Advanced Statistics HW4 (Solution)

Shujia Wong

2018-11-08

Exercises 1

Load the `WHEATUSA2004` data frame from the `PASWR2` package, then transform it into a tibble format:

```
library(PASWR2)
data("WHEATUSA2004")
WHEATUSA2004 <- as_tibble(WHEATUSA2004)
WHEATUSA2004
```

```
## # A tibble: 30 x 2
##    states acres
##    <fct>  <int>
##  1 AR       620
##  2 CA       320
##  3 CO      1700
##  4 DE        47
##  5 GA       190
##  6 ID       700
##  7 IL       900
##  8 IN       440
##  9 KS      8500
## 10 KY       380
## # ... with 20 more rows
```

(a) Find the quantiles, deciles, mean, maximum, minimum, interquartile range, variance, and standard deviation for the variable acres. Comment on what the most appropriate measures of center and spread would be for this variable. What is the USA's 2004 total harvested wheat surface area?

```
# Quantiles
  quantile(WHEATUSA2004$acres, probs = c(0.25,0.5,0.75))
```

```
##     25%     50%     75%
##  198.75  630.00 1213.75
```

```
# Deciles
  quantile(WHEATUSA2004$acres, probs = seq(0.1,0.9,0.1))
```

```
##    10%    20%    30%    40%    50%    60%    70%    80%    90%
##  135.0  180.0  263.5  416.0  630.0  824.0  982.5 1634.0 1925.0
```

```
# Descriptive statistics
WHEATUSA2004 %>%
  summarize(
   mean = mean(acres),
   median = median(acres),
   max = max(acres),
   min = min(acres),
   IQR = IQR(acres),
   var = var(acres),
   sd = sd(acres),
```

```
  sum = sum(acres)
  )
```

```
## # A tibble: 1 x 8
##    mean median   max   min   IQR      var    sd   sum
##   <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <int>
## 1 1149.    630  8500    47  1015 2980303. 1726. 34462
```

The histogram shows in (c) that the distribution of harvested wheat is skewed to the right. This skew is seen in how much larger the mean (1148.73) is versus the median (630). *When working with skewed distributions, the median is the most appropriate measure of center.*

The standard deviation is an appropriate measure of spread for unimodal symmetric distributions, while the MAD is often used when the median is reported to describe the center of a skewed data set. Hence, the most appropriate measures of spread is the MAD which value equals 667.17.

The USA's 2004 total harvested wheat surface area is 34462.

(b) Which states are below the 20th percentile? Which states are above the 80th percentile? In which quantile is WI (Wisconsin)?

The following states are below the 20th percentile:

```
WHEATUSA2004 %>%
filter(acres < quantile(acres, probs = 0.2))
```

```
## # A tibble: 5 x 2
##   states acres
##   <fct>  <int>
## 1 DE        47
## 2 MD       145
## 3 MS       135
## 4 NY       100
## 5 PA       135
```

The fowllowing states are above the 80th percentile:

```
WHEATUSA2004 %>%
  filter(acres > quantile(WHEATUSA2004$acres, probs = 0.8))
```

```
## # A tibble: 6 x 2
##   states acres
##   <fct>  <int>
## 1 CO      1700
## 2 KS      8500
## 3 NE      1650
## 4 OK      4700
## 5 TX      3500
## 6 WA      1750
```

Given values $x_{(1)}, x_{(2)}, ..., x_{(n)}$, the p^{th} quantile for the k^{th} order statistic, $p(k)$, is[1]

$$p(k) = \frac{k-1}{n-1}, k \leq n$$

[1]Ugarte, Maria Dolores, Ana F. Militino, and Alan T. Arnholt (2016). *Probability and Statistics with R* (Text book). 2nd. Boca Raton, FL: CRC Press. page 124.

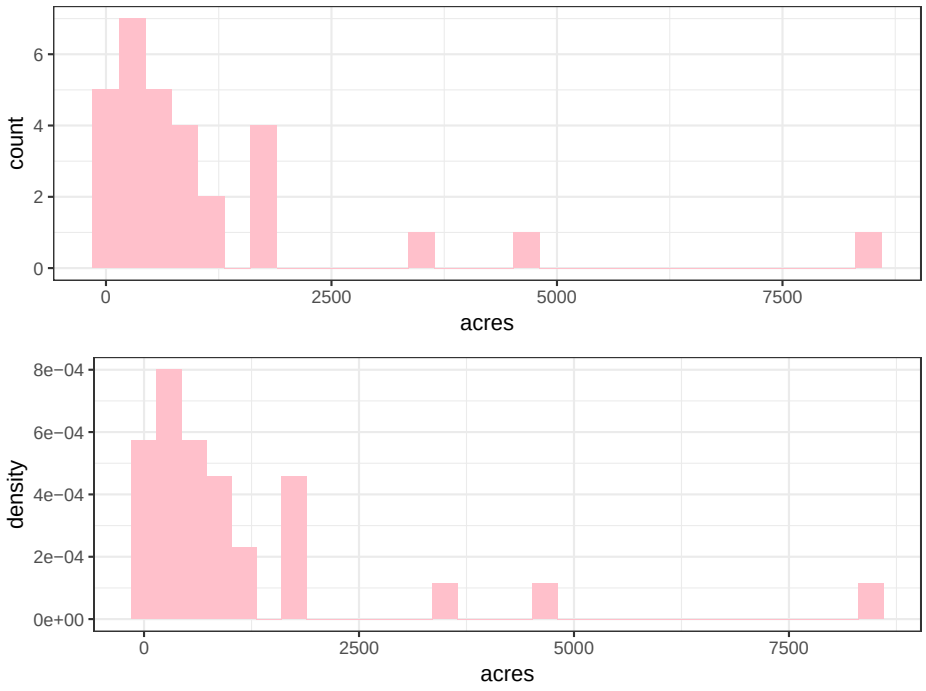
```
pk <- WHEATUSA2004 %>%
  filter(acres < WHEATUSA2004$acres[states == "WI"]) %>%
  summarize(
  pk = n()/(nrow(WHEATUSA2004)-1)
  )
pk
```

```
## # A tibble: 1 x 1
##       pk
##    <dbl>
## 1 0.276
```

The quantile of Wisconsin is 0.276.

(c) Create a frequency and a density histogram in the same graphics device using square plotting regions of the values in ACRES.

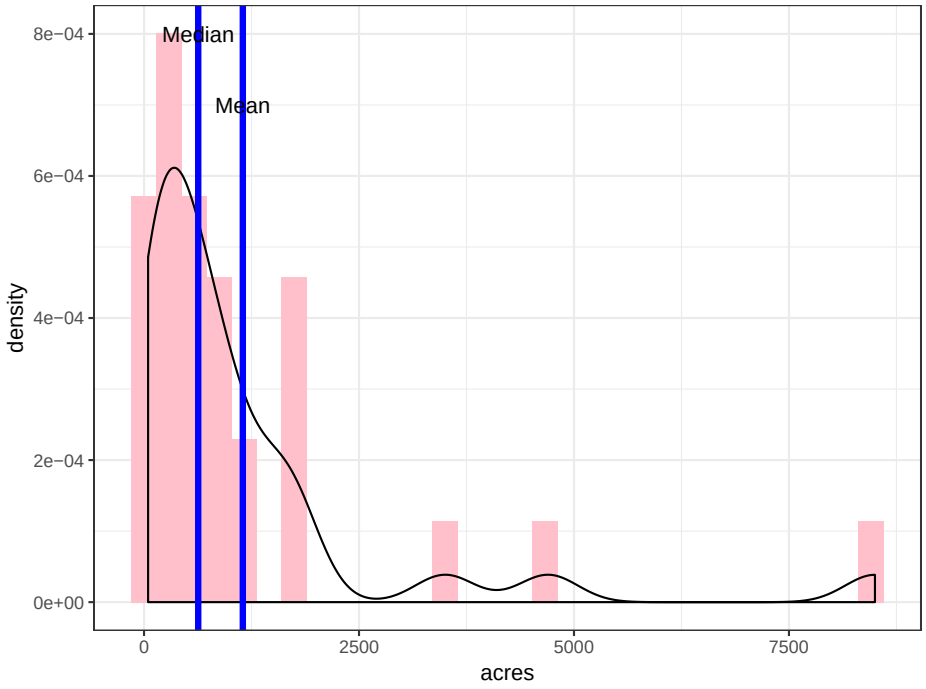
```
p1<-ggplot(WHEATUSA2004) +
  geom_histogram(aes(x=acres), fill = 'pink') +
  theme_bw()
p2<-ggplot(WHEATUSA2004)+
  geom_histogram(aes(x = acres, y = ..density..), fill = 'pink') +
  theme_bw()
multiplot(p1, p2)
```



(d) Add vertical lines to the density histogram from (c) to indicate the location of the mean and the median.

```
p1 <- ggplot(WHEATUSA2004, aes(x = acres, y = ..density..)) +
  geom_histogram(fill = 'pink') +
  geom_density() +
  theme_bw()
p1 + geom_vline(xintercept = c(median(WHEATUSA2004$acres),
                mean(WHEATUSA2004$acres)), size = 1.5, color = 'blue') +
```

```
    annotate("text", label = "Median", x = median(WHEATUSA2004$acres),
    y = 8e-04) +
    annotate("text", label = "Mean", x = mean(WHEATUSA2004$acres),
    y = 7e-04)
```



(e) Create a boxplot of the acres and locate the outliers.

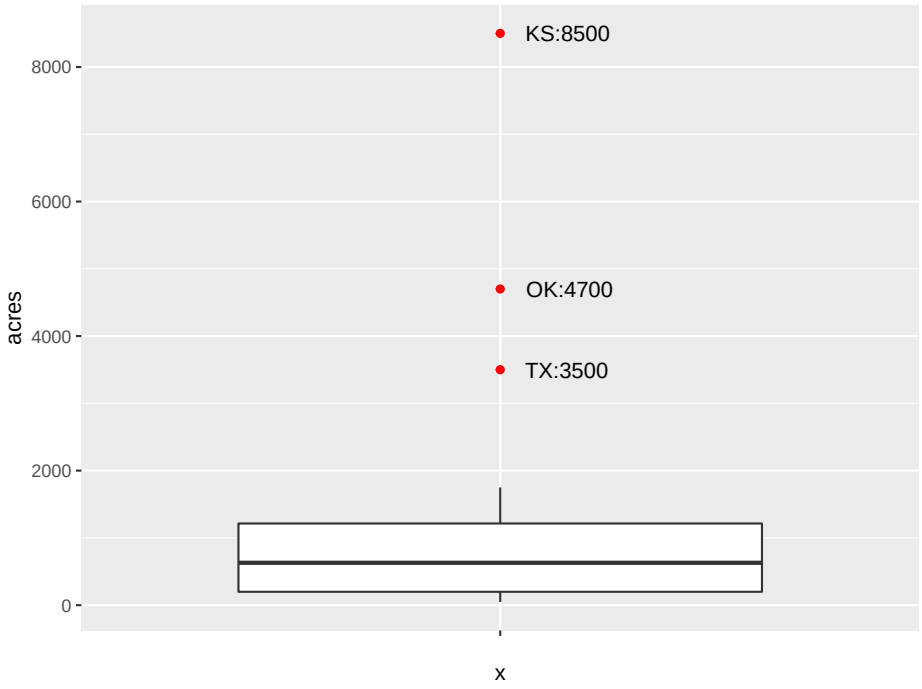
Outliers:

In a boxplot, visual points that display observations that fall more than 1.5 times the IQR from either edge of the box. These outlying points are unusual, so they are plotted individually[2].

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}

WHEATUSA2004 %>%
  mutate(outlier = ifelse(is_outlier(acres),
                          paste(states, acres, sep = ':') ,
                          as.character(NA))) %>%
  ggplot(aes(x = '', y = acres))+
  geom_boxplot(outlier.color = "red")+
  geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)
```

[2]Wickham, Hadley and Garrett Grolemund (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., page 96.



(f) Determine the state with the largest harvested wheat surface in acres. Remove this state from the data frame and compute the mean, median, and standard deviation of acres. How do these values compare to the values for these statistics computed in (a)?

```
# Find the largest state
largest <- WHEATUSA2004 %>%
  filter(rank(desc(acres)) == 1)
largest[1]
```

```
## # A tibble: 1 x 1
##   states
##   <fct>
## 1 KS
```

```
# The mean, median and sd after removing the largest state
WHEATUSA2004 %>%
  filter(!rank(desc(acres)) == 1) %>%
  summarize(
    median = median(acres),
    mean = mean(acres),
    sd = sd(acres)
  )
```

```
## # A tibble: 1 x 3
##   median  mean    sd
##    <int> <dbl> <dbl>
## 1    620  895. 1044.
```

The state with the largest harvested wheat surface in acres is KS. Once KS is removed, the mean and standard deviation are much smaller than those computed in part (a). Moreover, the median change little. This excercise shows that the mean is more *sensitive* to the extreme values while the median is more *robust*.

Exercises 2

Access the data from url http://www.stat.berkeley.edu/users/statlabs/data/babies.data and store the information in an object named `BABIES`. A description of the variables can be found at http://www.stat.berkeley.edu/users/statlabs/labs.html.

These data are a subset from a much larger study dealing with child health and development.

```
babies<-read.table("http://www.stat.berkeley.edu/users/statlabs/data/babies.data",header=T)
head(babies)
```

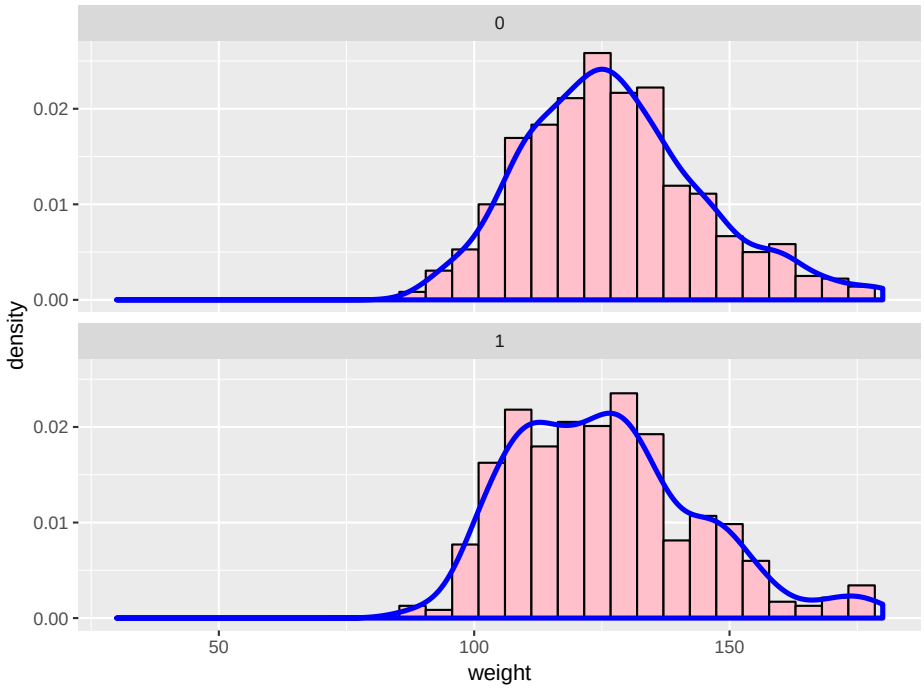
```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

(a) Create a "clean" data set that removes subjects if any observations on the subject are "unknown." Note that `bwt`, `gestation`, `parity`, `age`, `height`, `weight`, and `smoke` use values of 999, 999, 9, 99, 99, 999, and 9, respectively, to denote "unknown." Store the modified data set in an object named `CLEAN`.

```
CLEAN <- babies %>%
  filter(bwt != 999 & gestation != 999 & parity != 9 & age != 99 &
           height !=99 & weight != 999 & smoke != 9)
```

(b) Use the information in `CLEAN` to create a density histogram of the birth weights of babies whose mothers have never smoked (`smoke=0`) and another histogram placed directly below the first in the same graphics device for the birth weights of babies whose mothers currently smoke (`smoke=1`). Make the range of the x-axis 30 to 180 (ounces) for both histograms. Superimpose a density curve over each histogram.

```
CLEAN %>%
ggplot(aes(weight, y =..density..))+
  geom_histogram(color = "black",fill = "pink") +
  xlim(30, 180) +
  geom_density(size = 1.2, color = "blue") +
  facet_wrap(~smoke,ncol = 1)
```



(c) Based on the histograms in (b), characterize the distribution of baby birth weight for both non-smoking and smoking mothers.

The distribution of birth weight in the newborn of smoking mothers was flat and skew to the right, while the distribution looks normal for nonsmoking mothers. The proportion of babies born to smoking mothers who were overweight or underweight was higher.

(d) What is the mean weight difference between babies of smokers and non-smokers? Can you think of any reasons not to use the mean as a measure of center to compare birth weights in this problem?

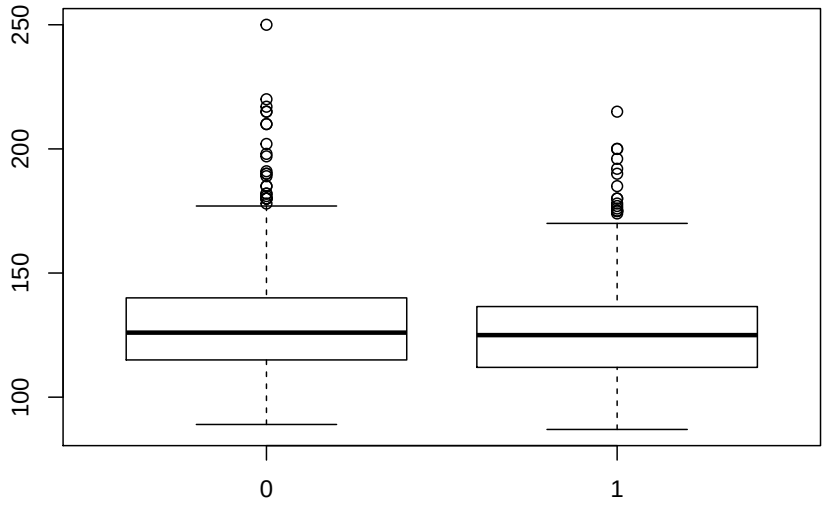
```
x1 <- CLEAN %>%
filter(smoke == 0)
x2 <- CLEAN %>%
filter(smoke == 1)
CLEAN %>%
  summarize(
  mean1 = mean(x1$weight),
  mean2 = mean(x2$weight),
  diff_mean = mean(x1$weight) - mean(x2$weight),
  median1 = median(x1$weight),
  median2 = median(x2$weight),
  diff_median = median(x1$weight)-median(x2$weight)
)
```

```
##      mean1    mean2 diff_mean median1 median2 diff_median
## 1 129.4797 126.9194   2.56033     126     125           1
```

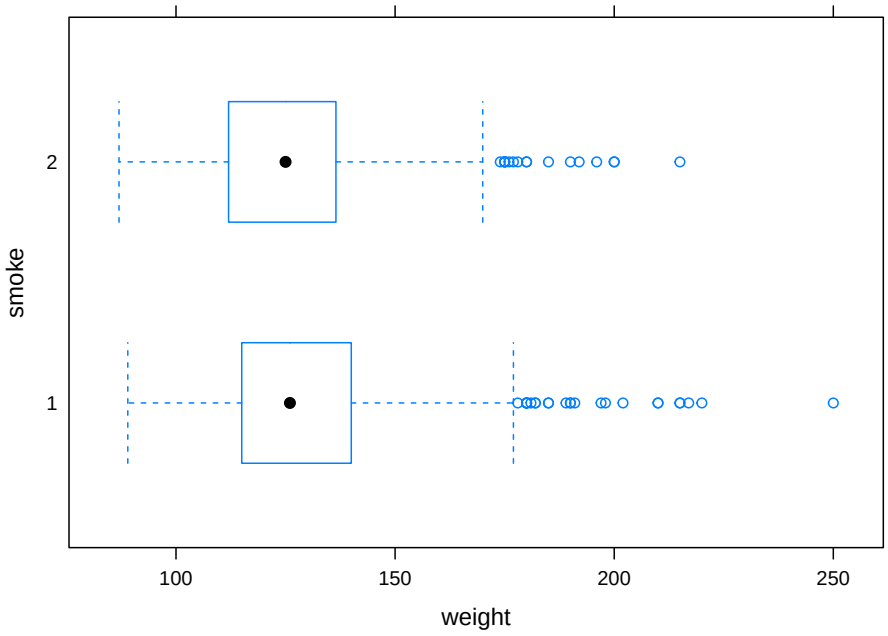
The mean weight difference between babies of smokers and non-smokers is 2.56, while the median difference is 1. The skewness of distribution for babies of smoking mothers, indicate that you should use the median, rather than the mean, as a measure of center to compare birth weights.

(e) Create side-by-side boxplots to compare the birth weights of babies whose mothers never smoked and those who currently smoke. Use traditional graphics (`boxplot()`), lattice graphics (`bwplot()`), and `ggplot` graphics to create the boxplots.

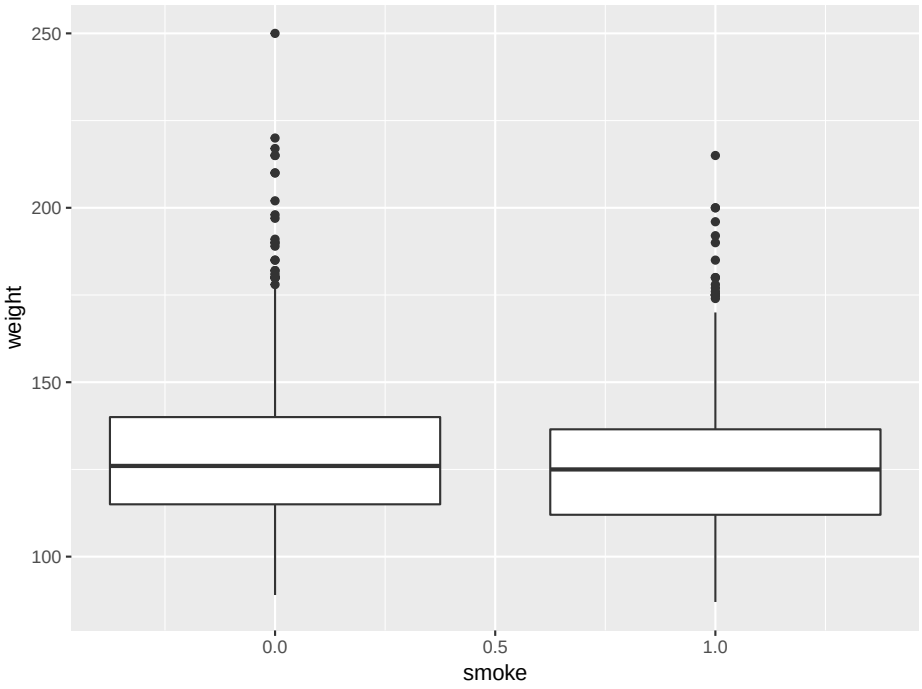
```
boxplot(weight ~ smoke, data = CLEAN)
```




```
library(lattice)
bwplot(smoke ~ weight, data = CLEAN)
```




```
ggplot(CLEAN)+
  geom_boxplot(aes(smoke,weight,group = smoke))
```


250
200
150
100
weight
0.0
0.5
1.0
smoke