Problem 2.3

Stephen Blatti

October 9, 2017

2.3 Load the data frame WheatSpain from the PASWR package.

```
library(PASWR2)
```

```
## Warning: package 'PASWR2' was built under R version 3.4.2
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

a. Find quantiles, deciles, mean, maximum, minimum, interquartile range, variance, and standard deviation of the variable hectares. Comment on the results. What was Spain's 2004 total harvested wheat area in hectares?

```
quantile(WHEATSPAIN$hectares)
```

```
##     0%    25%    50%    75%   100%
##     65   7203  25143 143250 619858
```

```
quantile(WHEATSPAIN$hectares, probs = seq(from = 0.1, to = 1.0, by = 0.1))
```

```
##      10%      20%      30%      40%      50%      60%      70%      80%
##    304.0   6329.4   9040.6  15397.6  25143.0  53481.2  88014.8 239389.2
##      90%     100%
## 410204.2 619858.0
```

```
mean(WHEATSPAIN$hectares)
```

```
## [1] 126561.5
```

```
median(WHEATSPAIN$hectares)
```

```
## [1] 25143
```

```
IQR(WHEATSPAIN$hectares)
```

```
## [1] 136047
```

```
var(WHEATSPAIN$hectares)
```

```
## [1] 38934822657
```

```
sd(WHEATSPAIN$hectares)
```

```
## [1] 197319.1
```

```
sum(WHEATSPAIN$hectares)
```

```
## [1] 2151546
```

```
  Spain's 2004 distribution of harvested wheat is skewed to the right - see that the m
ean 126561.5 the median is only 25143. The difference between Q1 and Q2 is also much s
maller than the difference between Q3 and Q2. The total harvested area is 2151546 hect
ares.
```

b. Create a function that calculates the quantiles, mean, variance, standard deviation, total, and the range of any variable.

```
calculate_stats <- function(x, ...){
  Q <- quantile(x)
  M <- mean(x)
  V <- var(x)
  SD <- sd(x)
  S <- sum(x)
  R <- diff(range(x))
  print(c(Quantiles = Q, Mean = M, Var = V, SD = SD, Total = S, Range = R))
  }
calculate_stats(WHEATSPAIN$hectares)
```

```
##   Quantiles.0%  Quantiles.25%  Quantiles.50%  Quantiles.75% Quantiles.100%
##   6.500000e+01   7.203000e+03   2.514300e+04   1.432500e+05   6.198580e+05
##           Mean            Var             SD          Total          Range
##   1.265615e+05   3.893482e+10   1.973191e+05   2.151546e+06   6.197930e+05
```

c. Which communities are below the 10th percentile in hectares? Which communities are above the 90th percentile? In which percentile is Navarra?

```
# bottom 10% of communities
below10 <- quantile(WHEATSPAIN$hectares, probs = 0.10)
WHEATSPAIN[WHEATSPAIN$hectares < below10, ]
```

```
##    community hectares acres
## 2   Asturias       65 160.6
## 17  Canarias      100 247.1
```

```
# top 10% of communities
above90 <- quantile(WHEATSPAIN$hectares, probs = 0.90)
WHEATSPAIN[WHEATSPAIN$hectares > above90, ]
```

```
##        community hectares   acres
## 10 Castilla-Leon   619858 1531703
## 16     Andalucia   558292 1379570
```

```
# Navarra
WHEATSPAIN[order(WHEATSPAIN$hectares), ]
```

```
##             community hectares     acres
## 2            Asturias       65     160.6
## 17           Canarias      100     247.1
## 3           Cantabria      440    1087.3
## 13       C.Valenciana     6111   15100.6
## 9            Baleares     7203   17799.0
## 14             Murcia     9500   23475.0
## 11             Madrid    13118   32415.3
## 1             Galicia    18817   46497.8
## 4             P.Vasco    25143   62129.7
## 6            La Rioja    34214   84544.6
## 5             Navarra    66326  163895.1
## 8            Cataluna    74206  183367.0
## 15        Extremadura   143250  353978.5
## 12 Castilla-La Mancha   263424  650934.9
## 7              Aragon   311479  769681.4
## 16          Andalucia   558292 1379569.6
## 10      Castilla-Leon   619858 1531702.5
```

```
nav_num <- which(WHEATSPAIN[order(WHEATSPAIN$hectares), ]$community=="Navarra")
p_nav <- (nav_num - 1) / (length(WHEATSPAIN[order(WHEATSPAIN$hectares), ]$community) -
1)
p_nav
```

```
## [1] 0.625
```

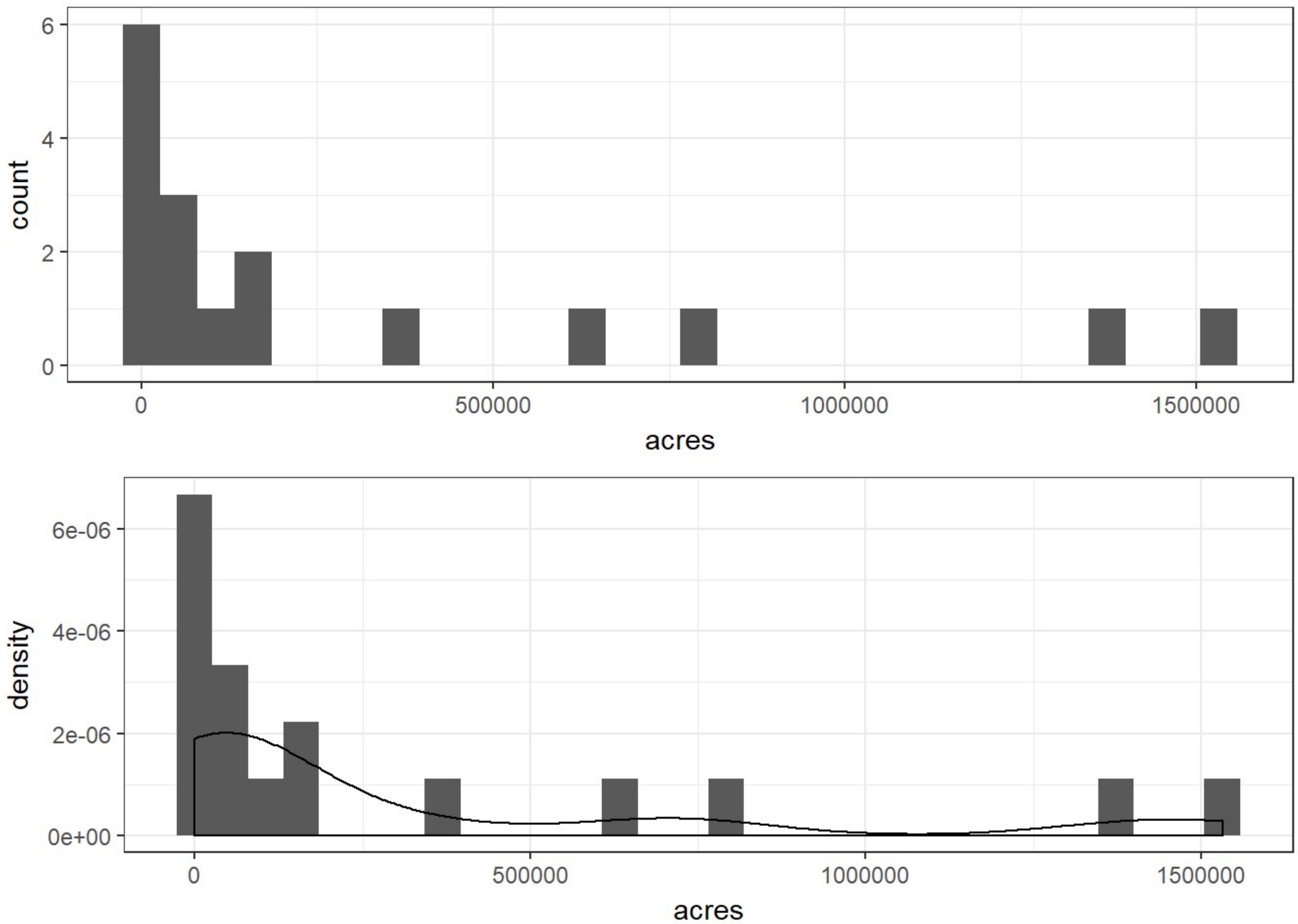
```
quantile(WHEATSPAIN$hectares, probs = p_nav)
```

```
## 62.5%
## 66326
```

d. Create and display in the same graphics device a frequency histogram of the variable acres and a density histogram of the variable acres. Superimpose a density curve over the 2nd histogram.

```
plot1 <- ggplot(data = WHEATSPAIN, aes(x = acres)) + geom_histogram() + theme_bw()
plot2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) + geom_histogram()
+ theme_bw() + geom_density()
multiplot(plot1, plot2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

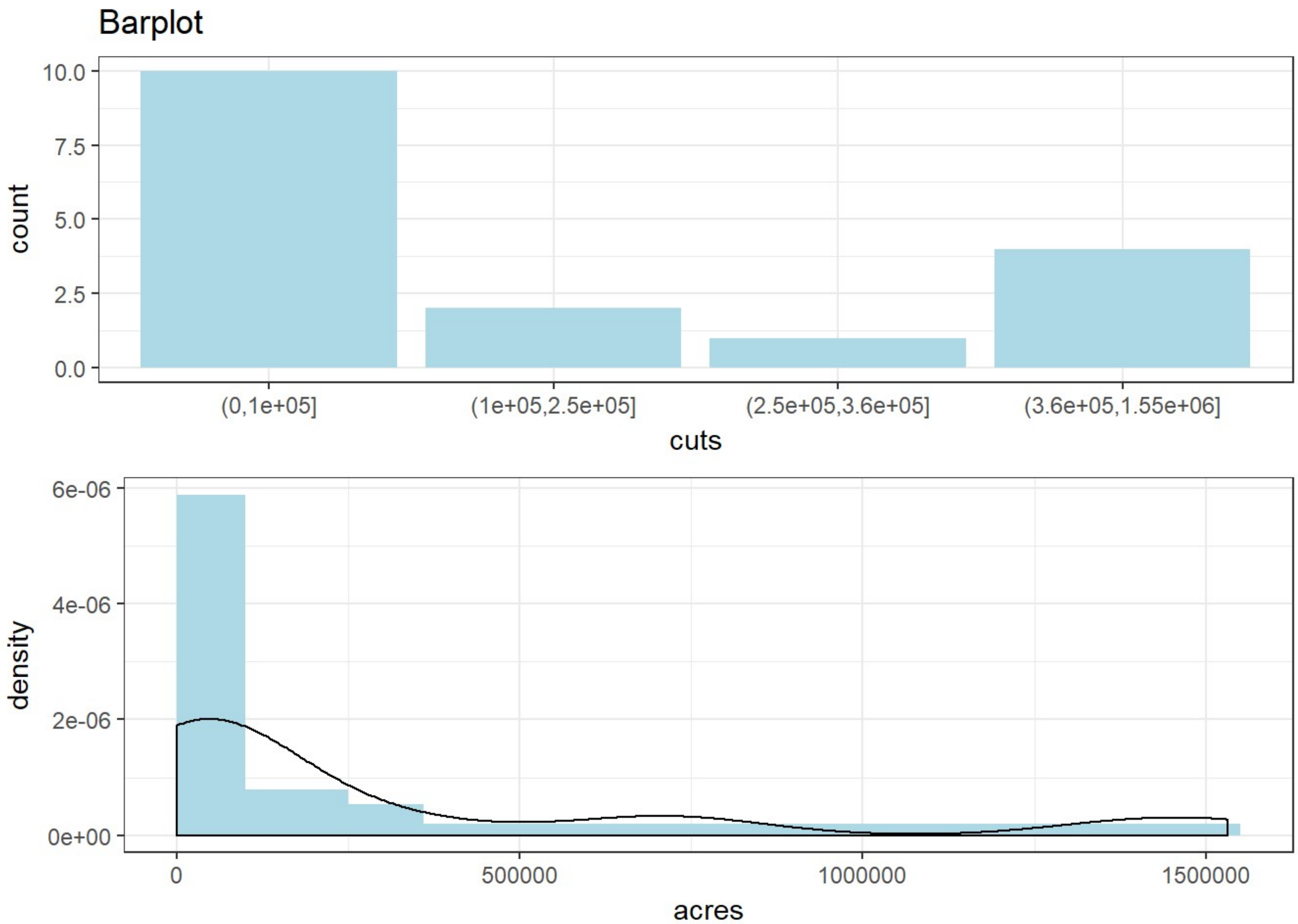


e) Explain why using breaks of 0; 100,000; 250,000; 360,000; and 1,550,000 automatically result in a density histogram.

```
# The breaks used are not equidistant, the default of his() is to then produce a densi
ty
#histogram.
```

f. Create and display in the same graphics device a barplot of acres and a density histogram of acres using break points of 0; 100,000; 250,000; 360,000; and 1,550,000.

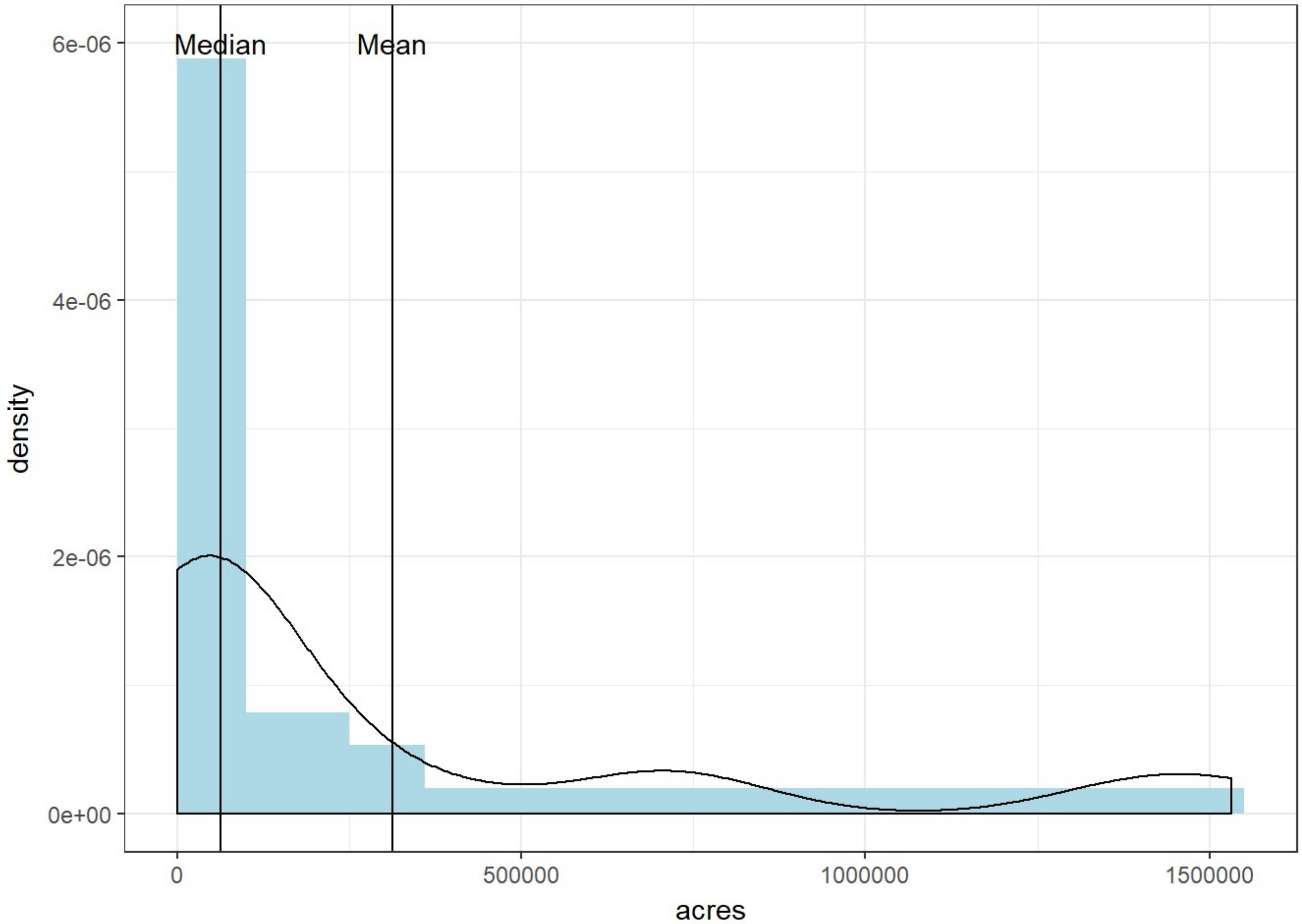
```
bins <- c(0, 100000, 250000, 360000, 1550000)
WHEATSPAIN$cuts <- cut(WHEATSPAIN$acres, breaks = bins)
plot1 <- ggplot(data = WHEATSPAIN, aes(x = cuts)) + geom_bar(fill = "lightblue") + the
me_bw() + labs(title = "Barplot")
plot2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) + geom_histogram(b
reaks = bins, fill = "lightblue") + theme_bw() + geom_density()
multiplot(plot1, plot2, layout = matrix(c(1, 2), byrow = TRUE, ncol = 1))
```



g)

Add vertical lines to the density histogram of acres to indicate the locations of the mean and the median.

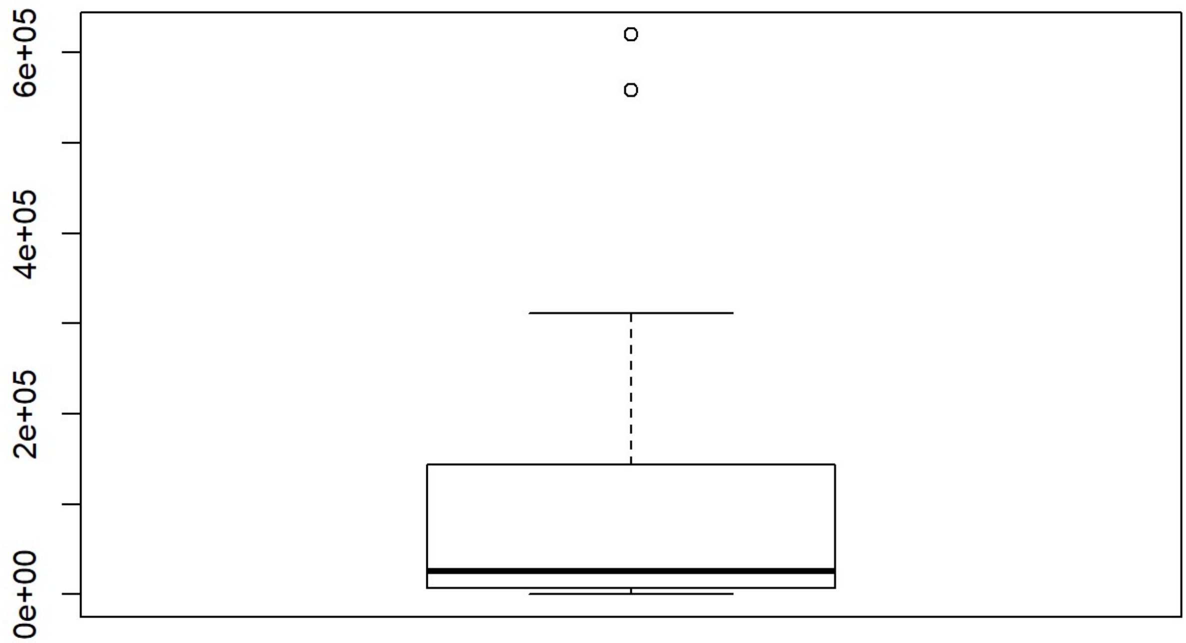
```
plot2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) + geom_histogram(b
reaks = bins, fill = "lightblue") + theme_bw() + geom_density()
plot2 + geom_vline(xintercept = c(median(WHEATSPAIN$acres),+ mean(WHEATSPAIN$acres)))
+ annotate("text", label = "Median", x = median(WHEATSPAIN$acres),y = 6e-06) + annotat
e("text", label = "Mean", x = mean(WHEATSPAIN$acres), y = 6e-06)
```



h)

Create a boxplot of hectares and label the communites that appear as outliers in the boxplot. (Hint: Use identity().)

```
with(data = WHEATSPAIN, boxplot(hectares))
with(data = WHEATSPAIN, identify(rep(1, length(hectares)), hectares, labels = communit
y))
```




```
## integer(0)
```

i. Determine the community with the largest harvested wheat surface area using either acres or hectares. Remove the community from the data frame and compute the mean, median, and standard deviation of hectares. How do these values compare to the values for these statistics computed in part (a)?

```
remove_CastillaLeon <- WHEATSPAIN[-10, ]
mean(WHEATSPAIN$hectares)
```

```
## [1] 126561.5
```

```
mean(remove_CastillaLeon$hectares)
```

```
## [1] 95730.5
```

```
median(WHEATSPAIN$hectares)
```

```
## [1] 25143
```

```
median(remove_CastillaLeon$hectares)
```

```
## [1] 21980
```

```
sd(WHEATSPAIN$hectares)
```

```
## [1] 197319.1
```

```
sd(remove_CastillaLeon$hectares)
```

```
## [1] 155864.7
```

The mean, median, and standard deviation are all smaller than those from part (a) where Castilla-Leon was included.