res=chisq.test(obs,correct=F)
> res

Pearson's Chi-squared test

data:  obs
X-squared = 0.90183, df = 1, p-value = 0.3423
```

Fisher's exact test

```
> fisher.test(obs)

Fisher's Exact Test for Count Data

data:  obs
p-value = 0.3845
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2898582 1.5137429
sample estimates:
odds ratio
 0.6970281
```

Fisher's exact test: tea tasting

A British woman claimed to be able to distinguish whether milk or tea was added to the cup first. To test, she was given 8 cups of tea, in four of which milk was added first. The null hypothesis is that there is no association between the true order of pouring and the woman's guess (she can't distinguish), the alternative that there is an association.

```
# fisher's exact test
> TeaTasting  = matrix(c(3, 1, 1, 3), nrow = 2,
+          dimnames = list(Guess = c("Milk", "Tea"),
+                          Truth = c("Milk", "Tea")))
> TeaTasting
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3
```

Fisher's exact test

```
> fisher.test(TeaTasting)
            Fisher's Exact Test for Count Data

data:  obs
p-value = 0.3845
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2898582 1.5137429
sample estimates:
odds ratio
 0.6970281
```