STAT 500: HW1

Jasmine Mou

9/18/2017

1. Classify the following variables as binary, discrete, or continuous. Also classify them as qualitative or quantitative. Example: Age in years. Answer: Continuous and quantitative

(a) Time in terms of AM or PM.
binary, as the time should either be in AM or in PM, and quanlitative.

(b) Angles as measured in degrees between 0 and 360.
continuous and quantitative.

(c) Bronze, Silver, and Gold medals as awarded at the Olympics.
discrete and qualitative.

(d) Height above sea level.
continous and quantitative.

(e) Number of patients in a hospital.
discrete and quantitative.

2. Do exploratory data analysis for "Pima Indian" data.

```
library(faraway)
data(pima)
```

(a) Find the dimension of the data (p and n).

```
dim <- dim(pima)
n <- dim[1]
p <- dim[2]
```

Thus the dimension of data is $n = 768$ and $p = 9$.

(b) Find the numerical summaries of each variables (min, 1Q, median, mean, 3Q, max).

```
t_summary <- apply(pima, 2, function(x) summary(x))
library(knitr)
kable(t_summary)
```

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
Min.	0.000	0.0	0.00	0.00	0.0	0.00	0.0780	21.00	0.000
1st Qu.	1.000	99.0	62.00	0.00	0.0	27.30	0.2438	24.00	0.000
Median	3.000	117.0	72.00	23.00	30.5	32.00	0.3725	29.00	0.000
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24	0.349
3rd Qu.	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41.00	1.000
Max.	17.000	199.0	122.00	99.00	846.0	67.10	2.4200	81.00	1.000

(c) Change some of the 0s to **NA** and find the numerical summaries of each variables again.
Here we change `glucose`, `triceps`, and `insulin`'s 0s to NA.

```
pima$glucose[pima$glucose==0] = NA
pima$triceps[pima$triceps==0] = NA
pima$insulin[pima$insulin==0] = NA
```

In the new summary, we notice the update of `Min.` values for these 3 variables.

```
t_summary <- apply(pima, 2, function(x) summary(na.omit(x)))
kable(t_summary)
```

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
Min.	0.000	44.0	0.00	7.00	14.00	0.00	0.0780	21.00	0.000
1st Qu.	1.000	99.0	62.00	22.00	76.25	27.30	0.2438	24.00	0.000
Median	3.000	117.0	72.00	29.00	125.00	32.00	0.3725	29.00	0.000
Mean	3.845	121.7	69.11	29.15	155.50	31.99	0.4719	33.24	0.349
3rd Qu.	6.000	141.0	80.00	36.00	190.00	36.60	0.6262	41.00	1.000
Max.	17.000	199.0	122.00	99.00	846.00	67.10	2.4200	81.00	1.000

(d) Which observation has the largest "diastolic blood pressure"? Give the row number.

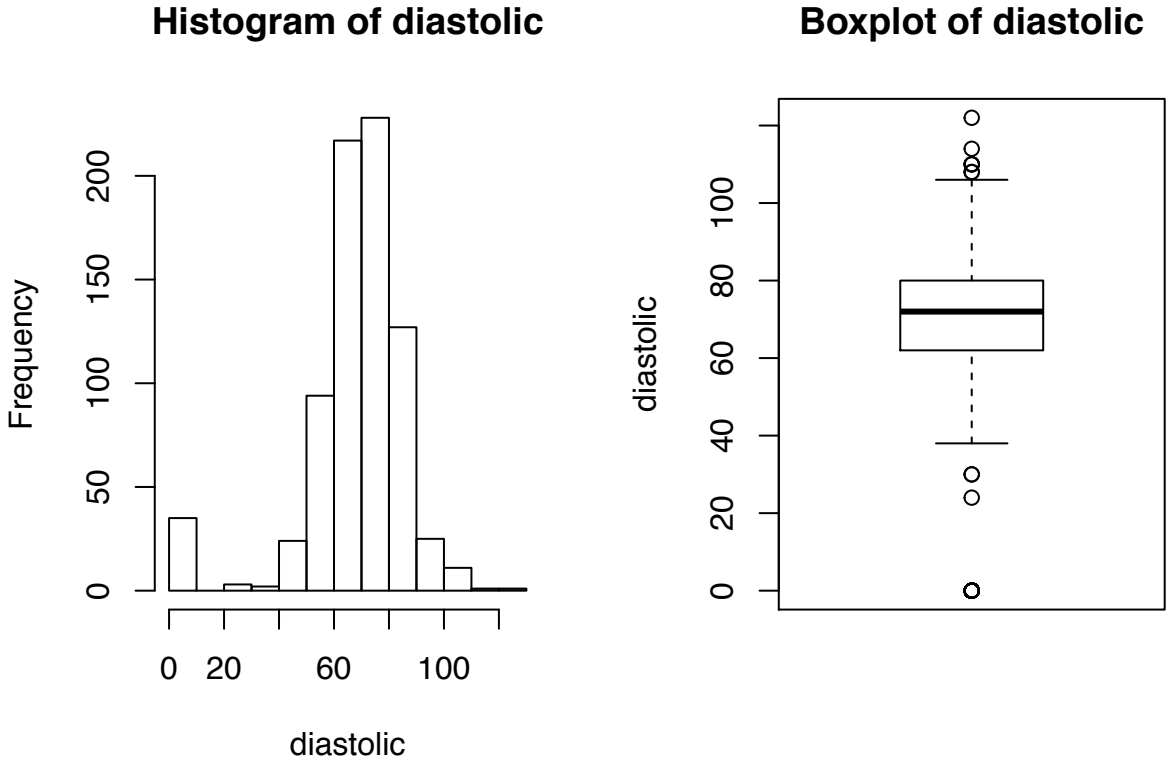
```
index_max_diastolic <- pima$diastolic==max(pima$diastolic, na.rm=TRUE)
pima[index_max_diastolic,]
```

```
##     pregnant glucose diastolic triceps insulin  bmi diabetes age test
## 107        1      96       122      NA      NA 22.4    0.207  27    0
```

The row number is 107.

(e) Draw the graphical summaries (histogram and boxplot) for a single variable "diastolic blood pressure". **(hint: you may want to use *par(mfrow = c(1,2))*)**

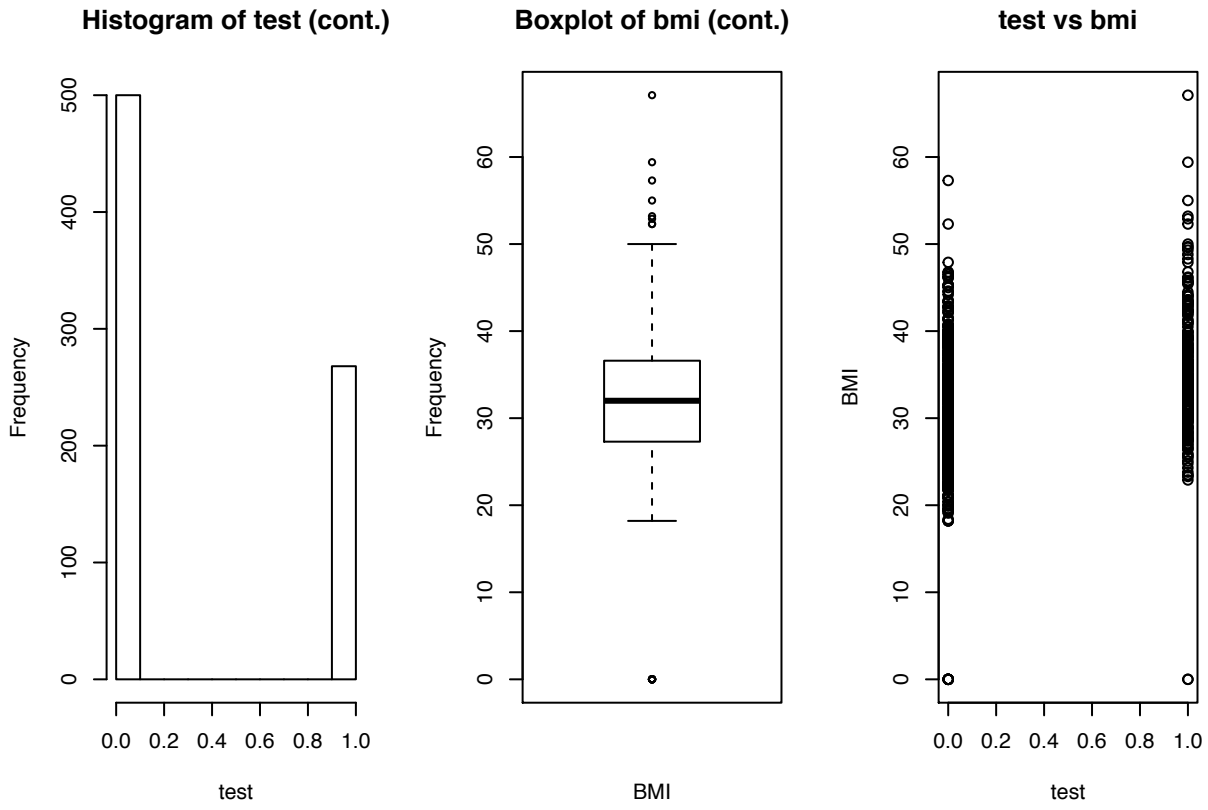
```
par(mfrow=c(1,2))
hist(pima$diastolic, xlab="diastolic", ylab="Frequency", main="Histogram of diastolic")
boxplot(pima$diastolic, ylab="diastolic", main="Boxplot of diastolic")
```



(f) Draw the graphical summaries (histogram and boxplot) for two variables; "test" and "bmi". You must present at least one graph showing the relationship between two variables. **(hint: you may want to use *par(mfrow = c(1,3)) and factor()*)**

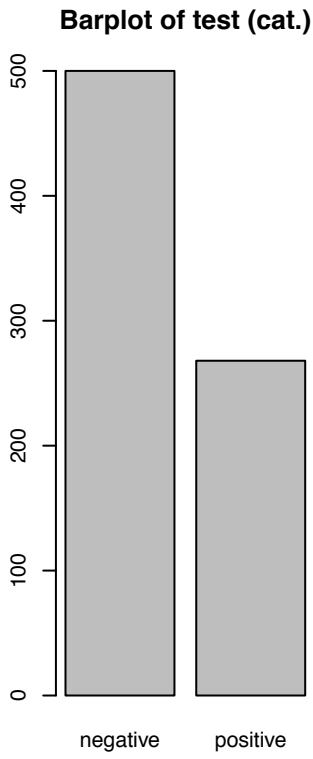
When viewing `test` *as continuous variables (0,1), we can summarize it with histogram. However, the clustered representation of* `test` *and* `bmi` *is not clear enough.*

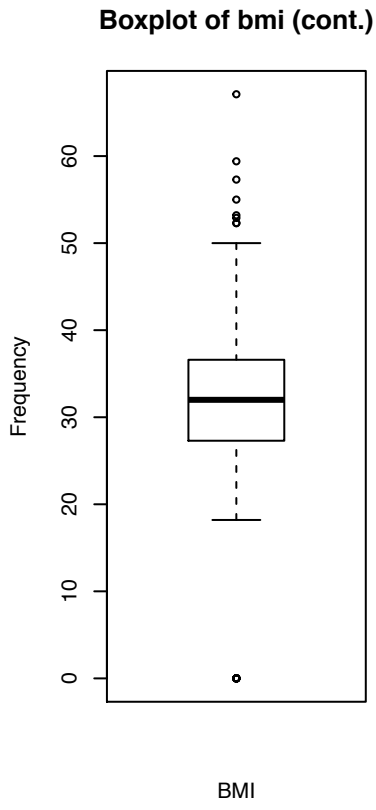
```
par(mfrow=c(1,3))
hist(pima$test, xlab="test", ylab="Frequency", main="Histogram of test (cont.)")
boxplot(pima$bmi, xlab="BMI", ylab="Frequency", main="Boxplot of bmi (cont.)")
plot(x=pima$test, y=pima$bmi, xlab="test", ylab="BMI", main="test vs bmi")
```

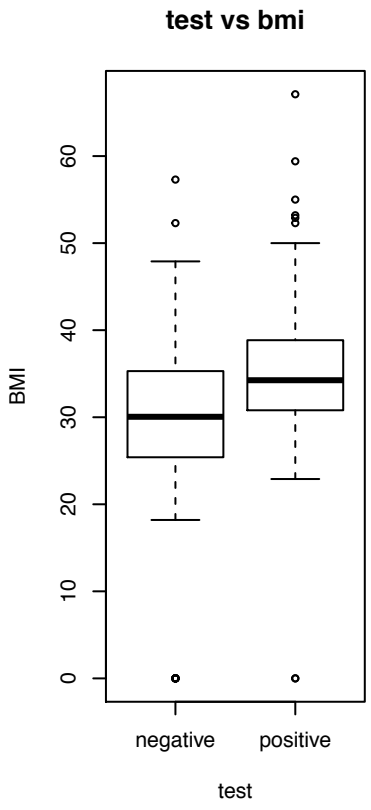


Thus we factorize the `test` *first before summarize the relationship between* `test` *and* `bmi`*. When* `test` *is converted to categorical variable rather than the original continous variable, histogram doesn't apply in representation of single variable* `test`*, and we use barplot instead.* **reference link**

```
par(mfrow=c(1,3))
pima$test=factor(pima$test, labels=c("negative", "positive"))
barplot(table(pima$test), main="Barplot of test (cat.)")
boxplot(pima$bmi, xlab="BMI", ylab="Frequency", main="Boxplot of bmi (cont.)")
plot(x=pima$test, y=pima$bmi, xlab="test", ylab="BMI", main="test vs bmi")
```


Barplot of test (cat.)
0
100
200
300
400
500
negative
positive


Boxplot of bmi (cont.)
Frequency
0
10
20
30
40
50
60
BMI


test vs bmi
BMI
0
10
20
30
40
50
60
negative
positive
test

Attachment: RMarkdown Codes

````
---
title: 'STAT 500: HW1'
author: "Jasmine Mou"
date: "9/18/2017"
output:
  pdf_document: default
---

1. Classify the following variables as binary, discrete, or continuous.
Also classify them as qualitative or quantitative. Example: Age in
years. Answer: Continuous and quantitative
(a) Time in terms of AM or PM.
*binary, as the time should either be in AM or in PM, and quanlitative.
*

(b) Angles as measured in degrees between 0 and 360.
*continuous and quantitative.*

(c) Bronze, Silver, and Gold medals as awarded at the Olympics.
*discrete and qualitative.*

(d) Height above sea level.
*continous and quantitative.*

(e) Number of patients in a hospital.
*discrete and quantitative.*


2. Do exploratory data analysis for ”Pima Indian” data.
```{r init}
library(faraway)
data(pima)
```

(a) Find the dimension of the data (p and n).
```{r (a)}
dim <- dim(pima)
n <- dim[1]
p <- dim[2]
```
*Thus the dimension of data is n = `r n` and p = `r p`.*

(b) Find the numerical summaries of each variables (min, 1Q, median,
mean, 3Q, max).
```{r (b), warning=FALSE}
t_summary <- apply(pima, 2, function(x) summary(x))
library(knitr)
kable(t_summary)
````
````

```

(c) Change some of the 0s to **NA** and find the numerical summaries of each variables again.
*Here we change `glucose`, `triceps`, and `insulin`'s 0s to NA.*
```{r (c)}
pima$glucose[pima$glucose==0] = NA
pima$triceps[pima$triceps==0] = NA
pima$insulin[pima$insulin==0] = NA
```
In the new summary, we notice the update of `Min.` values for these 3 variables.
```{r (c) cont}
t_summary <- apply(pima, 2, function(x) summary(na.omit(x)))
kable(t_summary)
```

(d) Which observation has the largest ”diastolic blood pressure”? Give the row number.
```{r (d)}
index_max_diastolic <- pima$diastolic==max(pima$diastolic, na.rm=TRUE)
pima[index_max_diastolic,]
```
The row number is `r which(index_max_diastolic)`.

(e) Draw the graphical summaries (histogram and boxplot) for a single variable ”diastolic blood pressure”. **(hint: you may want to use *par(mfrow = c(1,2))*)**
```{r (e)}
par(mfrow=c(1,2))
hist(pima$diastolic, xlab="diastolic", ylab="Frequency", main="Histogram of diastolic")
boxplot(pima$diastolic, ylab="diastolic", main="Boxplot of diastolic")
```

(f) Draw the graphical summaries (histogram and boxplot) for two variables; ”test” and ”bmi”. You must present at least one graph showing the relationship between two variables. **(hint: you may want to use *par(mfrow = c(1,3)) and factor()*)**
When viewing `test` as continuous variables (0,1), we can summarize it with histogram. However, the clustered representation of `test` and `bmi` is not clear enough.
```{r (f)}
par(mfrow=c(1,3))
hist(pima$test, xlab="test", ylab="Frequency", main="Histogram of test (cont.)")
boxplot(pima$bmi, xlab="BMI", ylab="Frequency", main="Boxplot of bmi (cont.)")
plot(x=pima$test, y=pima$bmi, xlab="test", ylab="BMI", main="test vs bmi")
```
```

Thus we factorize the `test` first before summarize the relationship between `test` and `bmi`. When `test` is converted to categorical variable rather than the original continous variable, histogram doesn't apply in representation of single variable `test`, and we use barplot instead. **[reference link](https://stackoverflow.com/a/24600821)**

```{r (f) cont}
par(mfrow=c(1,3))
pima$test=factor(pima$test, labels=c("negative", "positive"))
barplot(table(pima$test), main="Barplot of test (cat.)")
boxplot(pima$bmi, xlab="BMI", ylab="Frequency", main="Boxplot of bmi (cont.)")
plot(x=pima$test, y=pima$bmi, xlab="test", ylab="BMI", main="test vs bmi")
```