# Tutorial 1 Solutions

*STAT 3013/8027*

---

1. Rice: 2.31, 3.18. **Ans.** See the handwritten pages.
2. Consider the following:

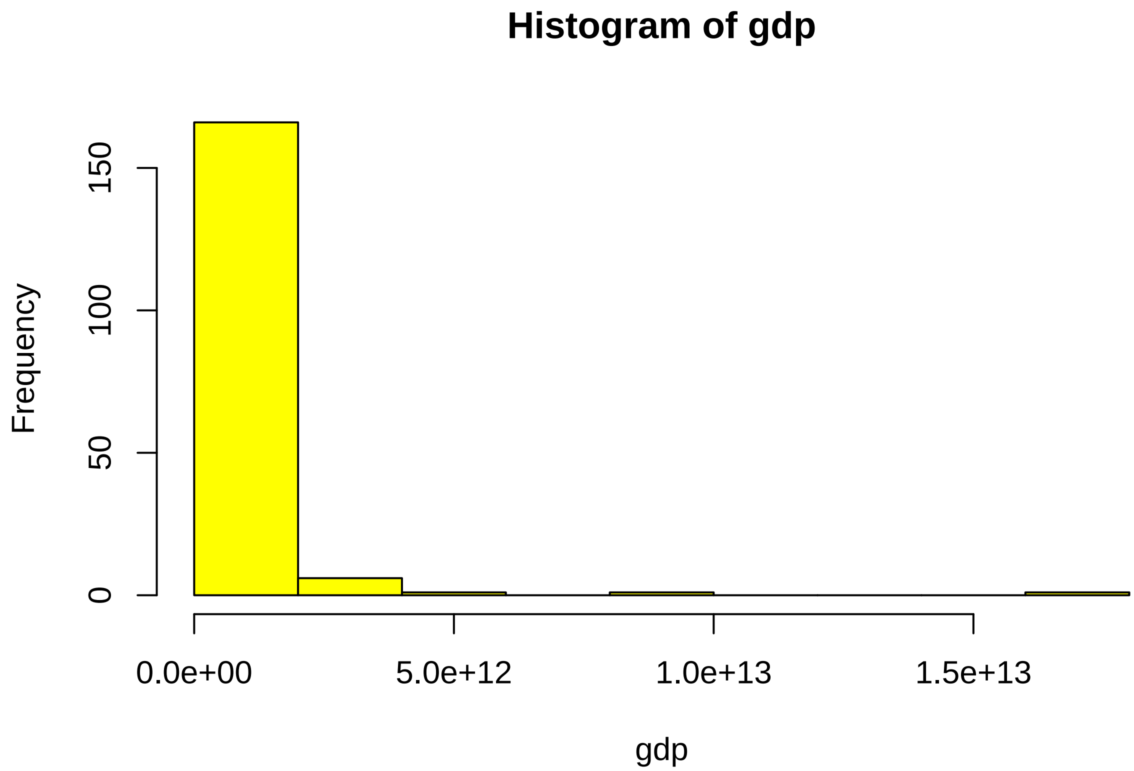
   a. Visually display the data and discuss. Try taking the natural log of the data (when statisticians say "log" they mean natural log).
   b. Compute a six number summary of the data.
   c. Based on the "box plot rule", determine if there are any outliers. Which countries are outliers? To use the rule examine the following: Are any values in the data below the $1^{\text{st}}$ Quartile - 1.5 IQR? Are any values in the data above the $3^{\text{rd}}$ Quartile + 1.5 IQR? IQR is the inter-quartile range.
   d. Let $Y = \log(\text{GDP})$. Suppose $Y \sim \text{normal}(\mu, \sigma^2)$. What is your best guess for $\mu$ and $\sigma^2$ as functions of $Y$ (call these $T_1$ and $T_2$)? What are the means (expected values) of $T_1$ and $T_2$?

**Ans.** First let's load in the data. Note that I removed the missing values (NA's). Missing data is an extensive and important topic in statistics; as such be careful about removing missing data. At the very least, discuss your full sample of data and then which cases were removed due to missing data.

```r
## GDP
gdp2013 <- read.table("gdp2013.txt", header=TRUE)
D <- gdp2013
D <- na.omit(D)
gdp <- D$Y2013
```

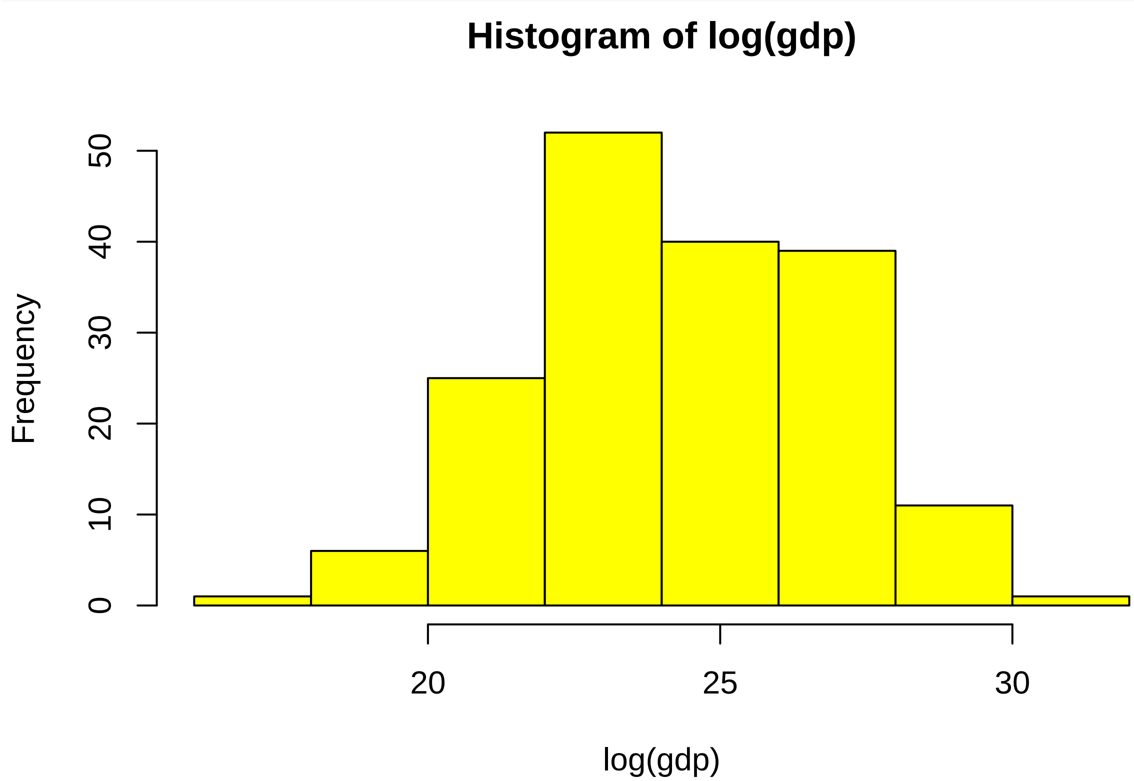
The histogram is unimodal, right skewed and appears to have some outliers.

```r
## visual display
hist(gdp, col="yellow")
```




If we take the log of the data it is far more symmetric.

```
## visual display
hist(log(gdp), col="yellow")
```



Let's get a six number summary (using the logged data)

```
log.gdp <- log(gdp)
summary(log.gdp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.46   22.73   24.15   24.23   26.13   30.45
```

Let's get the quartiles and the IQR.

```
##
Q <- quantile(log.gdp, prob=c(0.25, 0.5, 0.75), type=6)
Q
```

```
##      25%      50%      75%
## 22.71874 24.14522 26.13676
```

```
IQR <- unname(Q[3]-Q[1])
IQR
```

```
## [1] 3.418021
```

Let's identify outliers based on the box plot method.

```
##
high.outliers <- log.gdp[log.gdp > Q[3] + 1.5*IQR]
low.outliers <- log.gdp[log.gdp < Q[1] - 1.5*IQR]

high.outliers
```

```
## numeric(0)
```

```
low.outliers
```

```
## [1] 17.45664
```

```
# let's find the countries that have high outliers
D[log.gdp >= min(high.outliers),1]
```

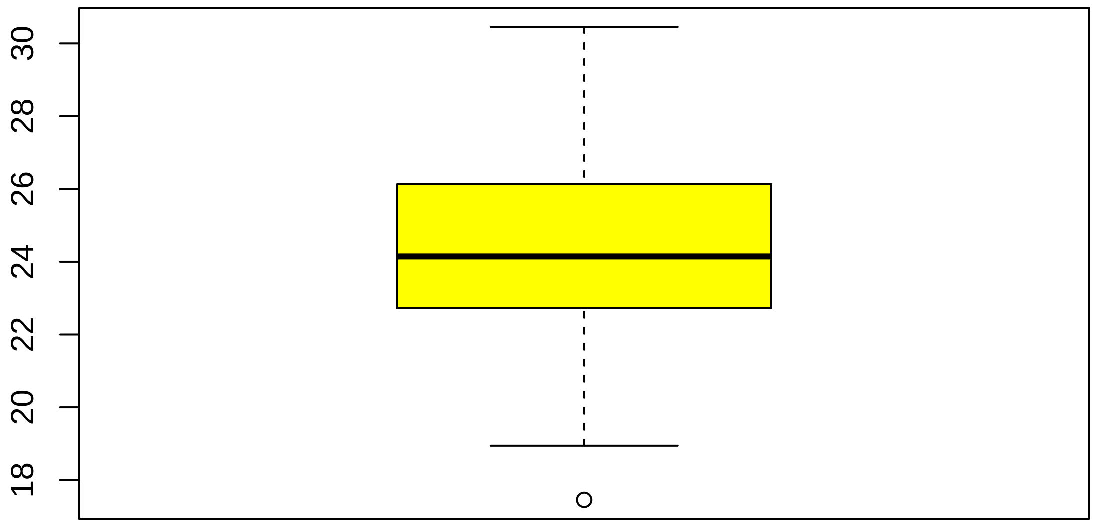
```
## Warning in min(high.outliers): no non-missing arguments to min; returning
## Inf
```

```
## factor(0)
## 214 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
# let's find the countries that have high outliers
D[log.gdp <= max(low.outliers),1]
```

```
## [1] Tuvalu
## 214 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
# let's make a boxplot.
boxplot(log.gdp, col="yellow")
```




- If we assume the data $Y_1, \ldots, Y_n \sim iid\ \ n(\mu, \sigma^2)$ then reasonable guesses for the population mean and variance are the sample mean and variances. So we have:

$$\bar{y} = 24.2263843 = \hat{\mu}$$

$$S^2 = 6.0934187 = \hat{\sigma}^2$$

A nice property of these estimators is that $E[\hat{\mu}] = E[\bar{Y}] = \mu$ and $E[\hat{\sigma}^2] = E[S^2] = \sigma^2$. We proved the latter in lecture and for fun let's do the former:

$$\begin{aligned} E[\hat{\mu}] = E[\bar{Y}] &= E[(1/n)(Y_1 + \cdots + Y_n)] \\ &= (1/n)E[(Y_1 + \cdots + Y_n)] \\ &= (1/n)(E[Y_1] + \cdots + E[Y_n]) \\ &= (1/n)nE[Y_1] = (1/n)n\mu = \mu \end{aligned}$$