Practical: Binary Outcomes

Alice Kamau
KEMRI-Wellcome Trust Research Programme

18 September 2019

We will use R to analyse categorical data in this practical session. We will compute different statistical tests of association from cross tabulations based on the birth weight data.

Part 1: Tests about proportions

Load the *birthweight2.csv* dataset into memory.

```r
setwd("/Users/akamau/Work/OneDrive - Kemri Wellcome Trust/Stats forum/Stat training")
bw.data <- read.csv("Data/birthweight2.csv", header=TRUE)
```

Load the R package 'descr'. This package is useful for descriptive statistics.

```r
if(!require(descr)) install.packages("descr"); library(descr)

descr(bw.data)
```

Test for a single proportion

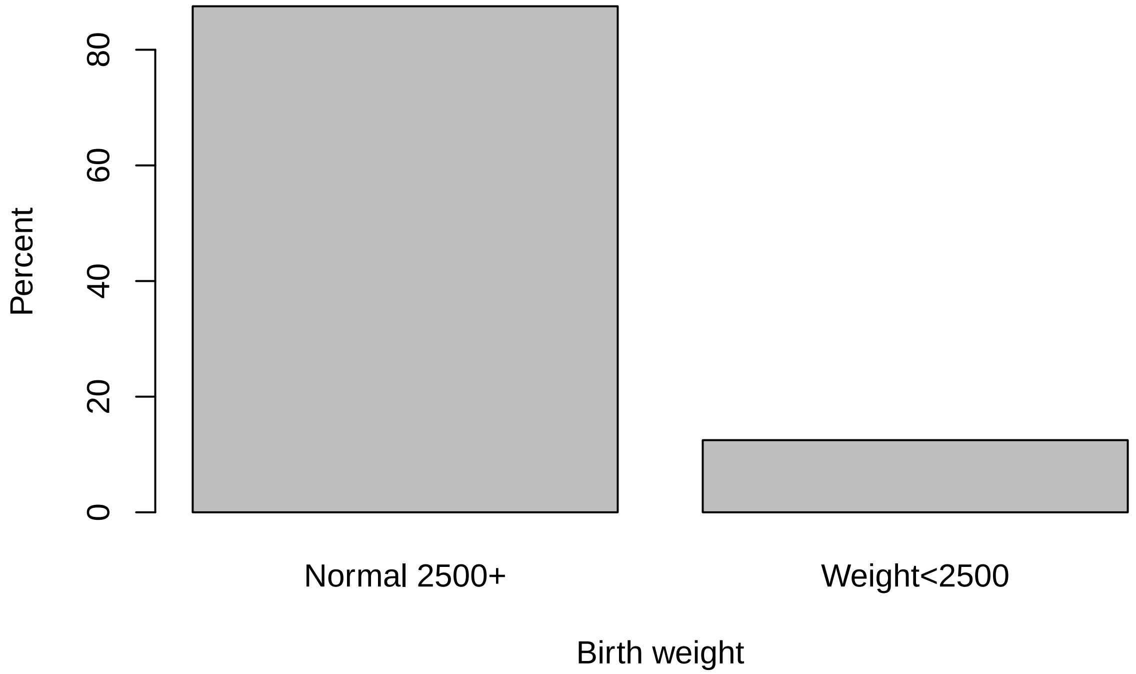
Consider the following hypothesis refering to the population from which the birth weight data was collected:

H_0 : The prevalence of normal birth weight in the population $= 40\%$

H_1 : The prevalence of normal birth weight in the population is $\neq 40\%$

To test the above H_0 in R proceed as follows:

```r
# Frequency tabulation of the birth weight variable
freq(bw.data$lbw,y.axis="percent",ylab="Percent",xlab="Birth weight")
```



```
## bw.data$lbw
##               Frequency Percent
## Normal 2500+        561   87.52
## Weight<2500          80   12.48
## Total               641  100.00
```

```
# Test of the hypothesis
prop.test(table(bw.data$lbw), p=0.4)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  table(bw.data$lbw), null probability 0.4
## X-squared = 601.12, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.4
## 95 percent confidence interval:
##  0.8465105 0.8992740
## sample estimates:
##       p
## 0.875195
```

1. Interpret the above output and state the decision you will take regarding the null hypothesis

2. Notice that in the `prop.test` output there is information that there was *continuity correction*. The function `prop.test` makes the correction by default. What do you think is the relevance of this correction?

In probability theory, a continuity correction is an adjustment that is made when a discrete distribution is approximated by a continuous distribution. which is as simple as adding or subtracting 0.5 to the discrete x-value

chi-square distribution - with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

The effect of this correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5

Now look at the help file for the function `prop.test` using

```
?prop.test
```

From what you learn about the function, edit the syntax of the above hypothesis test so that you test the following hypothesis:

H_0 : The prevalence of normal birth weight in the population $\geq 40\%$

H_1 : The prevalence of normal birth weight in the population is $< 40\%$

prop.test(table(bw.data$lbw), p=0.4, alternative = "less")

We do not have sufficient evidence to state that in the prevalence of normal birth weight in the population is not >=40%

3. Interpret your output and state the decision you will take regarding the null hypothesis

Test for difference in proportions

In the above example we tested the prevalence of normal birth weight in the population given a hypothesised value of 40%. In some cases the the question might involve comparing a proportion between two (or more) study groups.

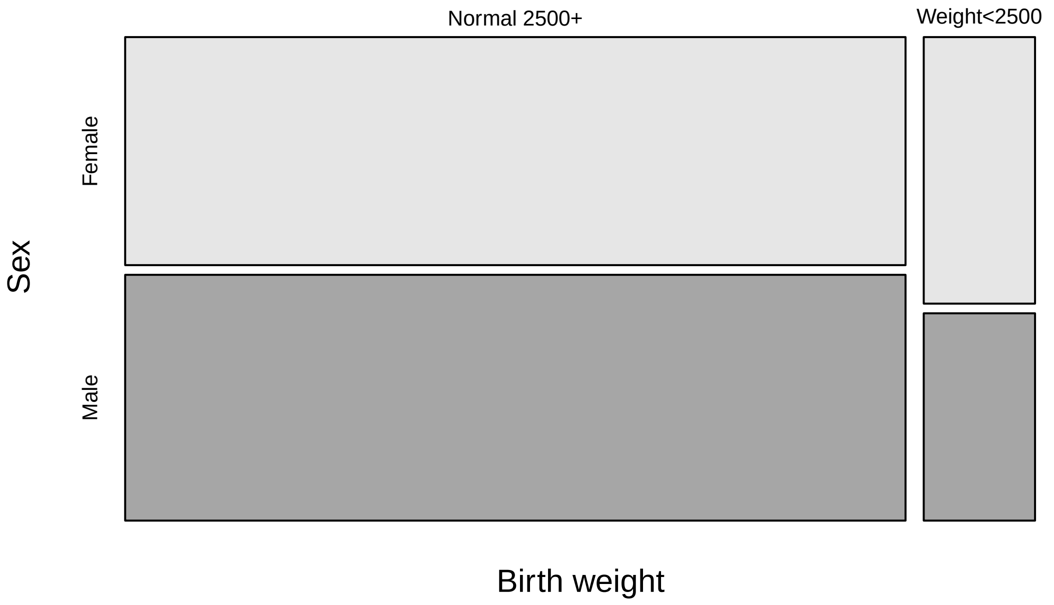
Consider the following hypothesis:

H_0 : The prevalence of normal birth weight is equal between male and female i.e $p_1 = p_2$

H_1 : The prevalence of normal birth weight is **not** equal between male and female i.e $p_1 \neq p_2$

4. The R code and output testing this hypothesis is provided. From the results, what is your conclusion?

```
# crosstab() is from the 'descr' package
crosstab(bw.data$sex,bw.data$lbw, prop.r=T, ylab="Sex",xlab="Birth weight")
```



```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## =================================================
##                bw.data$lbw
## bw.data$sex    Normal 2500+   Weight<2500   Total
## -------------------------------------------------
## Female                  270            45     315
##                       85.7%         14.3%   49.1%
## -------------------------------------------------
## Male                    291            35     326
##                       89.3%         10.7%   50.9%
## -------------------------------------------------
## Total                   561            80     641
## =================================================
```

```
prop.test(table(bw.data$sex,bw.data$lbw))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(bw.data$sex, bw.data$lbw)
## X-squared = 1.5372, df = 1, p-value = 0.215
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08982722  0.01883686
## sample estimates:
##    prop 1    prop 2
## 0.8571429 0.8926380
```

prop.r = row proportion to be inluded

Part 2: Contigency tables

Chi-squared test

1. Have a look at the help file for the chisq.test() function to familiarise yourself with what it expects as arguments.

```
?chisq.test
```

Consider Young Lives Study in Andhra Pradesh, India, that assessed the association between maternal common mental disorders (CMD) and infant stunting. The number of infants with stunted growth were 312 and 200 among mothers with and without CMD out of a total of 1311 and 557 respectively.

A Chi-square test can be performed as follows:

```
# create a contigency table of counts
# notice the substraction from the total to arrive at the numbers 999 and 357
table <- cbind(c(312,999), c(200,357))
table
```

```
##      [,1] [,2]
## [1,]  312  200
## [2,]  999  357
```

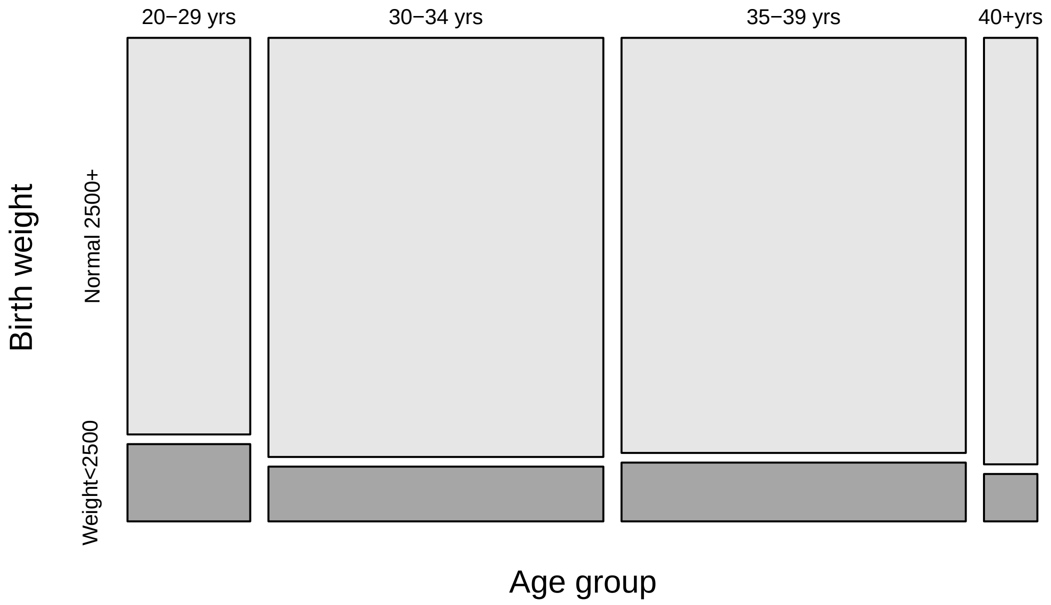
```r
# Run the Chi-square test of independence
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 28.199, df = 1, p-value = 1.095e-07
```

2. Provide an interpretation of the output, what is your conclusion?

3. Using the birthweight dataset test the association between low birth weight and maternal age group. Report your conclusion.

```r
crosstab(bw.data$lbw,bw.data$agegrp, prop.c=T, xlab="Age group",ylab="Birth weight")
```



```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |          Column Percent |
## |-------------------------|
##
## ==================================================================
##                  bw.data$agegrp
## bw.data$lbw      20-29 yrs   30-34 yrs   35-39 yrs   40+yrs   Total
## ------------------------------------------------------------------
## Normal 2500+           77         222         226       36     561
##                     83.7%       88.4%       87.6%    90.0%
## ------------------------------------------------------------------
## Weight<2500            15          29          32        4      80
##                     16.3%       11.6%       12.4%    10.0%
## ------------------------------------------------------------------
## Total                  92         251         258       40     641
##                     14.4%       39.2%       40.2%     6.2%
## ==================================================================
table <- table(bw.data$lbw,bw.data$agegrp)
chisq.test(table)
```

```
## Warning in chisq.test(table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 1.6556, df = 3, p-value = 0.6468
```

4. Why do you think you get a warning that the Chi-squared approximation may be incorrect?

It gave the warning because some of the expected values will be very small and therefore the approximations of p may not be right.

5. In the birthweight dataset test the association between *sex* and low birth weight (*lbw*). Report the direction of the association and your conclusion on the association between the variables.

table1 <- table(bw.data$lbw, bw.data$sex) chisq.test(table1)

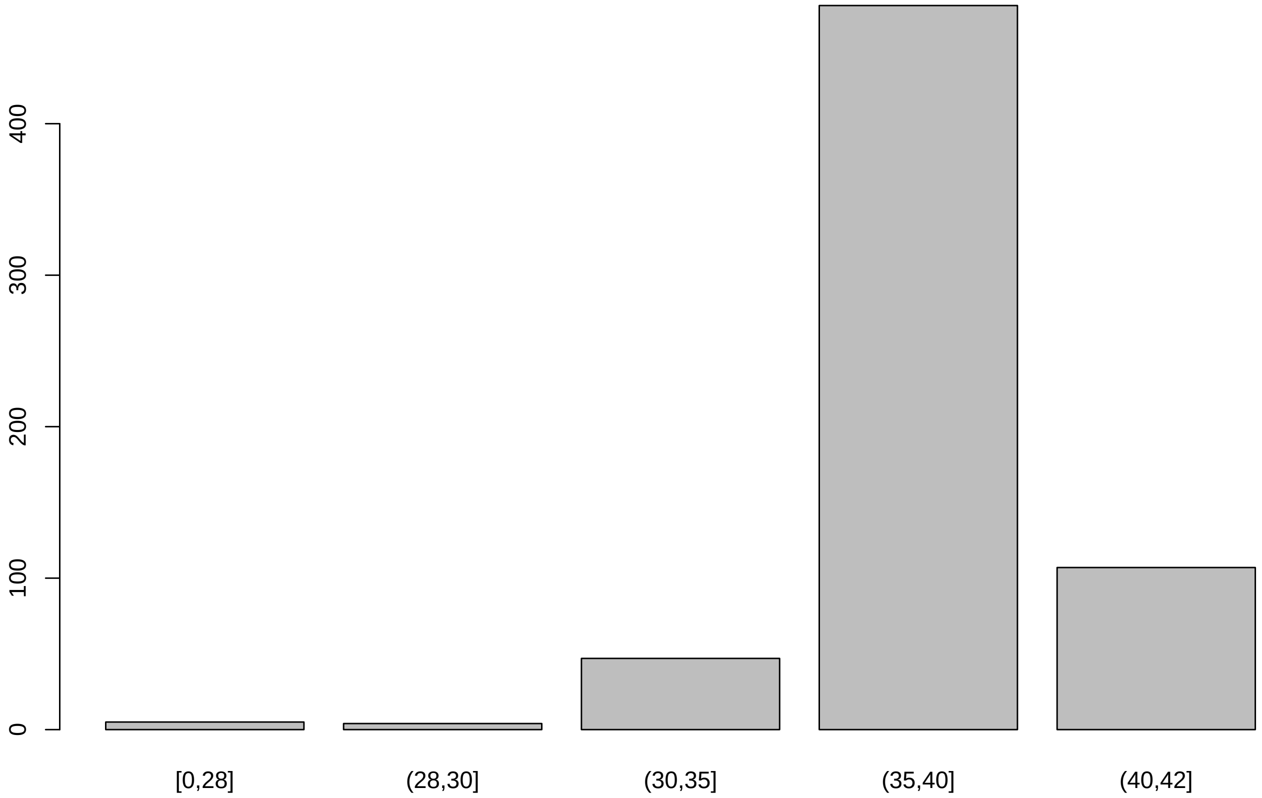
Fisher's exact test

6. Categorise gestation week (gestwks) into five groups. Cross tabulate the newly created variable and low birth weight *Outcome* variable

```
summary(bw.data$gestwks)
```

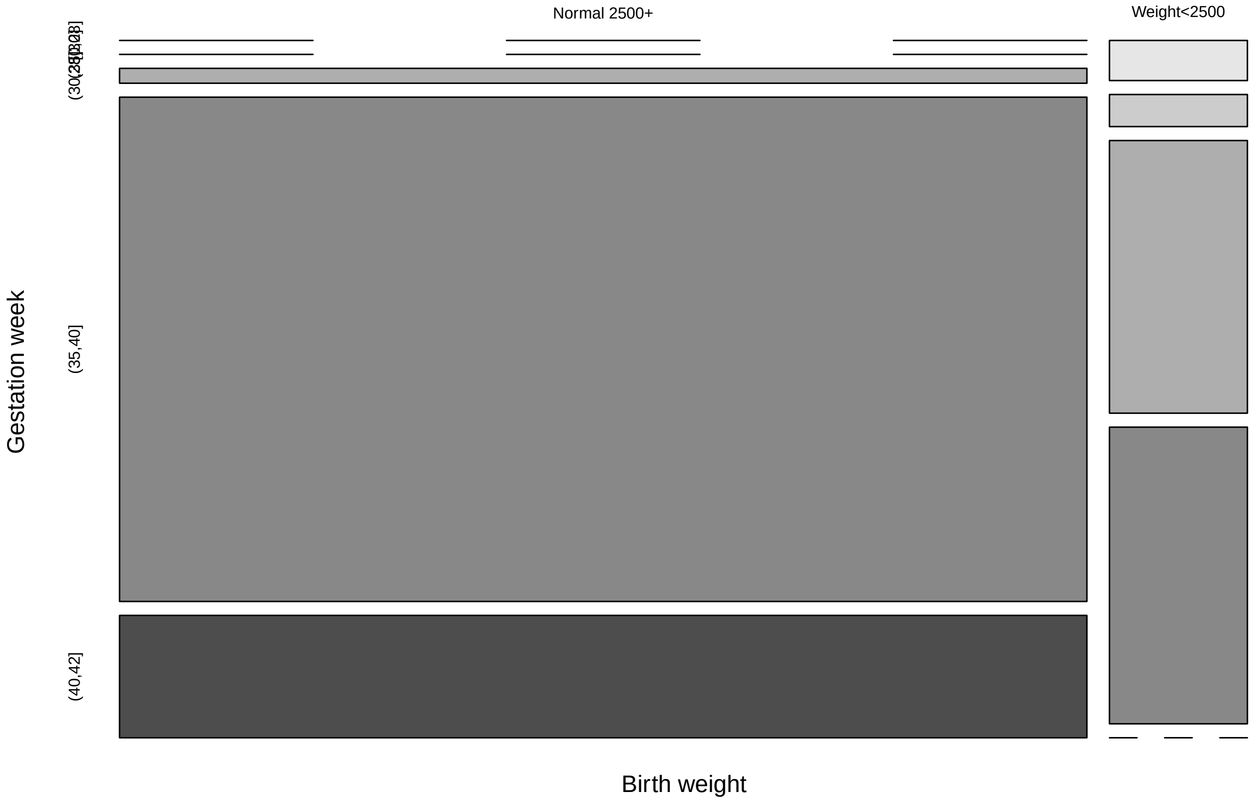
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.00   38.00   39.00   38.69   40.00   42.00
```

```
maxgestwks<-max(bw.data$gestwks)
bw.data$gestwks_gp <- cut(bw.data$gestwks, breaks=c(0,28,30,35,40,maxgestwks), include.lowest = TRUE)
freq(bw.data$gestwks_gp)
```



```
## bw.data$gestwks_gp
##         Frequency Percent
## [0,28]          5   0.780
## (28,30]         4   0.624
## (30,35]        47   7.332
## (35,40]       478  74.571
## (40,42]       107  16.693
## Total         641 100.000
```

```
crosstab(bw.data$gestwks_gp,bw.data$lbw, prop.r=T, xlab="Birth weight",ylab="Gestation week")
```



```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## ==========================================================
##                        bw.data$lbw
## bw.data$gestwks_gp    Normal 2500+   Weight<2500   Total
## ----------------------------------------------------------
## [0,28]                           0             5       5
##                               0.0%        100.0%    0.8%
## ----------------------------------------------------------
## (28,30]                          0             4       4
##                               0.0%        100.0%    0.6%
## ----------------------------------------------------------
## (30,35]                         13            34      47
##                              27.7%         72.3%    7.3%
## ----------------------------------------------------------
## (35,40]                        441            37     478
##                              92.3%          7.7%   74.6%
## ----------------------------------------------------------
## (40,42]                        107             0     107
##                             100.0%          0.0%   16.7%
## ----------------------------------------------------------
## Total                          561            80     641
## ==========================================================
```

Notice that there are cells with very low frequencies in the table.

7. Perform a Chi-square test to assess the association between the two variables

NB: This is an analysis for any type of outcome, in a more sanitised analysis we would not consider those who were lost to follow-up or those who withdrew in the tabulation

```
chisq.test(bw.data$gestwks_gp,bw.data$lbw, correct = FALSE)
```

```
## Warning in chisq.test(bw.data$gestwks_gp, bw.data$lbw, correct = FALSE):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  bw.data$gestwks_gp and bw.data$lbw
## X-squared = 242.38, df = 4, p-value < 2.2e-16
```

8. You get a warning message that the *Chi-squared approximation may be incorrect.* Why do you think so?

9. Perform a Fisher's exact test instead. Is there any difference in the p-value compared to the Chi-square one? If there is a difference why do you think one is larger/smaller than the other?

In R we can use the function **fisher.test** from the **stats** library.

*NB: The Fisher's exact test is memory intensive, you might need to increase the size of the working space for it to work. See **?fisher.test** for more details*

```
fisher.test(bw.data$gestwks_gp,bw.data$lbw)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  bw.data$gestwks_gp and bw.data$lbw
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```