Group 5, R Project 3

Kera Whitley, Laura Mathews, Nataliya Peshekhodko

2020-April-23

```
library(dplyr)
library (ggplot2)
library (gridExtra)
library (EnvStats)
```

First part: hypothesis test for the observed data

1.1. Read data

Read the data from txt file and save it in dataframe.

```
data <- read.delim("mpg.txt", header=TRUE, sep=" ")
```

1.2. Two sample t-test

Conduct two sample t-test twice: with equal variance and with unequal veriance. T-test with equal variance:

```
# Apply equal variance t-test
equal.var.test <- t.test( MPG~Country, data = data, var.equal=TRUE, conf.level=0.95)
equal.var.test
```

```
##
##  Two Sample t-test
##
## data:  MPG by Country
## t = 12.238, df = 326, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.454006 11.692605
## sample estimates:
## mean in group Japan    mean in group US
##            22.35443            12.28112
```

Based on the equal variance t.test **p-value**=$1.3562176 \times 10^{-28}$ is small (less then 5% cut off), we can reject the null hypothesis.

T-test with unequal variance:

```r
# Apply unequal variance t-test
unequal.var.test <- t.test( MPG~Country, data = data, var.equal=FALSE, conf.level = 0.95)
unequal.var.test
```

```
##
##  Welch Two Sample t-test
##
## data:  MPG by Country
## t = 12.99, df = 145.45, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.540639 11.605972
## sample estimates:
## mean in group Japan    mean in group US
##            22.35443            12.28112
```

Based on the unequal variance t.test **p-value**=$4.2477785 \times 10^{-26}$ is small (less then 5% cut off), we also can reject the null hypothesis.

1.3. How well the normality assumption met by the data

To test data for normality we will use 3 methods:

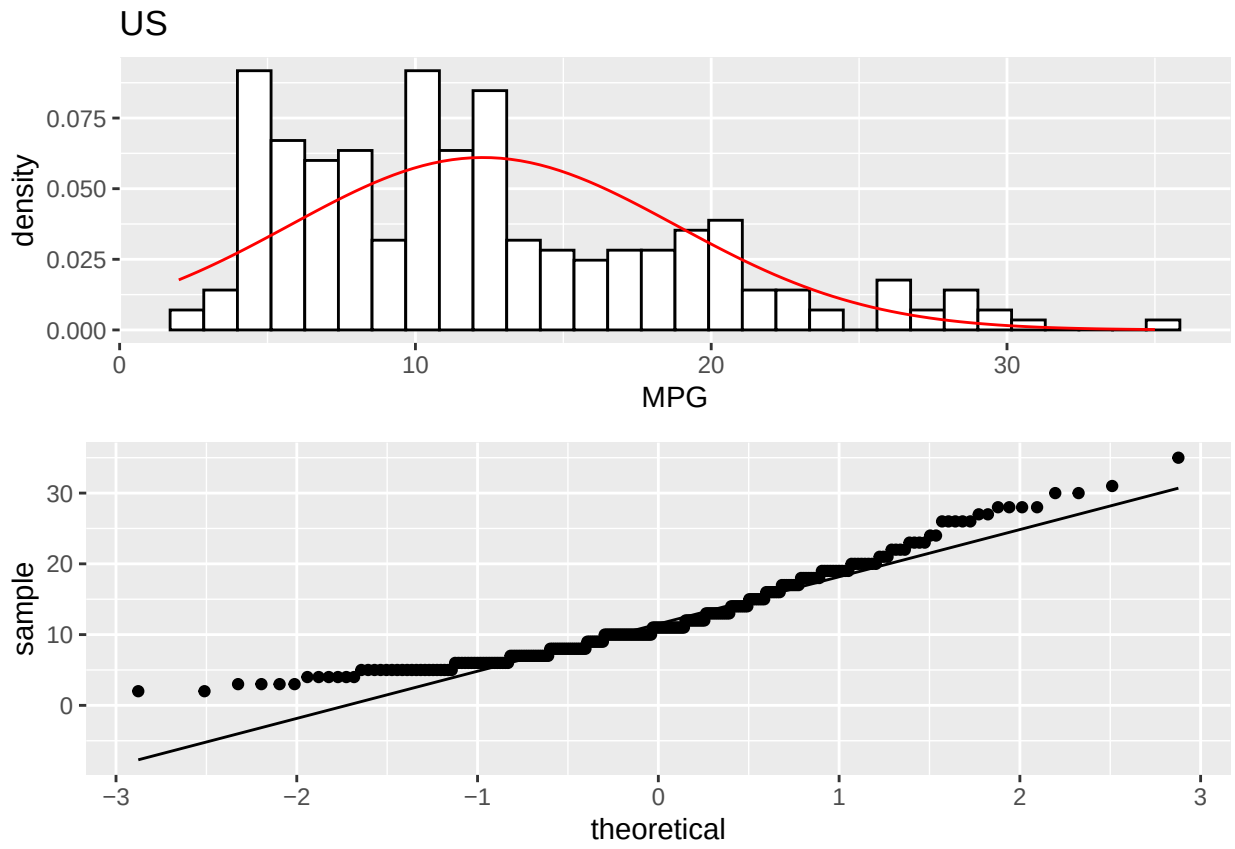
- Compare histogram of the data to a normal probability curve for the data
- Quantile-quantile plot
- Goodness of fit

First, split up data in two data sets based on the country.

```r
# Split up data based on the country name
us <- filter(data, Country == "US")
japan <- filter(data, Country == "Japan")
```

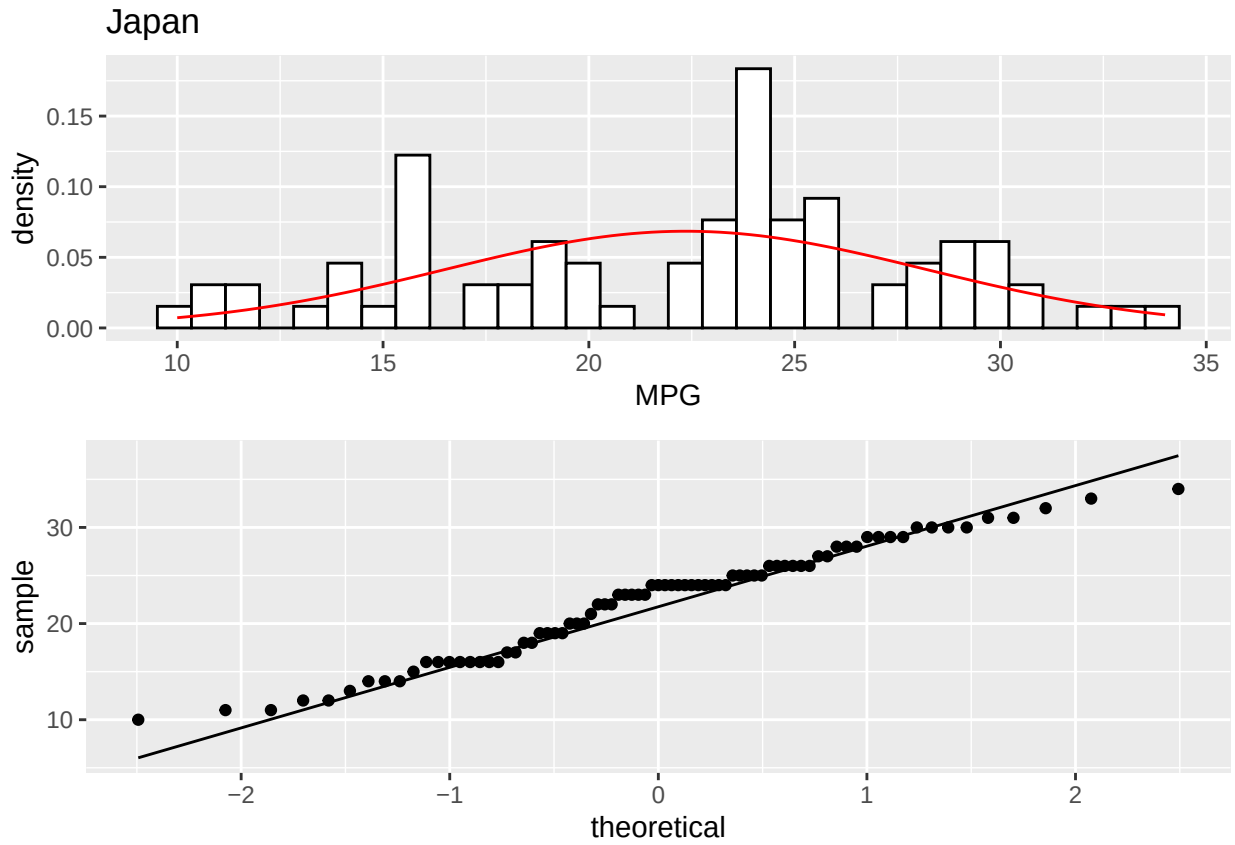
Plot histogram with overlaping normal curve and quantile-quantile plot for the US data.

```r
# Plot histogram
hist_us<- ggplot (us, aes(MPG))+geom_histogram(aes(y=..density..),colour="black", fill = "white") +
  ggtitle("US") + stat_function(fun=dnorm, args = list (mean = mean(us$MPG), sd=sd(us$MPG)),
                                colour = "red")
# Plot quantile-quantile plot
qq_us <- ggplot(us, aes(sample = MPG))+stat_qq()+stat_qq_line()
# Arrange two plots in one grid
grid.arrange(hist_us, qq_us, ncol=1, nrow=2)
```



Plot histogram with overlaping normal curve and quantile-quantile plot for the Japan data.

```
# Plot histogram
hist_jpn <- ggplot (japan, aes(MPG))+geom_histogram(aes(y=..density..),colour="black", fill = "white") +
  ggtitle ("Japan") + stat_function(fun=dnorm, args = list (mean = mean(japan$MPG), sd=sd(japan$MPG)),
                                    colour = "red")
# Plot quantile-quantile plot
qq_jpn <- ggplot(japan, aes(sample = MPG))+stat_qq()+stat_qq_line()
# Arrange two plots in one grid
grid.arrange(hist_jpn, qq_jpn, ncol=1, nrow=2)
```



Based on the graphical methods, data does not fit normality well.

Apply goodness of fit test to the data:

```
# Goodness of fit test for us and japan data
gofTest(us, distribution = "norm")
```

```
##
## Results of Goodness-of-Fit Test
## -------------------------------
##
## Test Method:                     Shapiro-Wilk GOF
##
## Hypothesized Distribution:       Normal
##
## Estimated Parameter(s):          mean = 7.140562
##                                  sd   = 6.913981
##
## Estimation Method:               mvue
##
## Data:                            us
##
## Sample Size:                     498
##
## Test Statistic:                  W = 0.7699273
##
```

```
## Test Statistic Parameter:        n = 498
##
## P-value:                         0
##
## Alternative Hypothesis:          True cdf does not equal the
##                                  Normal Distribution.
```

```
gofTest(japan, distribution = "norm")
```

```
##
## Results of Goodness-of-Fit Test
## -------------------------------
##
## Test Method:                     Shapiro-Wilk GOF
##
## Hypothesized Distribution:       Normal
##
## Estimated Parameter(s):          mean = 11.67722
##                                  sd   = 11.47152
##
## Estimation Method:               mvue
##
## Data:                            japan
##
## Sample Size:                     158
##
## Test Statistic:                  W = 0.788144
##
## Test Statistic Parameter:        n = 158
##
## P-value:                         7.538414e-14
##
## Alternative Hypothesis:          True cdf does not equal the
##                                  Normal Distribution.
```

Small p-values for both data mean that we can reject the null hypothesis (H0 - the data is consistent with normal distribution) and both data sets are not consistent with Normal distribution.

1.4. Preferable test for selected dataset

Our samples have unequal sizes, so unequal variance test is more preferable here.

We onserved that the data is violate the normality assumption, there are two options:

- We can transform our data so that the data becomes normally ditributed
- Run non-parametric test (Mann-Whitney U test) that does not require the assumption of normality.