# Chi-Squared Tests

Dr. Syring

February 8, 2020

# Tabulated Categorical Data

A common data format consists of the counts of observations falling into different categories, such as the hair and eye color data set in R:

```
HairEyeColor
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

# Multinomial Distribution

The binomial distribution can be used to model the probability an observation falls into one of two categories. If we generalize to the case of $k > 2$ categories we obtain the multinomial distribution (section 3.1), which can be used to describe categorical data. A multinomial r.v. $X$ is a vector $X = (X_1, \ldots, X_{k-1})$ giving the counts of observations in each of $k-1$ categories. The count in the $k^{th}$ category is then determined to be $n - \sum_{i=1}^{k-1} x_i$ where $n$ is the total number of observations. A multinomial r.v has pmf

$$P(X = x) = \frac{n!}{x_1! x_2! * \cdots * x_k!} p_1^{x_1} * \cdots * p_k^{x_k}$$

where $p_i$ gives the probability of the $i^{th}$ category.

# A relation between binomial and Chi-Squared

Recall that a binomial r.v. $Y_1$ has mean $np$ and variance $np(1-p)$ . Then, by the CLT

$$\frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{D} N(0,1)$$

Therefore,

$$\frac{(Y_1 - np_1)^2}{np_1(1-p_1)} \xrightarrow{D} \chi^2(1).$$

If we define $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$ we can write the above r.v. as

$$\frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}.$$

Now, generalize the above expression to a multinomial r.v. Suppose $X = (X_1, \ldots, X_{k-1})$ is a multinomial r.v. and define $X_k = n - \sum_{i=1}^{k-1} X_i$ and $p_k = 1 - \sum_{i=1}^{k-1} p_i$ . Then, we might guess that

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \xrightarrow{D} \chi^2(k-1).$$

This is actually true!

# Short proof sketch

Write

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k-1} \frac{(X_i - np_i)^2}{np_i} + \frac{(X_k - np_k)^2}{np_k}$$

$$= \sum_{i=1}^{k-1} \frac{(X_i - np_i)^2}{np_i} + \frac{(\sum_{i=1}^{k-1} X_i - np_i)^2}{np_k}$$

becuase $X_k = n - \sum_{i=1}^{k-1} X_i$ and $np_k = n - \sum_{i=1}^{k-1} np_i$ .

Then, confirm that the last expression above can be written

$$(X - np)^\top \Sigma^{-1} (X - np)$$

where $X$ is the column vector of $X_1, \ldots, X_{k-1}$ and $p$ is the column vector of $p_1, \ldots, p_{k-1}$ and $\Sigma$ is the matrix $n * [diag(p) - pp^\top]$ .

Since $(X - np)^\top \Sigma^{-1/2}$ is approximately standard normal, the quadratic form above is approximately chi-squared with degrees of freedom $k - 1$ .

# Testing for a specific Multinomial distribution

We can use the Chi-squared random variable

$$\sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}}$$

to test the null hypothesis

$$H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots, p_{k-1} = p_{0,k-1}$$

for a chosen vector $p_0 = (p_{01}, \ldots, p_{0,k-1})$ . The alternative hypothesis is simply that at least one of these category proportions $p_i$ is not $p_{0i}$ .

# Example: testing Mendel's theory of inheritance

The biologist Gregory Mendel hypothesized that yellow pea plants crossed with green pea plants would produce $75\%$ yellow and $25\%$ green child plants. Of $n = 8023$ hybrid seeds $2001$ grew into green plants and $6022$ grew into yellow pea plants.

$H_0 : p_1 = .25,\ p_2 = 0.75$ . The test statistic is

$$\frac{(2001 - 0.25 * 8023)^2}{0.25 * 8023} + \frac{(6022 - 0.75 * 8023)^2}{0.75 * 8023} = 0.015$$

If we test at $\alpha = 0.05$ then the $\chi^2(1)\ 95^{th}$ quantile is 3.84 so we do not reject $H_0$ .

# Testing equivalence of two Multinomial distributions

Suppose we have tabulated data like the hair and eye color data set in R that we model with a multinomial distribution.

```
HairEyeColor
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

There are two tables here, one for males and one for females. How could we test the null hypothesis that the distributions of hair and eye color are the same for males and females?

$$H_0 : p_{1i} = p_{2i}, \text{ for all } i$$

where $p_{1i}$ and $p_{2i}$ are the category $i$ probabilities for males and females.

The point estimate for each $p_{1i} = p_{2i} := p_i$ is the combined sample proportion $\frac{X_{1i}+X_{2i}}{n_1+n_2}$ . And, the test statistic

$$\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{\left(X_{ji}-n_j\left[\frac{X_{1i}+X_{2i}}{n_1+n_2}\right]\right)^2}{n_j\left[\frac{X_{1i}+X_{2i}}{n_1+n_2}\right]}$$

is approximately $\chi^2(k-1)$ . Why $k-$ df? There are $2k-2$ parameters, but under the null the distributions are equal so there are only $k-1$ "free" parameters.

Example computation for hair and eye color data:

```
df <- as.data.frame(HairEyeColor)
p.hat <- (df[1:16,4]+df[17:32,4])/sum(df[,4])
n.m <- sum(df[1:16,4])
n.f <- sum(df[17:32,4])
chi.sq.test.stat <- sum(((df[1:16,4]-n.m*p.hat[1:16])^2)/n.m*p.hat[1:16])+sum(((df[17:32,4]-n.f*p.
chi.sq.test.stat
```

```
## [1] 0.3425414
```

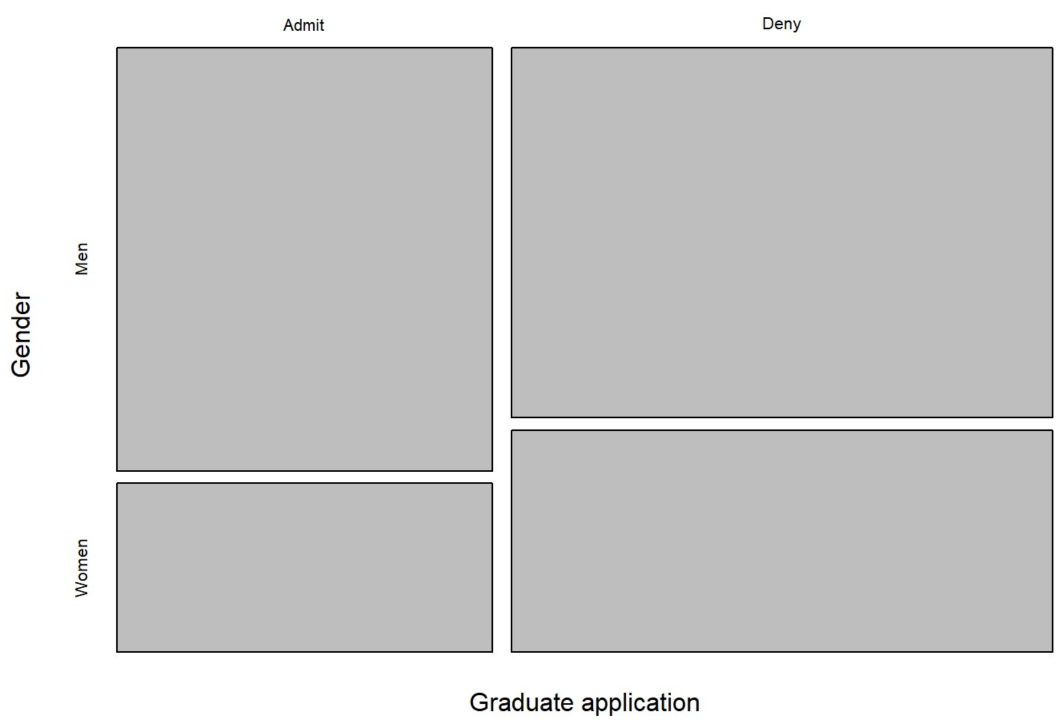
```
qchisq(.95,15)
```

```
## [1] 24.99579
```

# Chi-square tests of independence

A $2 \times k$ "contingency table" has two variables that can take on $2 \times k$ values and records the number of observations in each combination. For example, a $2 \times 2$ table is

```
UCB<-matrix(c(3738,4704,1494,2827),2,2,byrow=T)
rownames(UCB)<-c("Men","Women")
colnames(UCB)<-c("Admit","Deny")
mosaicplot(t(UCB),ylab="Gender",xlab="Graduate application", main="")
```




We may be interested in whether or not the chance of admission depends on gender. Let $p$ denote the probability of admission, and $p_M$ , $p_F$ denote the probability of admission for a Male and a Female applicant. Then, we want to test if $H_0 : p_M = p_F = p$ . If $p_{ij}$ denotes the $(i, j)$ cell probability in the table, then independence means $p_i j = p_i \cdot *p_{.j}$ where $p_{i\cdot} = p_{i1} + p_{i2}$ and $p_{.j} = p_{1j} + p_{2j}$ . Therefore, the test statistic is

$$\chi^2 = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} (X_{ij} - n_j[(X_{i1} + X_{i2})/(n_1 + n_2)])^2}{n_j[(X_{i1} + X_{i2})/(n_1 + n_2)]}$$

where $n_1$ and $n_2$ denote the number of males and females.

```
UCB<-matrix(c(3738,4704,1494,2827),2,2,byrow=T)
rownames(UCB)<-c("Men","Women")
colnames(UCB)<-c("Admit","Deny")
UCB
```

```
##       Admit Deny
## Men    3738 4704
## Women  1494 2827
```

```
p<-apply(UCB,1,sum)/sum(UCB)
q<-apply(UCB,2,sum)/sum(UCB)
p
```

```
##       Men     Women
## 0.6614432 0.3385568
```

```
q
```

```
##    Admit     Deny
## 0.409935 0.590065
```

```
expected<-outer(p,q,FUN="*")
expected*sum(UCB)
```

```
##            Admit     Deny
## Men   3460.671 4981.329
## Women 1771.329 2549.671
```

```
((UCB - expected*sum(UCB))^2)/(expected*sum(UCB))
```

```
##            Admit     Deny
## Men   22.22441 15.43993
## Women 43.42015 30.16521
```

```
sum(((UCB - expected*sum(UCB))^2)/(expected*sum(UCB)))
```

```
## [1] 111.2497
```

```
X2<-chisq.test(UCB)
X2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCB
## X-squared = 110.85, df = 1, p-value < 2.2e-16
```

The degrees of freedom are the number of free parameters minus the number of estimated parameters. We have a multinomial distribution with 4 categories, so there are 3 free parameters. We have estimated 2 parameters $\hat{p}_{i\cdot}$ and $\hat{p}_{\cdot j}$ so there are 3-2 = 1 df.

```
UCBAdmissions
```

```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
```

```
##            Gender
## Admit       Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##            Gender
## Admit       Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##            Gender
## Admit       Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##            Gender
## Admit       Male Female
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##            Gender
## Admit       Male Female
##   Admitted   22     24
##   Rejected  351    317
```

```
chisq.test(UCBAdmissions[,,1])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 1]
## X-squared = 16.372, df = 1, p-value = 5.205e-05
```

```
chisq.test(UCBAdmissions[,,2])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 2]
## X-squared = 0.085098, df = 1, p-value = 0.7705
```

```r
chisq.test(UCBAdmissions[,,3])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 3]
## X-squared = 0.63322, df = 1, p-value = 0.4262
```

```r
chisq.test(UCBAdmissions[,,4])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 4]
## X-squared = 0.22159, df = 1, p-value = 0.6378
```

```r
chisq.test(UCBAdmissions[,,5])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 5]
## X-squared = 0.80805, df = 1, p-value = 0.3687
```

```r
chisq.test(UCBAdmissions[,,6])
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCBAdmissions[, , 6]
## X-squared = 0.21824, df = 1, p-value = 0.6404
```