

# Contingency tables

The ASTA team

## Contents

<b>1</b>	<b>Contingency tables</b>	<b>1</b>
1.1	A contingency table . . . . .	1
<b>2</b>	<b>Independence</b>	<b>2</b>
2.1	Independence . . . . .	2
2.2	The Chi-squared test for independence . . . . .	2
2.3	Calculation of expected table . . . . .	3
2.4	Chi-squared ( $\chi^2$ ) test statistic . . . . .	3
2.5	$\chi^2$ -test template. . . . .	4
2.6	The function <code>chisq.test</code> . . . . .	4
<b>3</b>	<b>The <math>\chi^2</math>-distribution</b>	<b>5</b>
3.1	The $\chi^2$ -distribution . . . . .	5
<b>4</b>	<b>Agresti - Summary</b>	<b>6</b>
4.1	Summary . . . . .	6
<b>5</b>	<b>Standardized residuals</b>	<b>6</b>
5.1	Residual analysis . . . . .	6
5.2	Residual analysis in <b>R</b> . . . . .	7
5.3	Why not just use two-way ANOVA ? . . . . .	7
<b>6</b>	<b>Models for table data in <b>R</b></b>	<b>7</b>
6.1	Example . . . . .	7
6.2	Model specification . . . . .	8
6.3	Model specification in <b>R</b> . . . . .	8
6.4	Expected values and standardized residuals . . . . .	10

## 1 Contingency tables

### 1.1 A contingency table

- We return to the dataset `popularKids`, where we study **association** between **2 factors**: `Goals` and `Urban.Rural`.
- Based on a sample we make a cross tabulation of the factors and we get a so-called **contingency table** (`krydstabel`).

```
popKids <- read.delim("https://asta.math.aau.dk/datasets?file=PopularKids.dat")
library(mosaic)
tab <- tally(~Urban.Rural + Goals, data = popKids, margins = TRUE)
tab
```

```
##           Goals
```

```
## Urban.Rural Grades Popular Sports Total
##   Rural      57      50      42     149
##   Suburban   87      42      22     151
##   Urban     103      49      26     178
##   Total     247     141      90     478
```

### 1.1.1 A conditional distribution

- Another representation of data is the percent-wise distribution of `Goals` for each level of `Urban.Rural`, i.e. the sum in each row of the table is 100 (up to rounding):

```
tab <- tally(~Urban.Rural + Goals, data = popKids)
addmargins(round(100 * prop.table(tab, 1)),margin = 2)
```

```
##           Goals
## Urban.Rural Grades Popular Sports Sum
##   Rural      38      34      28 100
##   Suburban   58      28      15 101
##   Urban     58      28      15 101
```

- Here we will talk about the **conditional distribution** of `Goals` given `Urban.Rural`.
- An important question could be:
  - Are the goals of the kids different when they come from urban, suburban or rural areas? I.e. are the rows in the table significantly different?
- There is (almost) no difference between urban and suburban, but it looks like rural is different.

## 2 Independence

### 2.1 Independence

- Recall, that two factors are **independent**, when there is no difference between the population's distributions of one factor given the levels of the other factor.
- Otherwise the factors are said to be **dependent**.
- If we e.g. have the following conditional **population distributions** of `Goals` given `Urban.Rural`:

```
##           Goals
## Urban.Rural Grades Popular Sports
##   Rural      500      300      200
##   Suburban   500      300      200
##   Urban     500      300      200
```

- Then the factors `Goals` and `Urban.Rural` are independent.
- We take a sample and “measure” the factors  $F_1$  and  $F_2$ . E.g. `Goals` and `Urban.Rural` for a random child.
- The hypothesis of interest today is:

$$H_0 : F_1 \text{ and } F_2 \text{ are independent, } H_a : F_1 \text{ and } F_2 \text{ are dependent.}$$

### 2.2 The Chi-squared test for independence

- The relative frequencies in the sample gives an estimate of the unconditional distribution of `Goals`:

```
n <- margin.table(tab)
pctGoals <- round(100 * margin.table(tab, 2)/n, 1)
pctGoals
```

```
## Goals
```

```
## Grades Popular Sports
## 51.7 29.5 18.8
```

- If we assume independence, then this is also a guess of the conditional distributions of Goals given Urban.Rural.
- The corresponding expected counts in the sample are then:

```
##           Goals
## Urban.Rural Grades      Popular      Sports      Sum
## Rural      77.0 (51.7%)  44.0 (29.5%)  28.1 (18.8%) 149.0 (100%)
## Suburban   78.0 (51.7%)  44.5 (29.5%)  28.4 (18.8%) 151.0 (100%)
## Urban      92.0 (51.7%)  52.5 (29.5%)  33.5 (18.8%) 178.0 (100%)
## Sum        247.0 (51.7%) 141.0 (29.5%)  90.0 (18.8%) 478.0 (100%)
```

## 2.3 Calculation of expected table

pctexptab

```
##           Goals
## Urban.Rural Grades      Popular      Sports      Sum
## Rural      77.0 (51.7%)  44.0 (29.5%)  28.1 (18.8%) 149.0 (100%)
## Suburban   78.0 (51.7%)  44.5 (29.5%)  28.4 (18.8%) 151.0 (100%)
## Urban      92.0 (51.7%)  52.5 (29.5%)  33.5 (18.8%) 178.0 (100%)
## Sum        247.0 (51.7%) 141.0 (29.5%)  90.0 (18.8%) 478.0 (100%)
```

- We note that
  - The relative frequency for a given column is columnTotal divided by tableTotal. For example Grades, which is  $\frac{247}{478} = 51.7\%$ .
  - The expected value in a given cell in the table is then the cell's relative column frequency multiplied by the cell's rowTotal. For example Rural and Grades:  $149 \times 51.7\% = 77.0$ .
- This can be summarized to:
  - The expected value in a cell is the product of the cell's rowTotal and columnTotal divided by tableTotal.

## 2.4 Chi-squared ( $\chi^2$ ) test statistic

- We have an **observed table**:

tab

```
##           Goals
## Urban.Rural Grades Popular Sports
## Rural      57      50      42
## Suburban   87      42      22
## Urban     103      49      26
```

- And an **expected table**, if  $H_0$  is true:

```
##           Goals
## Urban.Rural Grades Popular Sports Sum
## Rural      77.0  44.0  28.1 149.0
## Suburban   78.0  44.5  28.4 151.0
## Urban      92.0  52.5  33.5 178.0
## Sum        247.0 141.0  90.0 478.0
```

- If these tables are “far from each other”, then we reject  $H_0$ . We want to measure the distance via the Chi-squared test statistic:
  - $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ : Sum over all cells in the table

- $f_o$  is the frequency in a cell in the observed table
- $f_e$  is the corresponding frequency in the expected table.
- We have:

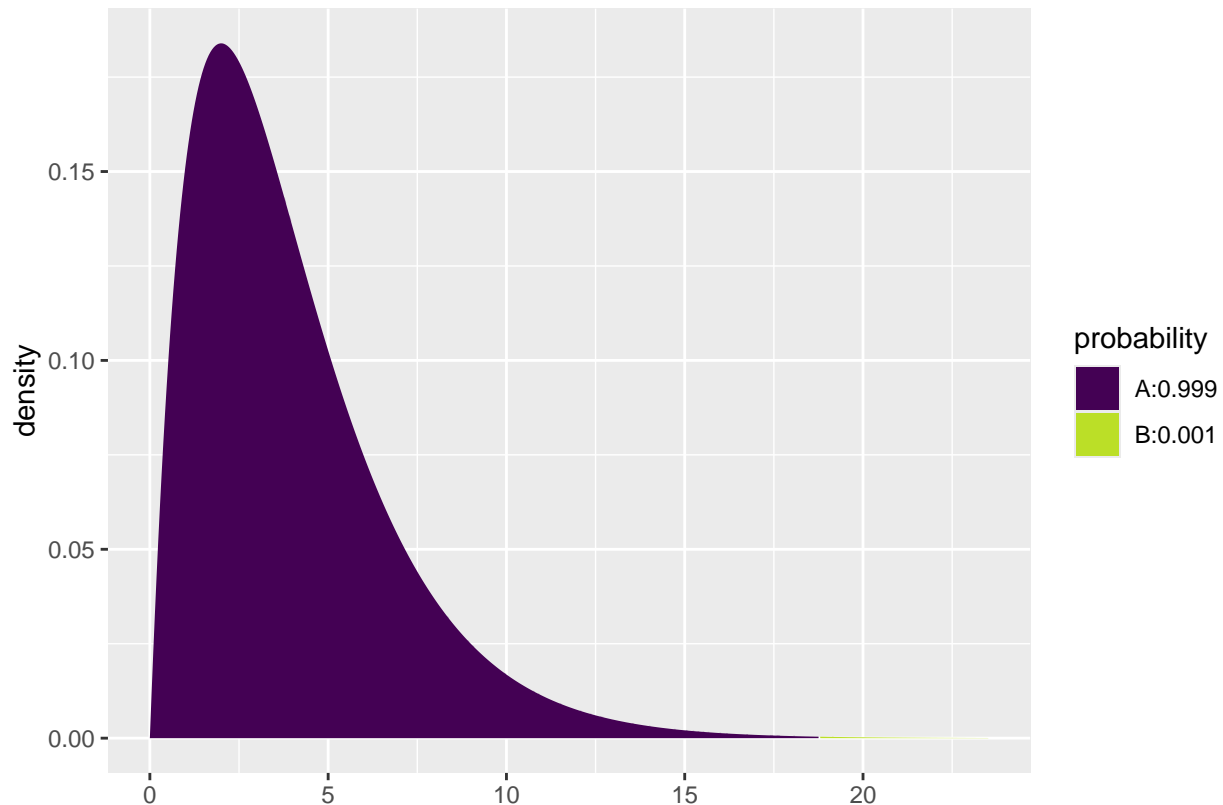
$$X_{obs}^2 = \frac{(57 - 77)^2}{77} + \dots + \frac{(26 - 33.5)^2}{33.5} = 18.8$$

- Is this a large distance??

## 2.5 $\chi^2$ -test template.

- We want to test the hypothesis  $H_0$  of independence in a table with  $r$  rows and  $c$  columns:
  - We take a sample and calculate  $X_{obs}^2$  - the observed value of the test statistic.
  - p-value: Assume  $H_0$  is true. What is then the chance of obtaining a larger  $X^2$  than  $X_{obs}^2$ , if we repeat the experiment?
- This can be approximated by the  $\chi^2$ -**distribution** with  $df = (r - 1)(c - 1)$  degrees of freedom.
- For Goals and Urban.Rural we have  $r = c = 3$ , i.e.  $df = 4$  and  $X_{obs}^2 = 18.8$ , so the p-value is:

```
1 - pdist("chisq", 18.8, df = 4)
```



```
## [1] 0.00086
```

- There is clearly a significant association between Goals and Urban.Rural.

## 2.6 The function `chisq.test`.

- All of the above calculations can be obtained by the function `chisq.test`.

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat
```

```
##
```

```
## Pearson's Chi-squared test
##
## data:  tab
## X-squared = 19, df = 4, p-value = 8e-04
```

```
testStat$expected
```

```
##           Goals
## Urban.Rural Grades Popular Sports
##   Rural      77  44.0  28.1
##   Suburban   78  44.5  28.4
##   Urban      92  52.5  33.5
```

- 
- The frequency data can also be put directly into a matrix.

```
data <- c(57, 87, 103, 50, 42, 49, 42, 22, 26)
tab <- matrix(data, nrow = 3, ncol = 3)
row.names(tab) <- c("Rural", "Suburban", "Urban")
colnames(tab) <- c("Grades", "Popular", "Sports")
tab
```

```
##           Grades Popular Sports
## Rural      57      50      42
## Suburban   87      42      22
## Urban     103      49      26
```

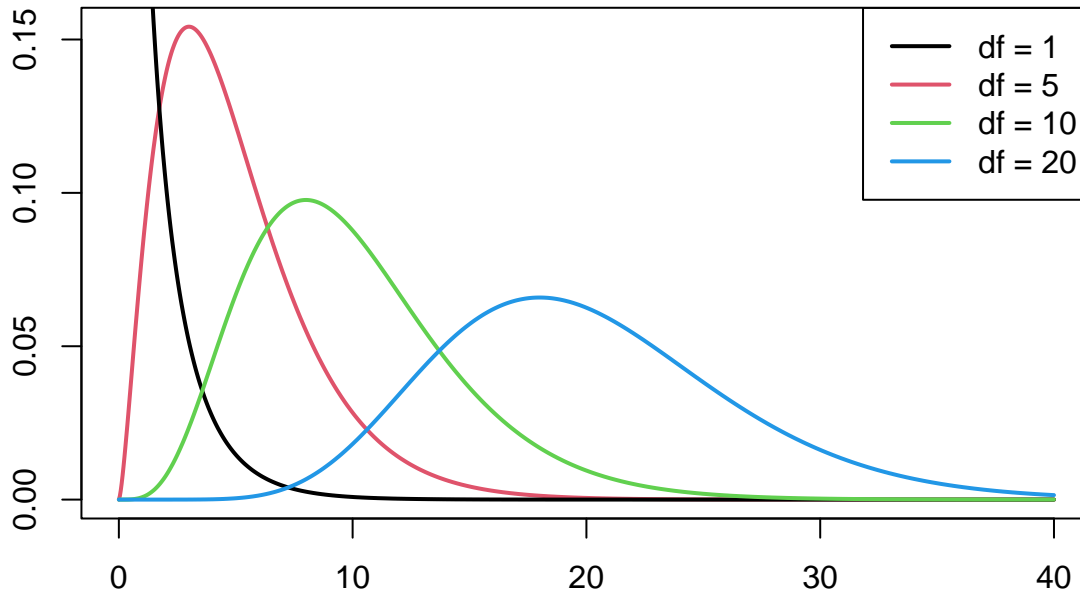
```
chisq.test(tab)
```

```
##
## Pearson's Chi-squared test
##
## data:  tab
## X-squared = 19, df = 4, p-value = 8e-04
```

### 3 The $\chi^2$ -distribution

#### 3.1 The $\chi^2$ -distribution

- The  $\chi^2$ -distribution with  $df$  degrees of freedom:
  - Is never negative.
  - Has mean  $\mu = df$
  - Has standard deviation  $\sigma = \sqrt{2df}$
  - Is skewed to the right, but approaches a normal distribution when  $df$  grows.



## 4 Agresti - Summary

### 4.1 Summary

- For the the Chi-squared statistic,  $X^2$ , to be appropriate we require that the expected values have to be  $f_e \geq 5$ .
- Now we can summarize the ingredients in the Chi-squared test for independence.

**TABLE 8.5: The Five Parts of the Chi-Squared Test of Independence**

- 
1. Assumptions: Two categorical variables, random sampling,  $f_e \geq 5$  in all cells
  2. Hypotheses:  $H_0$ : Statistical independence of variables  
 $H_a$ : Statistical dependence of variables
  3. Test statistic:  $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ , where  $f_e = \frac{(\text{Row total})(\text{Column total})}{\text{Total sample size}}$
  4.  $P$ -value:  $P =$  right-tail probability above observed  $\chi^2$  value,  
for chi-squared distribution with  $df = (r - 1)(c - 1)$
  5. Conclusion: Report  $P$ -value  
If decision needed, reject  $H_0$  at  $\alpha$ -level if  $P \leq \alpha$
- 

## 5 Standardized residuals

### 5.1 Residual analysis

- If we reject the hypothesis of independence it can be of interest to identify the significant deviations.
- In a given cell in the table,  $f_o - f_e$  is the deviation between data and the expected values under the null hypothesis.
- We assume that  $f_e \geq 5$ .
- If  $H_0$  is true, then the standard error of  $f_o - f_e$  is given by

$$se = \sqrt{f_e(1 - \text{rowProportion})(1 - \text{columnProportion})}$$

- The corresponding  $z$ -score

$$z = \frac{f_o - f_e}{se}$$

should in 95% of the cells be between  $\pm 2$ . Values above 3 or below -3 should not appear.

- In popKids table cell **Rural** and **Grade** we got  $f_e = 77.0$  and  $f_o = 57$ . Here  $\text{columnProportion} = 51.7\%$  and  $\text{rowProportion} = 149/478 = 31.2\%$ .

- We can then calculate

$$z = \frac{57 - 77}{\sqrt{77(1 - 0.517)(1 - 0.312)}} = -3.95$$

- Compared to the null hypothesis there are way too few rural kids who find grades important.
- In summary: The standardized residuals allow for cell-by-cell ( $f_e$  vs  $f_o$ ) comparison.

## 5.2 Residual analysis in R

- In R we can extract the standardized residuals from the output of `chisq.test`:

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat$stdres
```

```
##           Goals
## Urban.Rural Grades Popular Sports
##   Rural    -3.951   1.310  3.523
##   Suburban  1.767  -0.548 -1.619
##   Urban     2.087  -0.727 -1.819
```

## 5.3 Why not just use two-way ANOVA ?

- number of persons in different categories are *not* normally distributed
- variance typically larger the larger expected frequency
- underlying data are discrete (for each person, which column and row category does person belong to)
- these discrete variables are naturally modelled in terms of probabilities for different categories
- therefore hypothesis of independence becomes natural null hypothesis
- it is possible to model table frequencies as dependent variable using a regression model but then we need the framework of *generalized linear models* (see last slides)

Contingency table:

- *counts* of how many individuals fall within different categories for two (or more) categorical variables

Two-way ANOVA:

- a number of individuals/objects/... available for each combination of two categorical variables
- next a continuous variable is measured for each individual or object (this becomes the response variable)

# 6 Models for table data in R

## 6.1 Example

- We will study the dataset `HairEyeColor`.

```
HairEyeColor <- read.delim("https://asta.math.aau.dk/datasets?file=HairEyeColor.txt")
head(HairEyeColor)
```

```
##   Hair  Eye  Sex Freq
## 1 Black Brown Male   32
## 2 Brown Brown Male   53
```

```
## 3 Red Brown Male 10
## 4 Blond Brown Male 3
## 5 Black Blue Male 11
## 6 Brown Blue Male 50
```

- Data is organized such that the variable `Freq` gives the frequency of each combination of the factors `Hair`, `Eye` and `Sex`.
- For example: 32 observations are men with black hair and brown eyes.
- We are interested in the association between eye color and hair color ignoring the sex
- We aggregate data, so we have a table with frequencies for each combination of `Hair` and `Eye`.

```
HairEye <- aggregate(Freq ~ Eye + Hair, FUN = sum, data = HairEyeColor)
HairEye
```

```
##      Eye Hair Freq
## 1 Blue Black 20
## 2 Brown Black 68
## 3 Green Black 5
## 4 Hazel Black 15
## 5 Blue Blond 94
## 6 Brown Blond 7
## 7 Green Blond 16
## 8 Hazel Blond 10
## 9 Blue Brown 84
## 10 Brown Brown 119
## 11 Green Brown 29
## 12 Hazel Brown 54
## 13 Blue Red 17
## 14 Brown Red 26
## 15 Green Red 14
## 16 Hazel Red 14
```

## 6.2 Model specification

- We can write down a model for (the logarithm of) the expected frequencies by using dummy variables  $z_{e1}, z_{e2}, z_{e3}$  and  $z_{h1}, z_{h2}, z_{h3}$
- To denote the different levels of `Eye` and `Hair` (the reference level has all dummy variables equal to 0):

$$\log(f_e) = \alpha + \beta_{e1}z_{e1} + \beta_{e2}z_{e2} + \beta_{e3}z_{e3} + \beta_{h1}z_{h1} + \beta_{h2}z_{h2} + \beta_{h3}z_{h3}.$$

- Note that we haven't included an interaction term, which in this case implies, that we assume independence between `Eye` and `Hair` in the model.
- Since our response variable now is a count it is no longer a linear model (`lm`) as we have been used to (linear regression).
- Instead it is a so-called generalized linear model and the relevant R command is `glm`.

## 6.3 Model specification in R

```
model <- glm(Freq ~ Hair + Eye, family = poisson, data = HairEye)
```

- The argument `family = poisson` ensures that R knows that data should be interpreted as discrete counts and not a continuous variable.

```
summary(model)
```

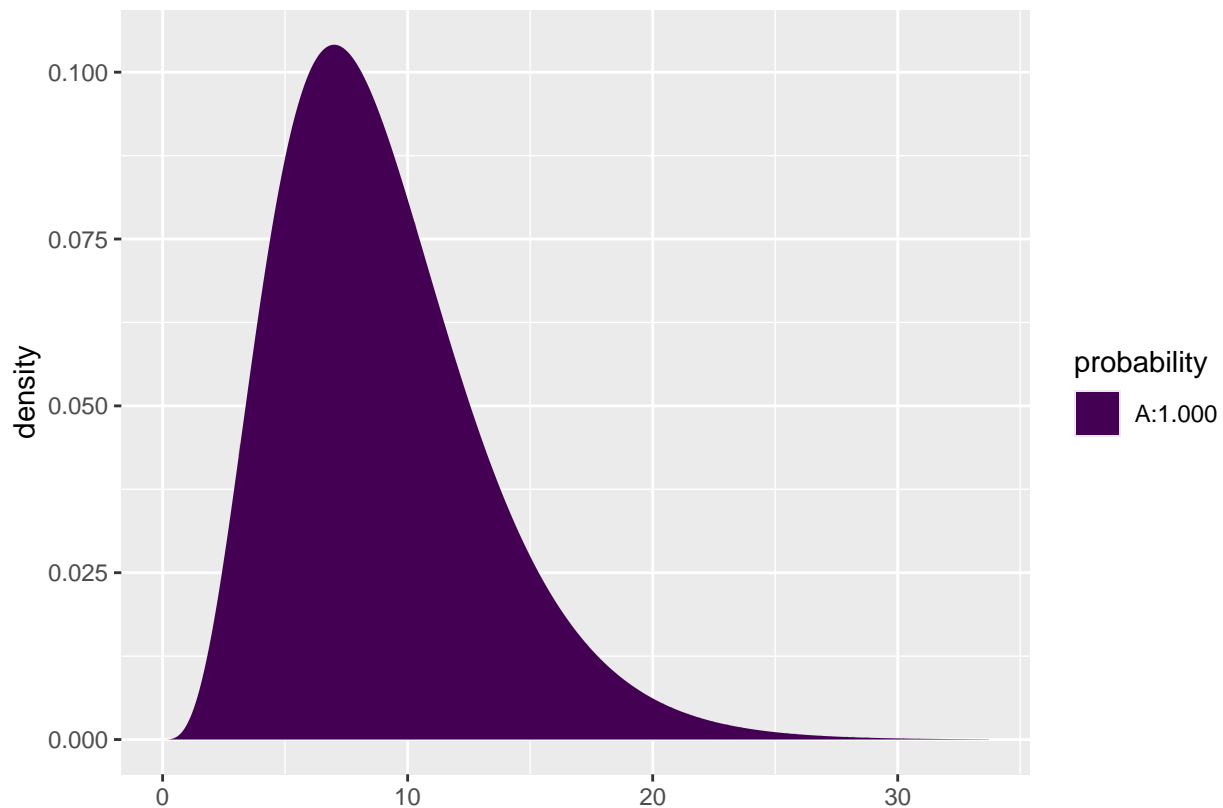
```
##
## Call:
```



```
## glm(formula = Freq ~ Hair + Eye, family = poisson, data = HairEye)
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.6693    0.1105  33.19 < 2e-16 ***
## HairBlond    0.1621    0.1309   1.24  0.216
## HairBrown    0.9739    0.1129   8.62 < 2e-16 ***
## HairRed     -0.4195    0.1528  -2.75  0.006 **
## EyeBrown     0.0230    0.0959   0.24  0.811
## EyeGreen    -1.2118    0.1424  -8.51 < 2e-16 ***
## EyeHazel    -0.8380    0.1241  -6.75  1.5e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 453.31 on 15 degrees of freedom
## Residual deviance: 146.44 on 9 degrees of freedom
## AIC: 241
##
## Number of Fisher Scoring iterations: 5
```

- A value of  $X^2 = 146.44$  with  $df = 9$  shows that there is very clear significance and we reject the null hypothesis of independence between hair and eye color.

```
1 - pdist("chisq", 146.44, df = 9)
```



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

## 6.4 Expected values and standardized residuals

- We also want to look at expected values and standardized (studentized) residuals.
- The null hypothesis predicts  $e^{3.67+0.02} = 40.1$  with brown eyes and black hair, but we have observed 68.
- This is significantly too many, since the standardized residual is 5.86.
- The null hypothesis predicts 47.2 with brown eyes and blond hair, but we have seen 7. This is significantly too few, since the standardized residual is -9.42.

```
HairEye$fitted <- fitted(model)
HairEye$resid <- rstudent(model)
HairEye
```

```
##      Eye Hair Freq fitted  resid
## 1  Blue Black   20  39.22 -4.492
## 2  Brown Black   68  40.14  5.856
## 3  Green Black    5  11.68 -2.508
## 4  Hazel Black   15  16.97 -0.583
## 5   Blue Blond   94  46.12  9.368
## 6  Brown Blond    7  47.20 -9.423
## 7  Green Blond   16  13.73  0.719
## 8  Hazel Blond   10  19.95 -2.936
## 9   Blue Brown   84 103.87 -3.437
## 10 Brown Brown  119 106.28  2.151
## 11 Green Brown   29  30.92 -0.511
## 12 Hazel Brown   54  44.93  2.023
## 13 Blue   Red   17  25.79 -2.399
## 14 Brown  Red   26  26.39 -0.101
## 15 Green  Red   14   7.68  2.368
## 16 Hazel  Red   14  11.15  0.961
```