Exercise getStartedR: solutions

Alessio Crippa

October 13, 2016

Get started with R, Exercises

The dataset *lowbwt* is about a study that aims to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Variables that were thought to be of importance were age, weight of the subject at her last menstrual period, smoking during pregnancy and race.

Questions

Data inspection

1. Load the **lowbwt** available at http://alecri.github.io/downloads/data/ (full address for Rdata file http://alecri.github.io/downloads/data/lowbwt.Rdata)

```
library(tidyverse)
## Load R dataset
load(url("http://alecri.github.io/downloads/data/lowbwt.Rdata"))
# check if it's loaded
ls()
```

```
## [1] "lowbwt"
```

```
## Alternatively, other data format can be used
lowbwt <- read.table("http://alecri.github.io/downloads/data/lowbwt.txt")
lowbwt <- read.csv("http://alecri.github.io/downloads/data/lowbwt.csv")
library(haven)
lowbwt <- read_dta("http://alecri.github.io/downloads/data/lowbwt.dta")
lowbwt <- read_sav("http://alecri.github.io/downloads/data/lowbwt.sav")
lowbwt <- read_sas("http://alecri.github.io/downloads/data/lowbwt.sas7bdat")
```

2. How many observations and variables are in the dataset?

```
# number of rows and columns
dim(lowbwt)
```

```
## [1] 189  11
```

```
# or
c(rows = nrow(lowbwt), cols = ncol(lowbwt))
```

```
## rows cols
##  189   11
```

```
# names of variables
names(lowbwt)
```

```
##  [1] "id"    "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"
##  [9] "ui"    "ftv"   "bwt"
```

```
glimpse(lowbwt)
```

```
## Observations: 189
## Variables: 11
## $ id    <dbl> 4, 10, 11, 13, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 2...
## $ low   <fctr> < 2500 g, < 2500 g, < 2500 g, < 2500 g, < 2500 g, < 250...
## $ age   <dbl> 28, 29, 34, 25, 25, 27, 23, 24, 24, 21, 32, 19, 25, 16, ...
## $ lwt   <dbl> 120, 130, 187, 105, 85, 150, 97, 128, 132, 165, 105, 91,...
## $ race  <fctr> Other, White, Black, Other, Other, Other, Other, Black,...
## $ smoke <fctr> Yes, No, Yes, No, No, No, No, No, No, Yes, Yes, Yes, No...
## $ ptl   <dbl> 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1, 0,...
## $ ht    <fctr> No, No, Yes, Yes, No, No, No, No, Yes, Yes, No, No, No,...
## $ ui    <fctr> Yes, Yes, No, No, Yes, No, Yes, No, No, No, No, Yes, No...
## $ ftv   <dbl> 0, 2, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 2, 2, 0, 0,...
## $ bwt   <dbl> 709, 1021, 1135, 1330, 1474, 1588, 1588, 1701, 1729, 179...
```

3. Sort the data by (increasing) age

```r
arrange(lowbwt, age)
```

```
## # A tibble: 189 × 11
##       id       low   age   lwt   race  smoke   ptl     ht     ui   ftv
##    <dbl>    <fctr> <dbl> <dbl> <fctr> <fctr> <dbl> <fctr> <fctr> <dbl>
## 1     78  < 2500 g    14   101  Other    Yes     1     No     No     0
## 2     81  < 2500 g    14   100  Other     No     0     No     No     2
## 3    213 >= 2500 g    14   135  White     No     0     No     No     0
## 4     57  < 2500 g    15   110  White     No     0     No     No     0
## 5     62  < 2500 g    15   115  Other     No     0     No    Yes     0
## 6    102 >= 2500 g    15    98  Black     No     0     No     No     0
## 7     25  < 2500 g    16   130  Other     No     0     No     No     1
## 8    143 >= 2500 g    16   110  Other     No     0     No     No     0
## 9    166 >= 2500 g    16   112  Black     No     0     No     No     0
## 10   167 >= 2500 g    16   135  White    Yes     0     No     No     0
## # ... with 179 more rows, and 1 more variables: bwt <dbl>
```

4. Categorize age in two groups (< 30, >= 30 years). Attach the proper labels to the new factor variable.

```r
lowbwt$agecat = factor(lowbwt$age >= 30, labels = c("< 30", ">= 30"))
table(lowbwt$agecat)
```

```
##
##  < 30 >= 30
##   162    27
```

5. Select and print white subjects whose child's birth weight is less than 1.5 kg

```r
filter(lowbwt, bwt < 1500)
```

```
## # A tibble: 5 × 12
##      id      low   age   lwt   race  smoke   ptl     ht     ui   ftv   bwt
##   <dbl>   <fctr> <dbl> <dbl> <fctr> <fctr> <dbl> <fctr> <fctr> <dbl> <dbl>
## 1     4 < 2500 g    28   120  Other    Yes     1     No    Yes     0   709
## 2    10 < 2500 g    29   130  White     No     0     No    Yes     2  1021
## 3    11 < 2500 g    34   187  Black    Yes     0    Yes     No     0  1135
## 4    13 < 2500 g    25   105  Other     No     1    Yes     No     0  1330
## 5    15 < 2500 g    25    85  Other     No     0     No    Yes     0  1474
## # ... with 1 more variables: agecat <fctr>
```

Univariate statistics

6. Summarize the continuous response variable birth weight. What is its mean and standard deviation?

```
summary(lowbwt$bwt)
```

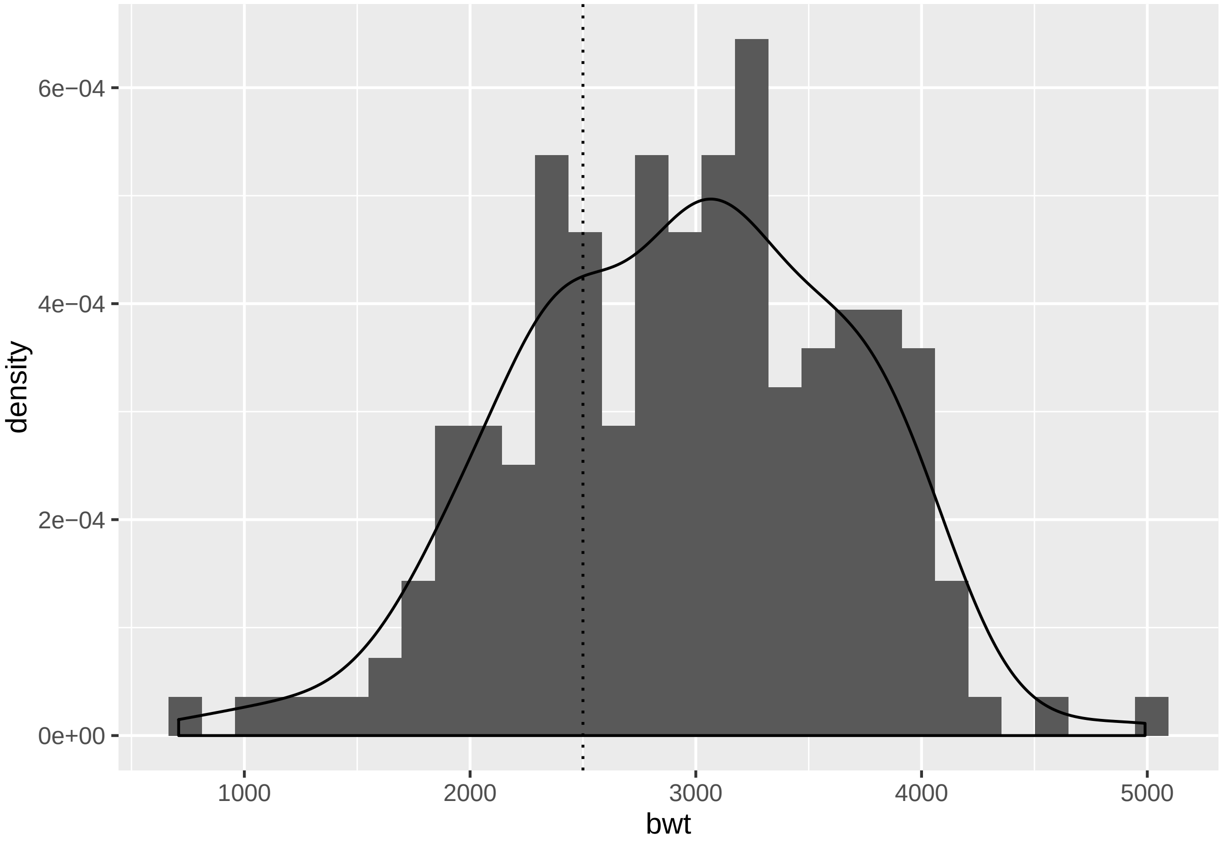
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     709    2414    2977    2945    3475    4990
```

```
c(mean = mean(lowbwt$bwt), std = sd(lowbwt$bwt))
```

```
##      mean       std
## 2944.6561  729.0224
```

7. Provide a graphical presentation of its distribution

```
ggplot(lowbwt, aes(x = bwt)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density() +
  geom_vline(xintercept = 2500, lty = "dotted")
```



8. Categorize birth weight in two groups: <2500 g and >= 2500 g (same as the *lwt* variable)

```
summary(lowbwt$bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     709    2414    2977    2945    3475    4990
```

```
lowbwt <- mutate(lowbwt, bwt_cat = cut(bwt, c(700, 2500, 5000), right = F,
                 levels = c(1, 0), labels = c("<2.5 kg", ">=2.5 kg")))
head(lowbwt$bwt_cat)
```

```
## [1] <2.5 kg <2.5 kg <2.5 kg <2.5 kg <2.5 kg <2.5 kg
## Levels: <2.5 kg >=2.5 kg
```

```
# check also with the variable 'low'
```

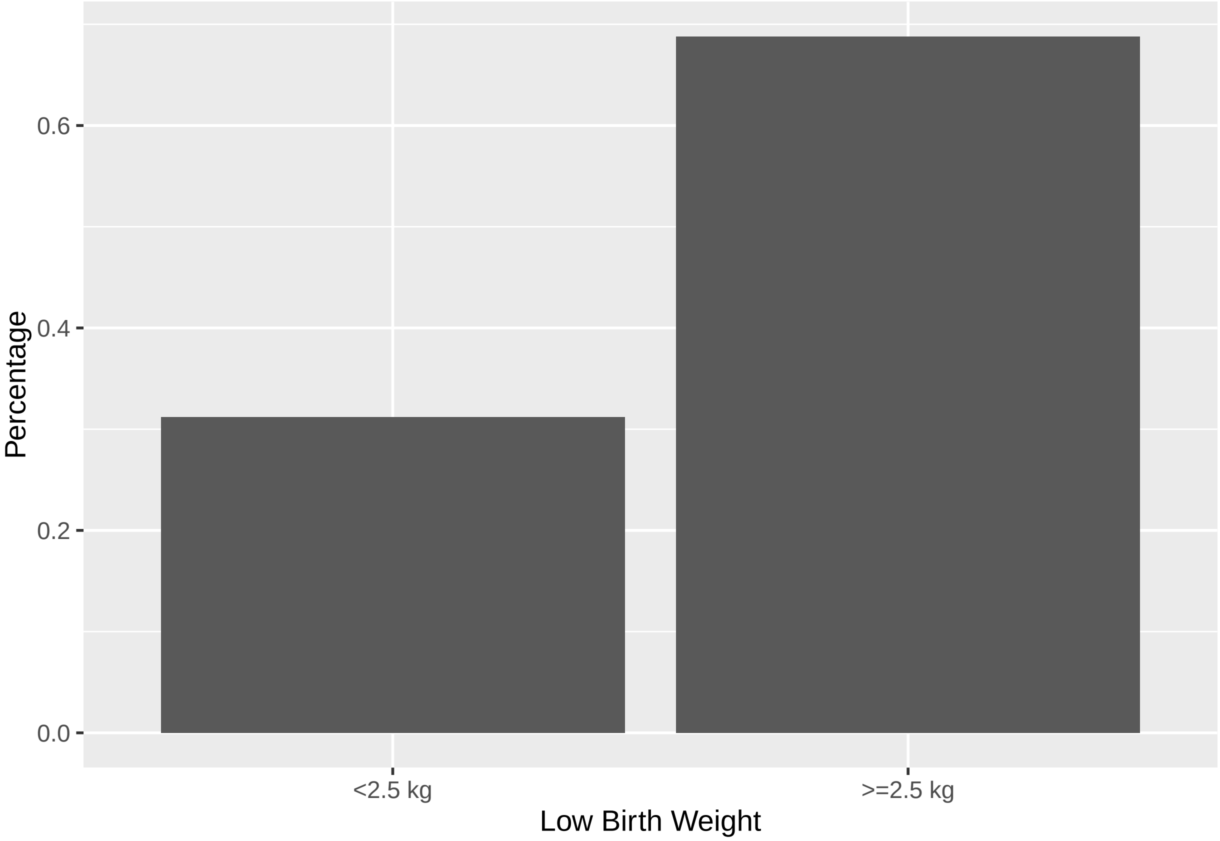
9. What is the percentage of women who had a baby weighting less than 2.5 kg?

```
tab <- table(lowbwt$bwt_cat)
prop.table(tab)
```

```
##
##    <2.5 kg  >=2.5 kg
## 0.3121693 0.6878307
```

10. Provide a graphical presentation for this binary variable

```
ggplot(lowbwt, aes(x = bwt_cat)) +
  geom_bar(aes(y = ..count../sum(..count..))) +
  labs(y = "Percentage", x = "Low Birth Weight")
```



Bivariate association

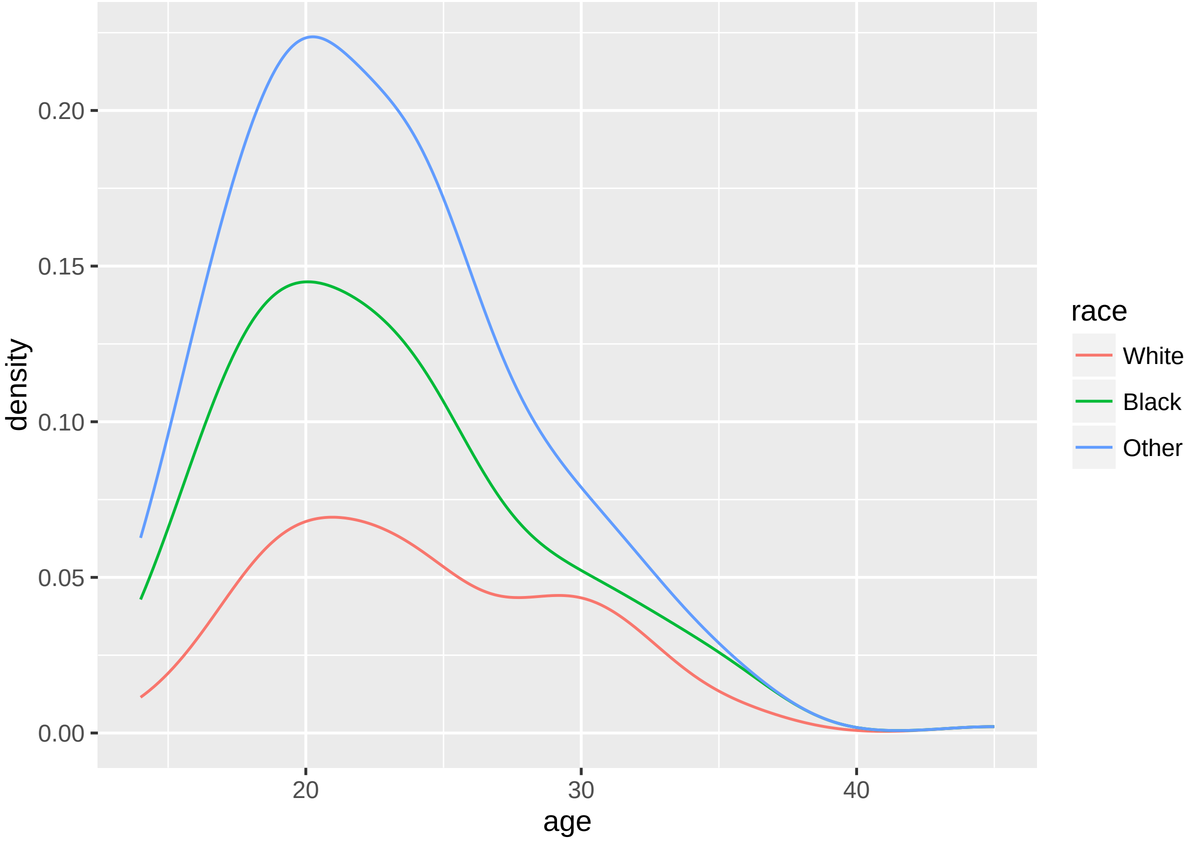
11. What is the mean and standard deviation of mother's age among white, black, and other races?

```
lowbwt %>% group_by(race) %>%
  summarise(mean = mean(age), std = sd(age))
```

```
## # A tibble: 3 × 3
##     race     mean      std
##   <fctr>    <dbl>    <dbl>
## 1  White 24.29167 5.654838
## 2  Black 21.53846 5.108665
## 3  Other 22.38806 4.535901
```

12. Present graphically the distribution of mother's age in the races subgroups

```
ggplot(lowbwt, aes(x = age, color = race)) +
  stat_density(geom = "line")
```



13. What is the percentage of smoking mothers among white, black, and other races?

```
tab <- with(lowbwt, table(race, smoke))
prop.table(tab, margin = 2)
```

```
##        smoke
## race           No       Yes
##   White 0.3826087 0.7027027
##   Black 0.1391304 0.1351351
##   Other 0.4782609 0.1621622
```

14. What is the difference in the mean birth weight comparing smoker vs non-smoker women? Test the hypothesis of equality of means. What do you conclude?

```
lowbwt %>% group_by(smoke) %>% summarize(mean(bwt))
```

```
## # A tibble: 2 × 2
##    smoke `mean(bwt)`
##   <fctr>       <dbl>
## 1     No    3054.957
## 2    Yes    2773.243
```

```
t.test(bwt ~ smoke, data = lowbwt)
```

```
##
##  Welch Two Sample t-test
##
## data:  bwt by smoke
## t = 2.7095, df = 170, p-value = 0.00743
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   76.46677 486.95979
## sample estimates:
##  mean in group No mean in group Yes
##          3054.957          2773.243
```

15. What is the risk of low birth weight among smoker and non-smoker women? Test the hypothesis of equality of proportions (no association). What do you conclude?

```
tab <- with(lowbwt, table(bwt_cat, smoke))
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 4.2359, df = 1, p-value = 0.03958
```

```
library(Epi)
with(lowbwt, twoby2(smoke, bwt_cat))
```

```
## 2 by 2 table analysis:
## ------------------------------------------------------
## Outcome   : <2.5 kg
## Comparing : No vs. Yes
##
##     <2.5 kg >=2.5 kg    P(<2.5 kg) 95% conf. interval
## No       29       86        0.2522    0.1812   0.3394
## Yes      30       44        0.4054    0.3001   0.5203
##
##                                     95% conf. interval
##              Relative Risk:  0.6220    0.4093   0.9453
##          Sample Odds Ratio:  0.4946    0.2643   0.9254
## Conditional MLE Odds Ratio:  0.4965    0.2522   0.9720
##     Probability difference: -0.1532   -0.2871  -0.0176
##
##              Exact P-value: 0.0362
##         Asymptotic P-value: 0.0276
## ------------------------------------------------------
```