Contingency table and count data

Patrick Breheny

November 3, 2016

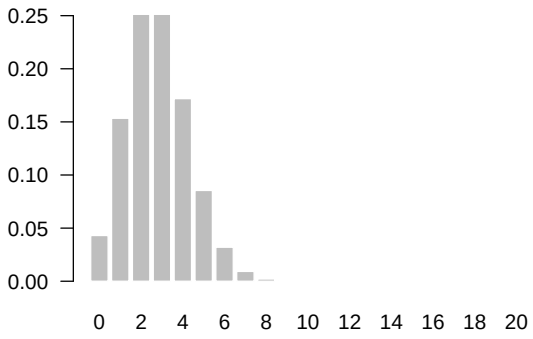
Today's lab will introduce the `R` functions for analyzing contingency tables as well as analyzing count data using the Poisson distribution.

1 Contingency tables

1.1 Fisher's exact test

The `R` function for carrying out Fisher's exact test is called `fisher.test`. To illustrate how it works, let's use one of the examples we saw in class when discussing cross-sectional studies. Below, we create a matrix `X` containing the data (here, counts of hospitalized patients according to whether they have a respiratory disease, a circulatory disease, both, or neither) in a 2×2 table, then analyze it using Fisher's exact test. First, I use the built-in `R` function, then I demonstrate where this p-value is coming from using the equations we derived in class.

```
> X <- rbind(c( 7,  29),
+            c(13, 208))
> fisher.test(X)


Fisher's Exact Test for Count Data

data:  X
p-value = 0.01174
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.193351 11.392898
sample estimates:
odds ratio
  3.833179

> ## By hand
> M <- sum(X[,1])
> N <- sum(X)
> n <- sum(X[1,])
> x <- 0:20
> Prob <- choose(M,x)*choose(N-M,n-x)/choose(N,n)
> names(Prob) <- x
> barplot(Prob, border="white", las=1)
```




```
> sum(Prob[Prob <= Prob["7"]])

[1] 0.01173709
```

Notice that `fisher.test` provides an estimate of the odds ratio along with a confidence interval. We didn't discuss this in class, but it is also possible to construct a confidence interval for the odds ratio by inverting Fisher's exact test. However, being able to test hypotheses other than indepdendence involves analysis of noncentral hypergeometric distributions, which are a bit beyond the scope of our course.

Still, it is worthwhile knowing that we can obtain these exact confidence intervals from `fisher.test`. Let's see how they compare with the approximate confidence interval that we derived in class.

```
> SE <- sqrt(sum(1/X))
> OR <- X[1,1]*X[2,2]/(X[1,2]*X[2,1])
> log(OR) + qnorm(c(0.025, 0.975))*SE      ## CI for log(OR)

[1] 0.3536013 2.3488048

> exp(log(OR) + qnorm(c(0.025, 0.975))*SE) ## CI for OR

[1]  1.424187 10.473045
```

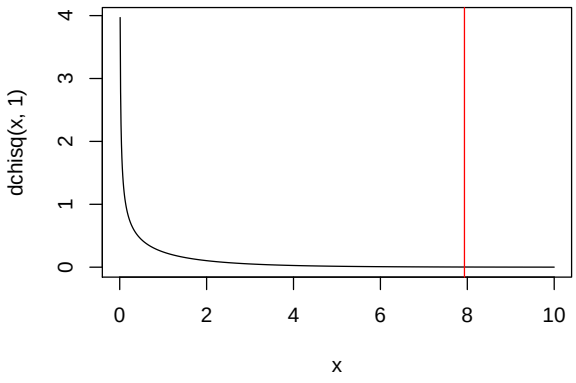
So, reasonably close, although clearly there are some differences. Differences between exact and approximate methods are to be expected here, since low counts are present in several cells. We see much better agreement, for example, in the CDC breast cancer data:

```
> Y <- rbind(c(4475, 65),
+            c(1597, 31))
> fisher.test(Y)$conf.int

[1] 0.8385228 2.0890648
attr(,"conf.level")
[1] 0.95

> SE <- sqrt(sum(1/Y))
> OR <- Y[1,1]*Y[2,2]/(Y[1,2]*Y[2,1])
> exp(log(OR) + qnorm(c(0.025, 0.975))*SE)

[1] 0.8679478 2.0576870
```

1.2 The χ^2 test

R also offers a function for carrying out the χ^2 test, called `chisq.test`. As before, I'll use the built-in R function, then demonstrate where the p-value is coming from using the equations we derived in class.

```
> chisq.test(X)

Warning in chisq.test(X): Chi-squared approximation may be incorrect


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

data:  X
X-squared = 6.1569, df = 1, p-value = 0.01309

> chisq.test(X, correct=FALSE)

Warning in chisq.test(X, correct = FALSE): Chi-squared approximation may be incorrect


Pearson's Chi-squared test

data:  X
X-squared = 7.9342, df = 1, p-value = 0.004851

> E <- outer(apply(X, 1, sum), apply(X, 2, sum))/sum(X) ## Expected cell counts
> sum((X-E)^2/E)

[1] 7.934195

> 1-pchisq(sum((X-E)^2/E), 1)

[1] 0.004850919

> x <- seq(0, 10, 0.01)
> plot(x, dchisq(x, 1), type="l")
> abline(v=sum((X-E)^2/E), col="red")
```



Two remarks here:

- First, note that we get a warning here for the χ^2 test. As you can verify by looking at `E`, one of the expected cell counts is only 2.8; `chisq.test` will warn you if any of the E_i values is below 5.
- Second, we didn't discuss the idea of a continuity correction in class. The most important thing to know in terms of practical usage is that a continuity "correction" doesn't necessarily make the test more accurate, but it always makes the test more conservative (increases the p-value). For example, although for the hospital data the continuity-corrected test was in closer agreement with Fisher's exact test than the regular χ^2 test, this is not the case for the CDC breast cancer data:

```
> fisher.test(Y)$p.value

[1] 0.1991656

> chisq.test(Y)


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

data:  Y
X-squared = 1.451, df = 1, p-value = 0.2284

> chisq.test(Y, correct=FALSE)


Pearson's Chi-squared test

data:  Y
X-squared = 1.7457, df = 1, p-value = 0.1864
```

Note that the `chisq.test` function does not provide an estimate or confidence interval for the odds ratio (and there is no option to request one).

1.3 Larger tables

Before moving on to Poisson/count data, let's take a quick look at larger tables (i.e., not 2×2 tables). For example, here are the results from the 2000 General Social Survey in terms of gender and party identification:

```
> Z <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
> dimnames(Z) <- list(gender = c("Female","Male"),
+                     party = c("Democrat","Independent", "Republican"))
> Z

        party
gender   Democrat Independent Republican
  Female      762         327        468
  Male        484         239        477
```

The χ^2 test proceeds in exactly the same way for larger tables, with the exception of a change in the degrees of freedom:

```
> chisq.test(Z, correct=FALSE)


Pearson's Chi-squared test

data:  Z
X-squared = 30.07, df = 2, p-value = 2.954e-07

> E <- outer(apply(Z, 1, sum), apply(Z, 2, sum))/sum(Z)
> 1-pchisq(sum((Z-E)^2/E), 2)

[1] 2.953589e-07
```

The syntax of `fisher.test` is the same for larger tables as it is for smaller ones. The internal calculations, however, are quite different as the distribution of counts no longer follows a simple hypergeometric distribution.

```
> fisher.test(Z)


Fisher's Exact Test for Count Data

data:  Z
p-value = 3.027e-07
alternative hypothesis: two.sided
```

Note that (a) the p-value is nearly identical to that of the χ^2 test, and (b) we no longer get an odds ratio. There is an obvious reason for this: we cannot take a ratio when we have more than two categories we are comparing. We can only discuss the odds ratios for smaller comparisons within the larger table:

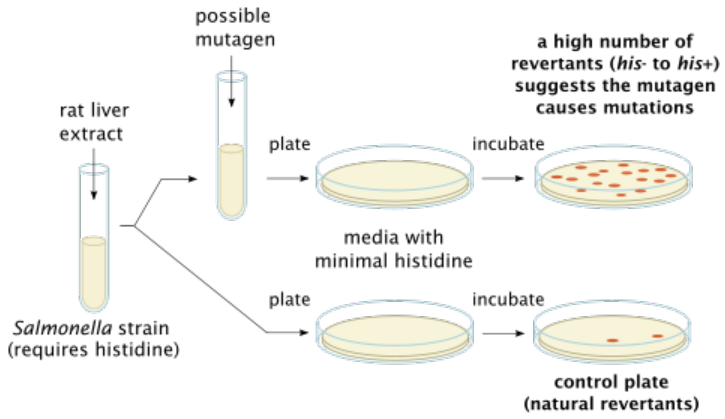
```
> fisher.test(Z[,-2])


Fisher's Exact Test for Count Data

data:  Z[, -2]
p-value = 6.806e-08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.347341 1.910944
sample estimates:
odds ratio
  1.604345

> fisher.test(Z[,-3])


Fisher's Exact Test for Count Data

data:  Z[, -3]
p-value = 0.1786
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
 0.9349959 1.4151128
sample estimates:
odds ratio
    1.1506
```

Thus, women are 60% more likely to identify as Democrats (vs. Republicans) compared to men, but there is no clear evidence from this survey that women are any more likely to identify as Independents (vs. Republicans) than men are.

2 Count data

The function in `R` for analyzing one- and two-sample count data using the Poisson distribution is called `poisson.test`. To get some added variety, let's consider a different sort of experimental setup than what we considered in class. The Poisson distribution is widely used to model plate counts in laboratory settings. One common example is the Ames test for mutagenicity:



Let's analyze data from an Ames test concerning the mutagenicity of a substance called harmine:

```
> x <- c(16, 21, 22)      ## Control
> y <- c(22, 23, 33, 37) ## 100 micrograms of harmine
> X <- sum(x)
> Y <- sum(y)
```

Here, due to the additivity of the Poisson distribution, we have $X \sim \text{Pois}(3\lambda)$ and $Y \sim \text{Pois}(4\mu)$. To analyze using `poisson.test`:

```
> poisson.test(c(Y, X), c(4, 3))


Comparison of Poisson rates

data:  c(Y, X) time base: c(4, 3)
count1 = 115, expected count1 = 99.429, p-value = 0.01749
alternative hypothesis: true rate ratio is not equal to 1
95 percent confidence interval:
```

```
 1.059044 2.036378
sample estimates:
rate ratio
  1.461864
```

Thus, the data are incompatible with the hypothesis that harmine does not affect mutation rate; this study indicates that harmine increases the natural mutation rate by 46%, with a 95% confidence interval for the rate ratio of $[1.06, 2.04]$. Note that a t-test of the same data is considerably less powerful:

```
> t.test(x, y)


Welch Two Sample t-test

data:  x and y
t = -2.1919, df = 4.2892, p-value = 0.08891
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.289645   2.122978
sample estimates:
mean of x mean of y
 19.66667  28.75000
```

Finally, if we wanted to obtain separate intervals for each rate, we can use `poisson.test` for that also:

```
> poisson.test(Y, 4)


Exact Poisson test

data:  Y time base: 4
number of events = 115, time base = 4, p-value < 2.2e-16
alternative hypothesis: true event rate is not equal to 1
95 percent confidence interval:
 23.73607 34.51002
sample estimates:
event rate
     28.75

> poisson.test(X, 3)


Exact Poisson test

data:  X time base: 3
number of events = 59, time base = 3, p-value < 2.2e-16
alternative hypothesis: true event rate is not equal to 1
95 percent confidence interval:
 14.97118 25.36857
sample estimates:
event rate
  19.66667
```

```
> ## Equivalent to
> f.u <- function(lam) ppois(X, 3*lam) - 0.025
> uniroot(f.u, c(20,30))$root

[1] 25.36859

> f.l <- function(lam) 1-ppois(X-1, 3*lam) - 0.025
> uniroot(f.l, c(1, 20))$root

[1] 14.97118
```