Exploratory analysis of categorical data

Alice Kamau

7/5/2021

Learning objectives

- Explain key procedures for the analysis of categorical data
- Use R to perform tests on proportions for one, two or k categorical variables
- Interpret the results of tests on proportions for one, two or k categorical variables

Understanding categorical variables

- When we calculate summaries of categorical variables we are aiming to describe the sample distribution of the variable, just as with numeric variables.
- The general question we need to address is, 'what are the relative frequencies of different categories?'
- Since a categorical variable takes a finite number of possible values, the simplest thing to do is tabulate the number of occurances of each type.
- Load required package

```
library(tidyverse)
```

- Set the directory

```
setwd("/Users/akamau/Documents/I-StaR/Course 1/Day5/Presentation/")
```

- Load the data

```
bw_df <- read.csv("Data/birthweight2.csv")
```

table() & prop.table()

- table() & prop.table() is a quick way to pull together row/column frequencies and proportions for categorical variables

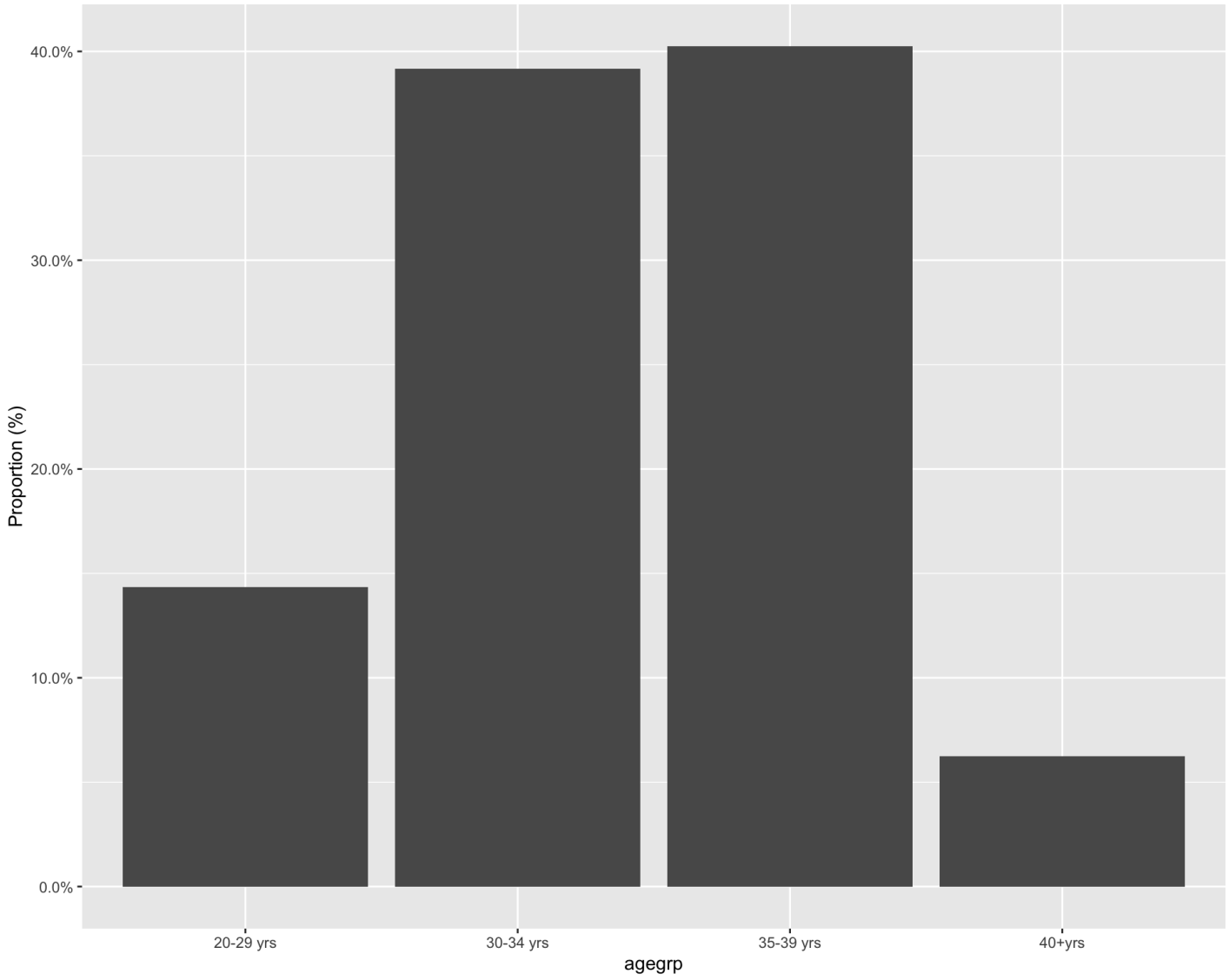
```
table(bw_df$lbw2)
```

```
##
##   0   1
## 561  80
```

```
prop.table(table(bw_df$lbw2))
```

```
##
##        0        1
## 0.875195 0.124805
```

```
ggplot(bw_df, aes(agegrp)) + geom_bar(aes(y = (..count..)/sum(..count..))) +
    scale_y_continuous(labels = scales::percent) + ylab("Proportion (%)")
```



R functions: binom.test() & prop.test()

- The R functions binom.test() and prop.test() can be used to perform one-proportion test:
- binom.test(): compute exact binomial test. Recommended when sample size is small
- prop.test(): can be used when sample size is large (N > 30). It uses a normal approximation to binomial
- The syntax of the two functions are exactly the same. The simplified format is as follow:

```
# syntax
binom.test(x, n, p = 0.5, alternative = "two.sided")
prop.test(x, n, p = NULL, alternative = "two.sided",
          correct = TRUE)

x: the number of of successes
n: the total number of trials
p: the probability to test against.
correct: a logical indicating whether Yates' continuity correction should be applied where possible.
```

One sample proportion test

- One sample proportion test is used to compare an observed proportion to a theoretical one, when there are only two categories.

```
# Assumming a normal approximation H0: Proportion of normal
# birthweight = 90%
prop.test(sum(bw_df$lbw2 == 0), length(bw_df$lbw2 == 0), p = 0.9,
    correct = T)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(bw_df$lbw2 == 0) out of length(bw_df$lbw2 == 0), null probability 0.9
## X-squared = 4.1109, df = 1, p-value = 0.04261
## alternative hypothesis: true p is not equal to 0.9
## 95 percent confidence interval:
##  0.8465105 0.8992740
## sample estimates:
##         p
## 0.875195
```

```
# Using exact binomial test
binom.test(sum(bw_df$lbw2 == 0), length(bw_df$lbw2 == 0), p = 0.9)
```

```
##
##  Exact binomial test
##
## data:  sum(bw_df$lbw2 == 0) and length(bw_df$lbw2 == 0)
## number of successes = 561, number of trials = 641, p-value =
## 0.04099
## alternative hypothesis: true probability of success is not equal to 0.9
## 95 percent confidence interval:
##  0.8470910 0.8997842
## sample estimates:
## probability of success
##               0.875195
```

- The function returns:
  - *the value of Pearson's chi-squared test statistic.*
  - *a p-value*
  - *a 95% confidence intervals*
  - *an estimated probability of success (the proportion of children with normal weight)*

Recap of basic tools for analysing binary data

- Descriptive: Bar charts and tabulations
- Analytic: Use of prop.test() assuming normal approximation or using binom.test() based on the exact distribution

Exercise

- Use birthweight2
- Check the variables, and explore the data.
- Generate a barplot of lbw stratified by sex
- Get the proportion of low birth weight babies and 95% CI.
- Get the proportion of lbw babies (and 95% CI) by sex.
- Test this hypothesis p=0.90 (90% normal BW) for female babies and male babies separately

Solution

- Check the variables, and explore the data.

```
names(bw_df)
```

```
##  [1] "id"       "matage"   "ht"       "gestwks"  "sex"      "bweight"
##  [7] "ethnic"   "lbw"      "agegrp"   "lbw2"     "agegrp1"
```

```
str(bw_df)
```

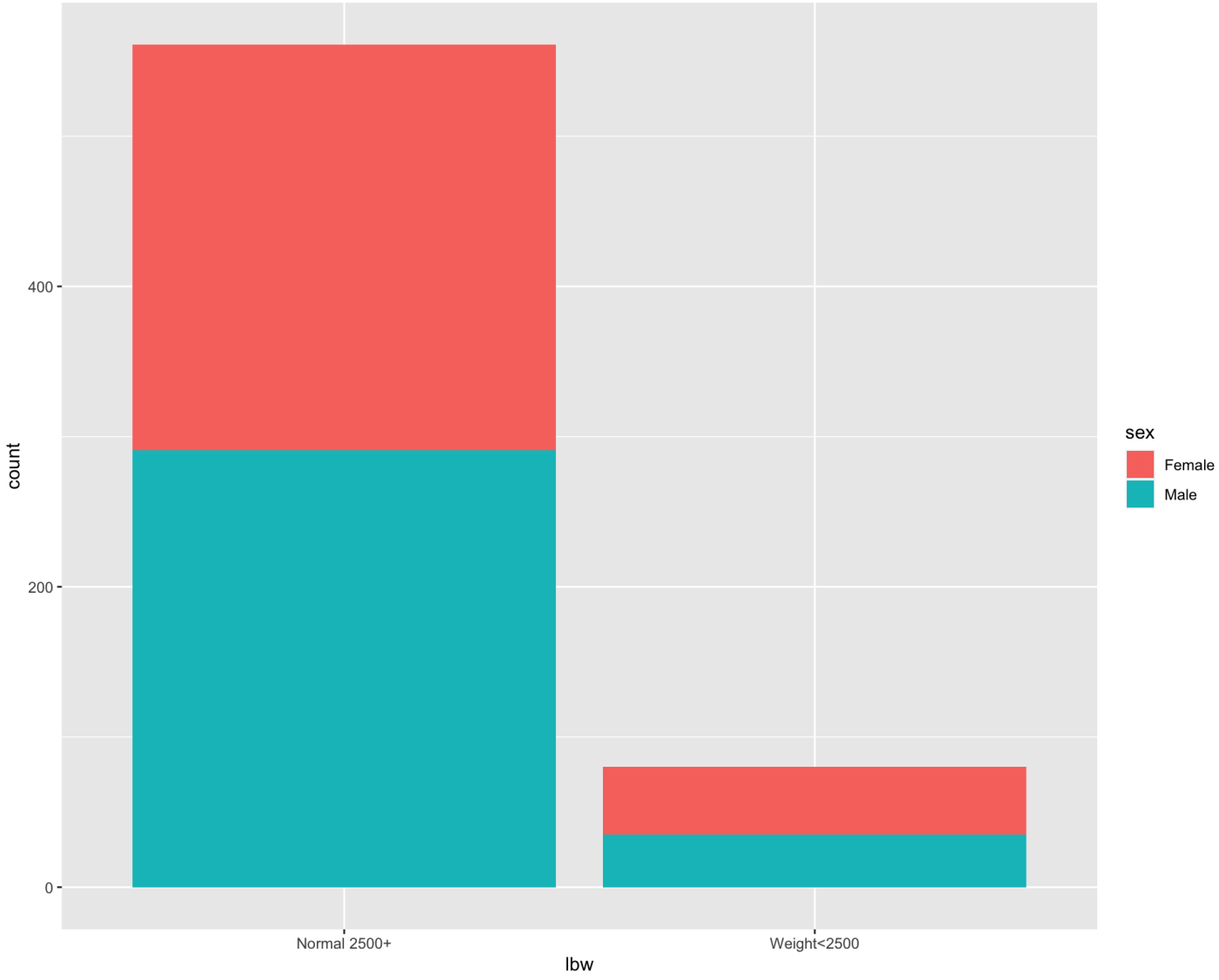
```
## 'data.frame':    641 obs. of  11 variables:
##  $ id     : int  107 579 438 570 569 210 105 528 382 403 ...
##  $ matage : int  23 23 24 24 25 25 25 25 25 25 ...
##  $ ht     : int  2 2 1 2 1 1 2 2 1 2 ...
##  $ gestwks: int  39 41 36 39 31 38 38 39 39 40 ...
##  $ sex    : chr  "Female" "Female" "Female" "Female" ...
##  $ bweight: int  3680 3120 2720 2550 1320 3260 3340 3040 3210 3380 ...
##  $ ethnic : int  1 4 3 4 4 1 1 4 3 3 ...
##  $ lbw    : chr  "Normal 2500+" "Normal 2500+" "Normal 2500+" "Normal 2500+" ...
##  $ agegrp : chr  "20-29 yrs" "20-29 yrs" "20-29 yrs" "20-29 yrs" ...
##  $ lbw2   : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ agegrp1: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(bw_df)
```

```
##    id matage ht gestwks    sex bweight ethnic          lbw    agegrp
## 1 107     23  2      39 Female    3680      1 Normal 2500+ 20-29 yrs
## 2 579     23  2      41 Female    3120      4 Normal 2500+ 20-29 yrs
## 3 438     24  1      36 Female    2720      3 Normal 2500+ 20-29 yrs
## 4 570     24  2      39 Female    2550      4 Normal 2500+ 20-29 yrs
## 5 569     25  1      31 Female    1320      4  Weight<2500 20-29 yrs
## 6 210     25  1      38   Male    3260      1 Normal 2500+ 20-29 yrs
##   lbw2 agegrp1
## 1    0       1
## 2    0       1
## 3    0       1
## 4    0       1
## 5    1       1
## 6    0       1
```

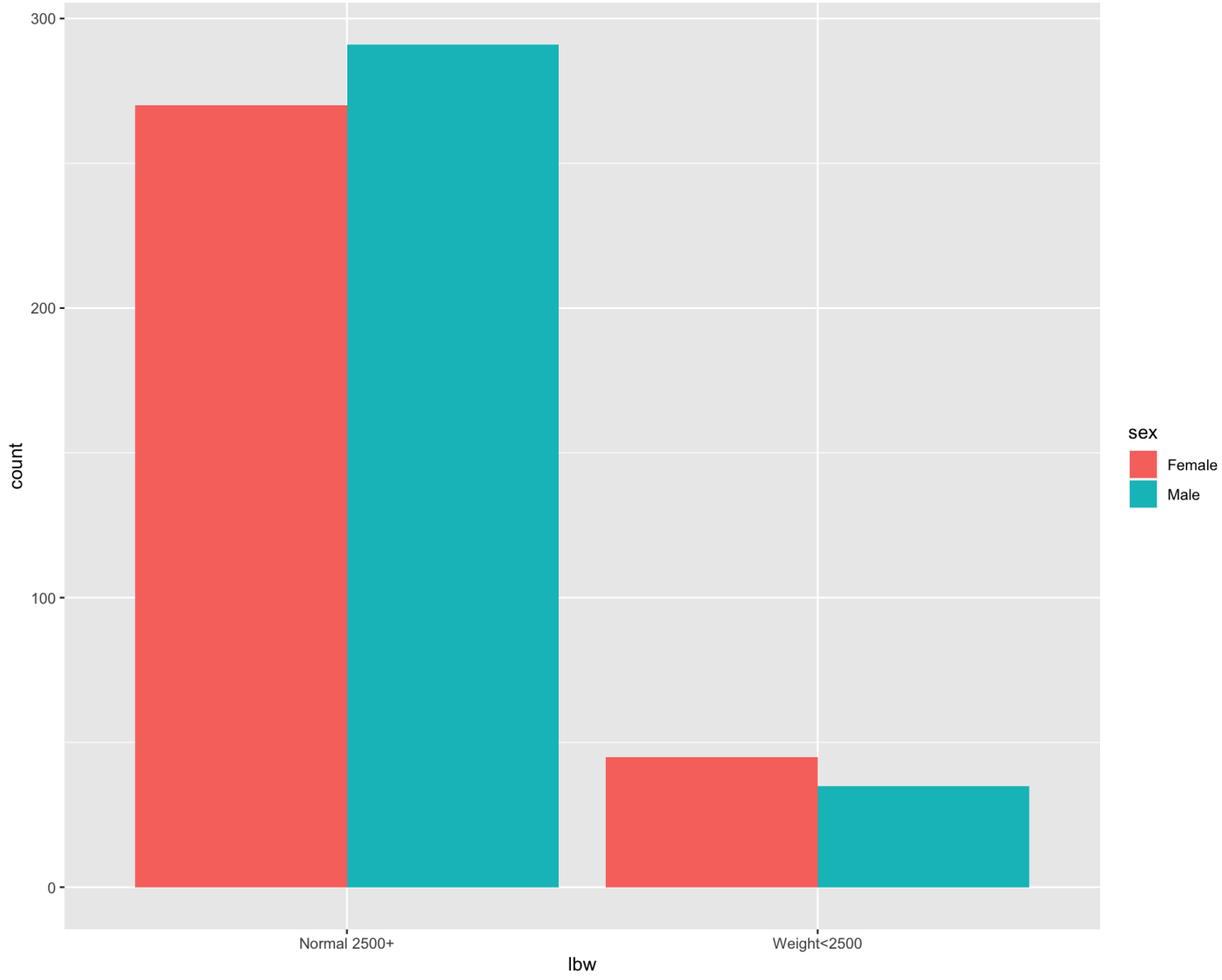
- Generate a barplot of lbw stratified by sex (default is stacked)

```
ggplot(bw_df, aes(x = lbw)) + geom_bar(aes(fill = sex))
```



- Generate a barplot of lbw stratified by sex (side by side)

```
ggplot(bw_df, aes(lbw)) + geom_bar(aes(fill = sex), position = "dodge")
```



- Get the proportion of low birth weight babies and 95% CI.

```
prop.test(sum(bw_df$lbw2 == 1), length(bw_df$lbw2 == 1))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(bw_df$lbw2 == 1) out of length(bw_df$lbw2 == 1), null probability 0.5
## X-squared = 359.44, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1007260 0.1534895
## sample estimates:
##       p
## 0.124805
```

- Get the proportion of lbw babies (and 95% CI) by sex.

```
table(bw_df$sex)
```

```
##
## Female   Male
##    315    326
```

```
bw_female <- bw_df %>% filter(sex == "Female")
prop.table(table(bw_female$lbw2))
```

```
##
##         0         1
## 0.8571429 0.1428571
```

```
bw_male <- bw_df %>% filter(sex == "Male")
prop.table(table(bw_male$lbw2))
```

```
##
##        0        1
## 0.892638 0.107362
```

Females

- Test this hypothesis p=0.90 (90% normal BW) for female babies and male babies separately

```
prop.test(sum(bw_female$lbw2 == 0), length(bw_female$lbw2 ==
    0), p = 0.9, correct = T)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(bw_female$lbw2 == 0) out of length(bw_female$lbw2 == 0), null probability 0.9
## X-squared = 5.9612, df = 1, p-value = 0.01462
## alternative hypothesis: true p is not equal to 0.9
## 95 percent confidence interval:
##  0.8124482 0.8928825
## sample estimates:
##         p
## 0.8571429
```

Males

```
prop.test(sum(bw_male$lbw2 == 0), length(bw_male$lbw2 == 0),
    p = 0.9, correct = T)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(bw_male$lbw2 == 0) out of length(bw_male$lbw2 == 0), null probability 0.9
## X-squared = 0.12304, df = 1, p-value = 0.7258
## alternative hypothesis: true p is not equal to 0.9
## 95 percent confidence interval:
##  0.8526247 0.9230946
## sample estimates:
##       p
## 0.892638
```

Associations between categorical variables

- The general question we need to address is, "do different combinations of categories seem to be under or over represented?"
- We need to understand which combinations are common and which are rare.
- The simplest thing we can do is 'cross-tabulate' the number of occurrences of each combination.
- The resulting table is called a contingency table.
- The counts in the table are sometimes referred to as frequencies.

```r
tab1 <- table(bw_df$sex, bw_df$lbw2)
prop.table(tab1, 1)  ## row proportions
```

```
##
##                  0         1
##   Female 0.8571429 0.1428571
##   Male   0.8926380 0.1073620
```

```r
prop.table(tab1, 2)  ## column proportions
```

```
##
##                  0         1
##   Female 0.4812834 0.5625000
##   Male   0.5187166 0.4375000
```

```r
##
prop.table(table(bw_df$sex, bw_df$lbw2), 1)
```

```
##
##                  0         1
##   Female 0.8571429 0.1428571
##   Male   0.8926380 0.1073620
```

Using dplyr & tidyr: Crosstabs

- A good reasons for not just using the base table() command is when you dealing with missing data

```
## frequency
bw_df %>% group_by(sex, lbw2) %>% summarise(n = n()) %>% spread(lbw2,
    n) %>% kable()
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups` argument.
```

sex	0	1
Female	270	45
Male	291	35

```
## proportion

bw_df%>%
  group_by(sex, lbw2)%>%
  summarize(n=n())%>%
  mutate(prop=n/sum(n))%>%
  subset(select=c("sex","lbw2","prop"))%>%   #drop the frequency value
  spread(lbw2, prop) %>%
  kable()
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups` argument.
```

sex	0	1
Female	0.8571429	0.1428571
Male	0.8926380	0.1073620

Including some missing data

```
bw_df %>%
  mutate(
    ethnic2 = ifelse(ethnic==2, NA, ethnic),
    agegrp2=ifelse(agegrp=="40+yrs",NA,agegrp)
  ) %>%
  group_by(ethnic2, agegrp2) %>%
  tally() %>%
  data.frame() %>% ### <-- go from tbl_df to data.frame
  spread(agegrp2, n)
```

```
##   ethnic2 20-29 yrs 30-34 yrs 35-39 yrs <NA>
## 1       1        28       103       108   21
## 2       3        28        70        55    6
## 3       4        22        56        58    5
## 4      NA        14        22        37    8
```

Using base R table() command

- base R ignores this missing value in the output table

```
bw_df$ethnic2 <- ifelse(bw_df$ethnic == 2, NA, bw_df$ethnic)
bw_df$agegrp2 <- ifelse(bw_df$agegrp == "40+yrs", NA, bw_df$agegrp)
table(bw_df$ethnic2, bw_df$agegrp2)
```

```
##
##     20-29 yrs 30-34 yrs 35-39 yrs
##   1        28       103       108
##   3        28        70        55
##   4        22        56        58
```

Comparing proportions

- To test the hypothesis that the proportions are different, there are several ways to do this:
  - *Using chi-squared test*
  - *Using fishers exact test*
  - *Two sample proportion test*

Chi-squared test - Comparing proportions

- Comparing two (or more) proportions - the Chi-squared test uses Expected numbers.
- Chi-squared test is valid for any contingency table
- Assumptions: sufficient numbers in each cell of the table
  - *State the null hypothesis: No association between the two variables.*
  - *Calculate the Chi-squared statistic from the Observed and Expected numbers*
  - *Obtain the p-value for the data, under H0*

```
chisq.test(bw_df$sex, bw_df$lbw2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  bw_df$sex and bw_df$lbw2
## X-squared = 1.5372, df = 1, p-value = 0.215
```

```
# If Chi-squared test not valid then get R to test the null
# hypothesis H0 using the Fishers exact test.
fisher.test(bw_df$sex, bw_df$lbw2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  bw_df$sex and bw_df$lbw2
## p-value = 0.1895
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.436256 1.187129
## sample estimates:
## odds ratio
##  0.7220222
```

Two sample proportion test

```
prop.test(table(bw_df$sex, bw_df$lbw2))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(bw_df$sex, bw_df$lbw2)
## 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
```

Exercise

- Use birthweight2, with outcome low birth weight (lbw)
- Ensure you have the variable that shows 1= LBW, 0=Normal
- Compare the proportion with low birth weight by the ethnic groups.
- Tabulate and test if lbw differs by ethnic.
- Tabulate the low birth weight by hypertension status of mothers (variable is called ht)
- Look at the association between lbw and hypertension (ht), using the chi-squared test

Solutions

- Ensure you have the variable that shows 1= LBW, 0=Normal

```
table(bw_df$lbw2)
```

```
##
##   0   1
## 561  80
```

- Compare the proportion with low birth weight by the ethnic groups.

```
prop.table(table(bw_df$ethnic, bw_df$lbw2), 1)
```

```
##
##            0         1
##   1 0.8846154 0.1153846
##   2 0.8765432 0.1234568
##   3 0.8427673 0.1572327
##   4 0.8936170 0.1063830
```

- Test if lbw differs by ethnic.

```
## Chi-squared test can be used for larger tables, with more
## categories (e.g. ethnic).
chisq.test(bw_df$ethnic, bw_df$lbw2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  bw_df$ethnic and bw_df$lbw2
## X-squared = 2.1814, df = 3, p-value = 0.5356
```

```
prop.test(table(bw_df$ethnic, bw_df$lbw2))
```

```
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  table(bw_df$ethnic, bw_df$lbw2)
## X-squared = 2.1814, df = 3, p-value = 0.5356
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.8846154 0.8765432 0.8427673 0.8936170
```

- Tabulate the low birth weight by hypertension status of mothers (variable is called ht)

```
## frequency
bw_df %>% group_by(ht, lbw2) %>% summarise(n = n()) %>% spread(lbw2,
    n) %>% kable()
```

```
## `summarise()` has grouped output by 'ht'. You can override using the `.groups` argument.
```

ht	0	1
1	62	27
2	499	53

```r
## proportion

bw_df%>%
  group_by(ht, lbw2)%>%
  summarize(n=n())%>%
  mutate(prop=n/sum(n))%>%
  subset(select=c("ht","lbw2","prop"))%>%   #drop the frequency value
  spread(lbw2, prop)%>%
  kable()
```

```
## `summarise()` has grouped output by 'ht'. You can override using the `.groups` argument.
```

ht	0	1
1	0.6966292	0.3033708
2	0.9039855	0.0960145

- Look at the association between lbw and hypertension (ht), using the chi-squared test

```
chisq.test(bw_df$ht, bw_df$lbw2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  bw_df$ht and bw_df$lbw2
## X-squared = 28.301, df = 1, p-value = 1.038e-07
```

```
prop.test(table(bw_df$ht, bw_df$lbw2))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(bw_df$ht, bw_df$lbw2)
## X-squared = 28.301, df = 1, p-value = 1.038e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3124998 -0.1022128
## sample estimates:
##    prop 1    prop 2
## 0.6966292 0.9039855
```