Set 2: Sections 2.1, 2.2, 2.3, 2.4

Histograms:

- **a graphical descriptive statistic**
- **applicable given univariate data** $x_1, \ldots, x_n$
- **able to observe centrality, shape(skewness, symmetry), dispersion, outliers**
- **we encourage intervals of equal length**
- **generated by statistical software, Excel**

Histograms:

- **data are weights of students in kg: 47, 55, 79, 63, 64, 67, 54, 59, 58, 84, 70, 61, 65, 59**

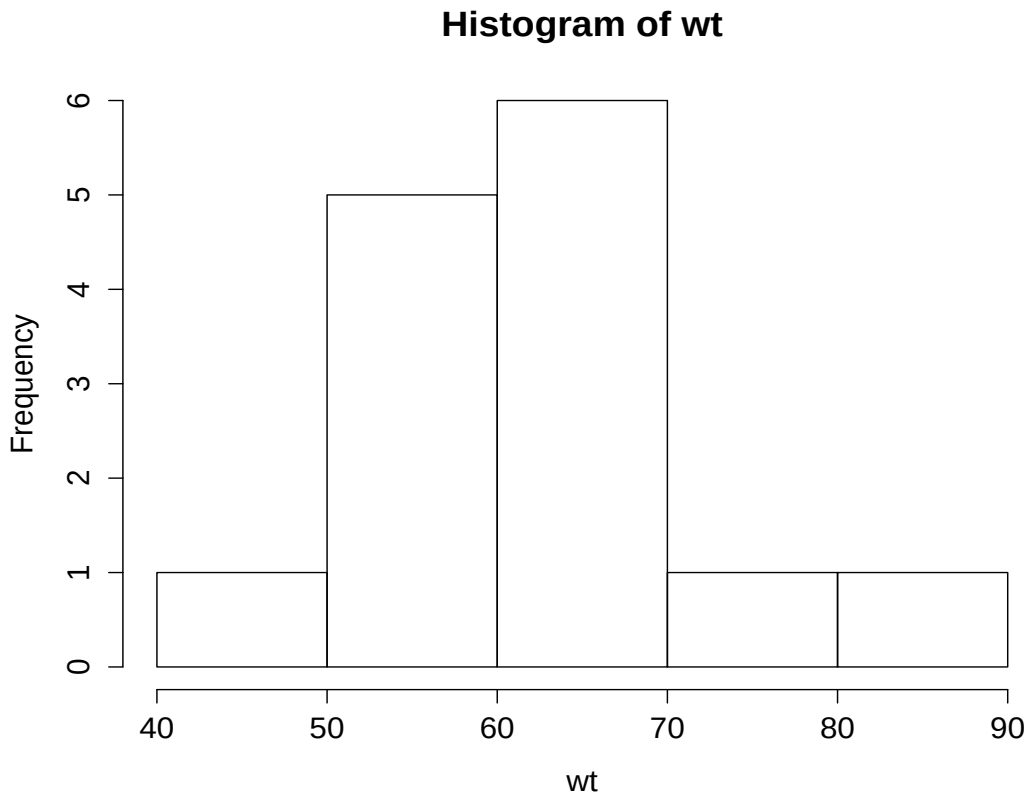


Figure 1: Histogram of weights of students

```
wt = c(47, 55, 79, 63, 64, 67, 54, 59, 58, 84, 70, 61, 65, 59)
hist(wt)
```

Issues in constructing histograms:

- **always label axes and provide a title**
- **be aware of the scale of the vertical axis**

Sample mean $\bar{x}$:

- **a numerical descriptive statistic of centrality**
- **applicable given univariate data** $x_1, \ldots, x_n$
- $\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x_i}{n}$

Sample median $\tilde{x}$:

- **a numerical descriptive statistic of centrality**
- **applicable given univariate data** $x_1, \ldots, x_n$
- $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \textbf{if } n \textbf{ odd} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right)/2 & \textbf{if } n \textbf{ even} \end{cases}$

```
> mean(wt); median(wt)
[1] 63.21429
[1] 62
```

egs.

The median is more *robust* than the mean wrt outliers.

Consider a sample of n house prices:

- $\bar{x} = \$850,000$
- $\tilde{x} = \$700,000$
- **Why do the statistics differ?**

Exercise: Can you approximate the median and mean from a histogram?

Variability (dispersion) in data:

- **Consider the following two datasets**
  - **Dataset 1: -2, -1, 0, 1, 2**
  - **Dataset 2: -300, -100, 0, 100, 300**

Sample range R:

- **a numerical descriptive statistic of variability**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $R = x_{(n)} - x_{(1)}$
- **based on only two data values**

Sample variance s^2:

- **a numerical descriptive statistic of variability**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(\sum x_i^2) - n\bar{x}^2}{n-1}$
- $s^2 \geq 0$; $s^2 = 0$ **corresponds to** $x_1 = \cdots = x_n$
- **large s^2 corresponds to widely spread data**
- **note that denominator is $n-1$ instead of n**
- **think about why the difference $x_i - x_n$ is squared**
- **distinguish between the two formulae**
- **note that s^2 is measured in squared units**
- **the sample standard deviation is given by s**

Sample variance and standard deviation:

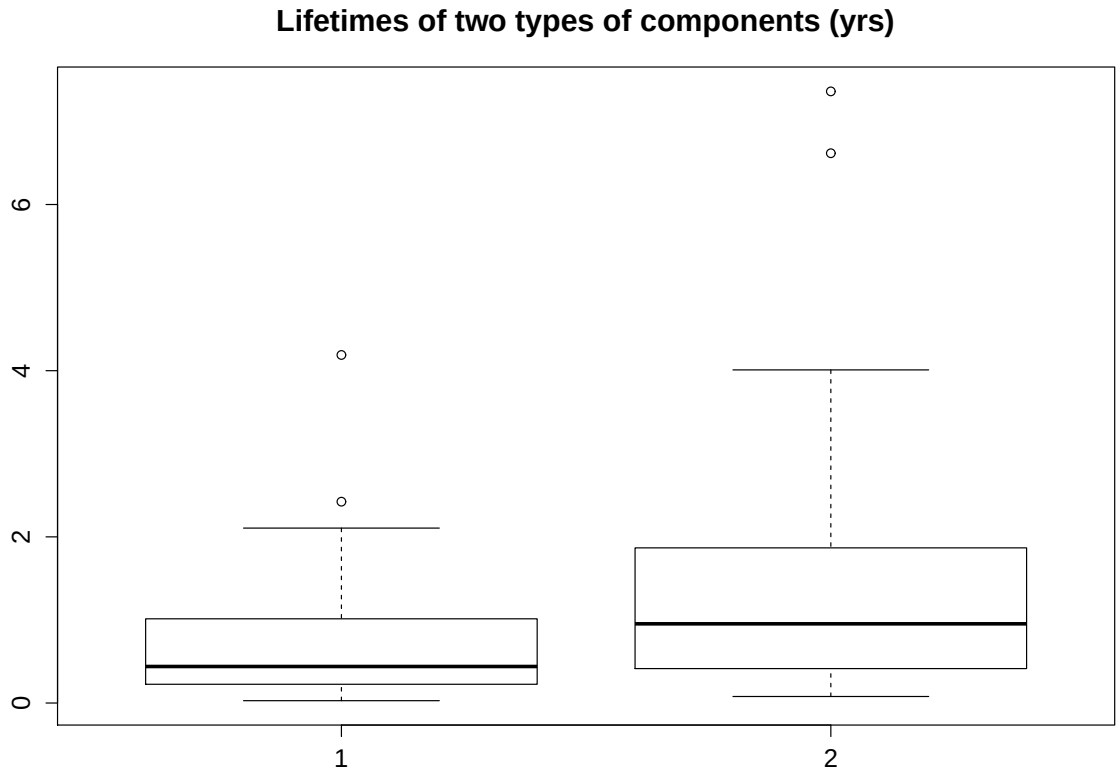
```
D1=c(-2, -1, 0, 1,2)  #Dataset 1
D2=c(-300, -100, 0, 100, 300)  #Dataset 2
> var(D1) #sample variance
[1] 2.5
> sqrt(var(D1))  #sample standard deviation
[1] 1.581139

> var(D2); sqrt(var(D2))
[1] 50000   #sample variance
[1] 223.6068  #sample standard deviation
```

Exercise: Do you get the same answer with your calculator?

Boxplots:

- **a graphical descriptive statistic**
- **applicable given univariate data (in groups)**
- **generated by statistical software**
- **calculations require $\tilde{x}$, lower fourth, upper fourth**
- **interpreting boxplots is our focus**
- **boxplots are not as popular as they should be**



Exercise: How do location/scale changes affect $\bar{x}$ and s^2:

- **i.e.** $x_i \to y_i = a + bx_i$
- **e.g. changing Celsius data to Fahrenheit**