# Stat 427/527: Advanced Data Analysis I

Chapter 1: Summarizing and Displaying Data

August, 2017

# Random variables (r.v.)

- r.v: a variable whose value is subject to variations due to chance.
- two broad categories of r.v.s: **qualitative** and **quantitative**.
  - Qualitative data includes categorical outcomes:
    *Nominal outcome* is one of several categories
    Ex: sex: female and male
    *Ordinal Outcome* is one of several ordered categories.
    Ex: strongly agree, agree, neutral, disagree, strongly disagree.
  - Quantitative data includes numeric outcomes:
    *Discrete Outcome* is one of a fixed set of numerical values.
    Ex: Number of children.
    *Continuous Outcome* is any numerical value.
    Ex: Birthweight.

## Random Sampling and data description

- **Recall**: we are looking at ways to summarize data
  - Numerical summaries:
    measures of center
    (mean, median, mode)
    measures of spread
    (sample variance, range, IQR)
  - Graphical summaries:
    Stem and leaf plots
    Histograms
    Box Plots

## Random Sampling and data description

- **Recall**: we are looking at ways to summarize data
  - Numerical summaries:
    measures of center
    (mean, median, mode)
    measures of spread
    (sample variance, range, IQR)
  - Graphical summaries:
    Stem and leaf plots
    Histograms
    Box Plots

# 6-1 Numerical Summaries

## Definition: Sample Mean

If the $n$ observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6\text{-}1)$$

EX: # earthquakes of magnitude 7 or greater for years 1980-1990:

18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{18 + 14 + \ldots 23}{20} = 14.75$$

## Definition: Median

First we need to order the data

6,7 ,8, 8, 10, 11,11,12,14, 15, 15, 15, 16, 16, 18, 20, 22, 23, 23, 25

and then choose that valued that divides the data in 2 halves.

$$\text{If } n \text{ is even, then } Median = \frac{X_{\frac{n}{2}} + X_{\frac{n+1}{2}}}{2}$$

$$\text{If } n \text{ is odd, then } Median = X_{\frac{n}{2}}$$

Ex: n=20 is even, so Median is (15+15)/2=15

## Definition: Mode

The mode is is the value that occurs the most frequently in a data set or a probability distribution

In our example, hence the mode is 15.

### Remark:

The sample mean is affected by large values in the observations. Hence, if the data are highly skewed, it might not be the best measure to use. Instead, the median is a more robust measure, because it is always half way the data, no matter the value assumed by our observations.

## Measures of spread or variability

### Definition: Sample Variance

If $x_1, x_2, \ldots, x_n$ is a sample of $n$ observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (6\text{-}3)$$

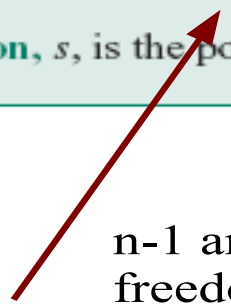
The **sample standard deviation**, $s$, is the positive square root of the sample variance.

## Definition: Sample Variance

If $x_1, x_2, \ldots, x_n$ is a sample of $n$ observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (6\text{-}3)$$

The **sample standard deviation**, $s$, is the positive square root of the sample variance.



n-1 are the degrees of freedom. We loose one degree of freedom for using the sample mean instead of the true mean

## Definition: Sample Variance

If $x_1, x_2, \ldots, x_n$ is a sample of $n$ observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (6\text{-}3)$$

The **sample standard deviation**, $s$, is the positive square root of the sample variance.

In our example, we obtain

$$s^2 = 32.72 \quad s = \sqrt{32.72} = 5.72$$

## Definition: Five Number Summary

| Min | Q_1 | Median | Q_3 | Max |
|---|---|---|---|---|
| 6 | | 15 | | 25 |

## Definition: Five Number Summary

| Min | Q_1 | Median | Q_3 | Max |
|---|---|---|---|---|
| 6 | **10.5** | 15 | **19** | 25 |

**Q_1: First Quartile**

is the median of the first ½ of the data

**Q_2: Second Quartile**

is the median of the 2nd ½ of the data

- Range: $R = \max - \min = 25 - 6 = 19$
- Interquartile Range: $IQR = Q_3 - Q_1 = 19 - 10.5 = 8.5$
- The interquartile range is less sensitive to the extreme values in the sample than is the ordinary sample range

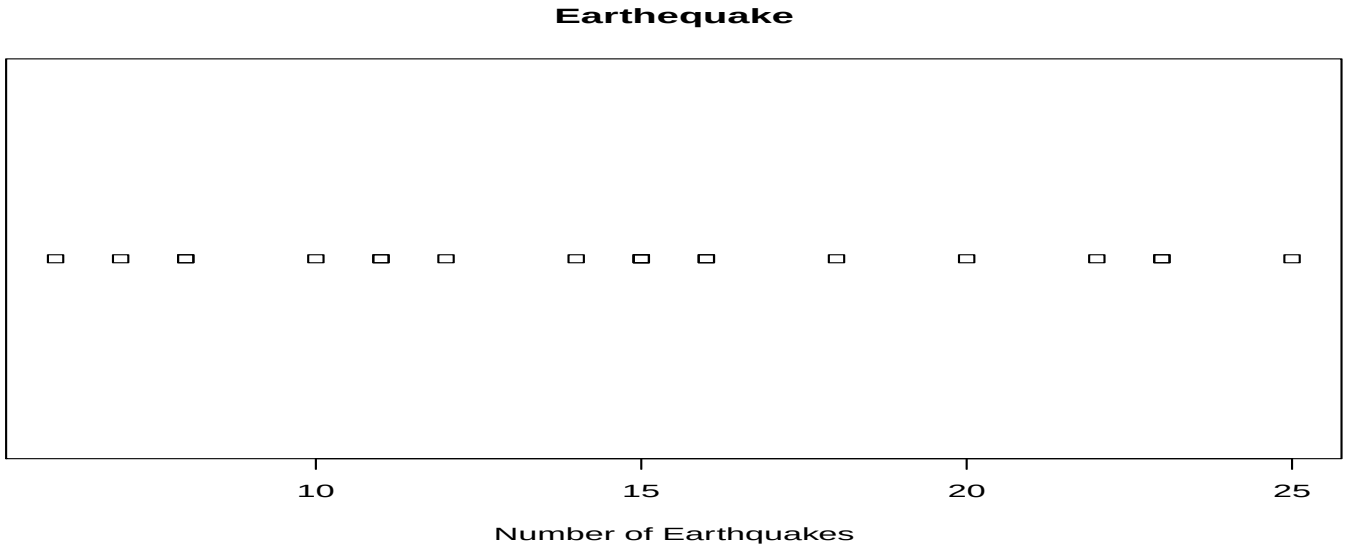
```
> ####Earthquake data,
# earthquakes of magnitude 7 or greater for years 1980-1990:
> eq<-c(18,14,10,15,8,15,6,11,8,7,12,11,23,
            16,15,25,22,20,16,23)
> eq
 [1] 18 14 10 15  8 15  6 11  8  7 12 11 23
      16 15 25 22 20 16 23
> #### mean
> mean(eq)
[1] 14.75
> #### sample variance
> var(eq)
[1] 32.72368
> #### sample standard deviation
> sd(eq)
[1] 5.720462
> #or
> sqrt(var(eq))
[1] 5.720462
```

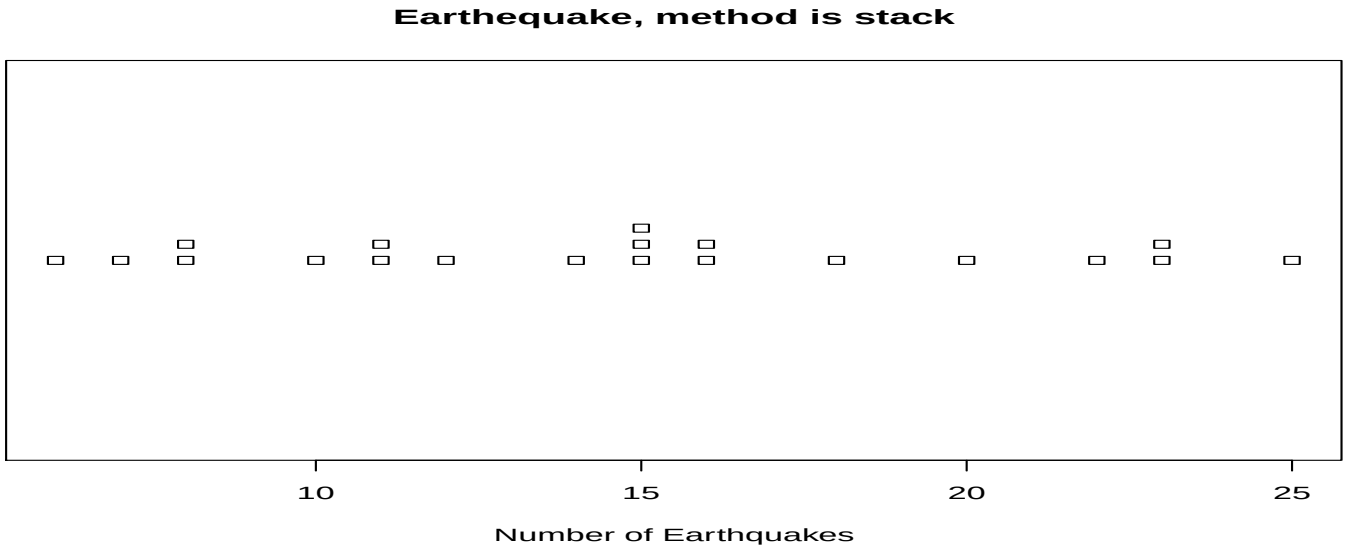
```
> #### sorting
> sort(eq)
 [1]  6  7  8  8 10 11 11 12 14 15 15 15 16 16
      18 20 22 23 23 25
> #### quartiles
> median(eq)
[1] 15
> fivenum(eq)
[1]  6.0 10.5 15.0 19.0 25.0
> ##Range
> fivenum(eq)[5] - fivenum(eq)[1]
[1] 19
> ##IQR
> fivenum(eq)[4] - fivenum(eq)[2]
[1] 8.5
> diff(fivenum(eq)[c(2,4)])
[1] 8.5
```

# Graphical summaries

- Dotplots–The dotplot breaks the range of data into many small-equal width intervals, and counts the number of observations in each interval.

```
#### stripchart-ggplot
# stripchart (dotplot) using R base graphics
# main is the title, xlab is x-axis label
 (ylab also available)
# par() gives graphical options
# mfrow = "multifigure by row or column"
# 2 rows, 1 column
par(mfrow=c(2,1))
stripchart(eq, main="Earthequake",
 xlab="Number of Earthquakes")
stripchart(eq, method="stack", main="Earthequake,
method is stack",xlab="Number of Earthquakes")
```


Earthequake
10
15
20
25
Number of Earthquakes


Earthequake, method is stack
10
15
20
25
Number of Earthquakes

## Graphical summaries

## 6-2. Stem-and-Leaf Diagrams

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number $x_i$ consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

### Steps for Constructing a Stem-and-Leaf Diagram

(1) Divide each number $x_i$ into two parts: a **stem**, consisting of one or more of the leading digits and a **leaf**, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

```
> #### stem-and-leaf
> # stem-and-leaf plot
> stem(eq)

  The decimal point is 1 digit(s) to the right of the |

  0 | 6788
  1 | 01124
  1 | 555668
  2 | 0233
  2 | 5
```

```
> # scale=2 makes plot roughly twice as wide
> stem(eq, scale=2)

  The decimal point is at the |

   6 | 00
   8 | 00
  10 | 000
  12 | 0
  14 | 0000
  16 | 00
  18 | 0
  20 | 0
  22 | 000
  24 | 0
```

## 6-3 Frequency Distributions and Histograms

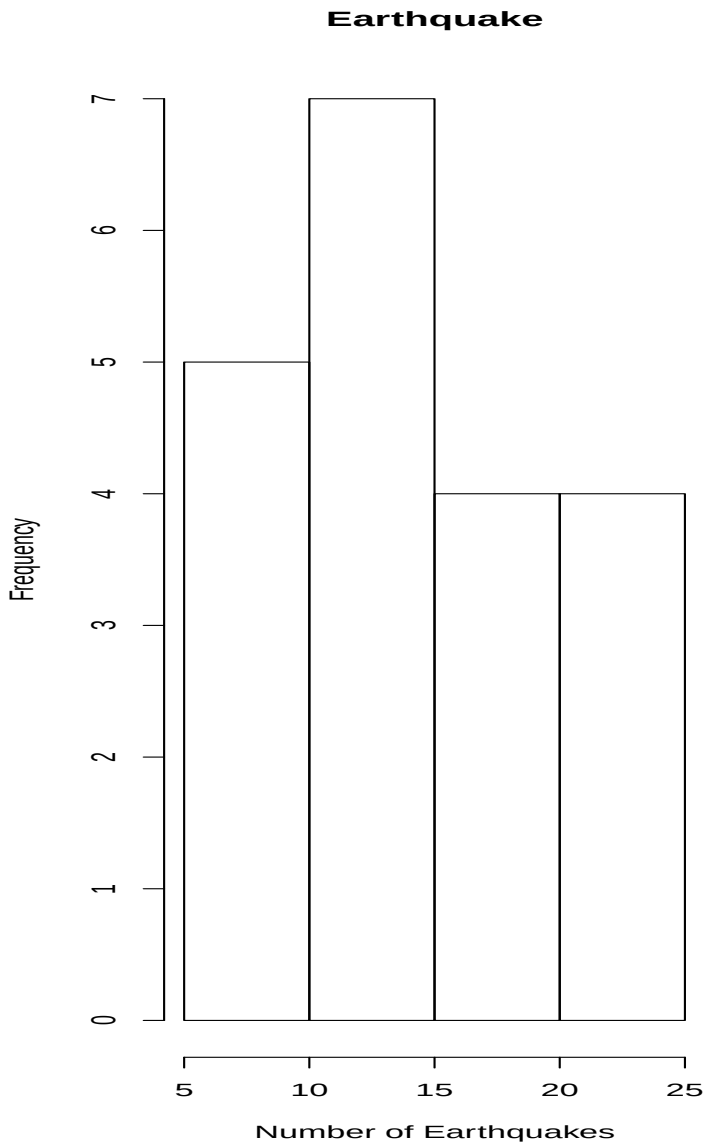
- A **frequency distribution** is a more compact summary of data than a stem-and-leaf diagram.

- To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**, or **bins**.

**Constructing a Histogram (Equal Bin Widths):**

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

```
#### hist
# histogram using R base graphics
par(mfrow=c(1,2))
hist(eq, main="Earthquake", xlab="Number of Earthquakes")
# breaks are how many bins-1 to use
# freq=FALSE changes the vertical axis to density,
# so the total area of the bars is now equal to 1
hist(eq, breaks = 10, freq = FALSE,
main="Histogram, density")
```


Earthquake
Frequency
0
1
2
3
4
5
6
7
5
10
15
20
25
Number of Earthquakes

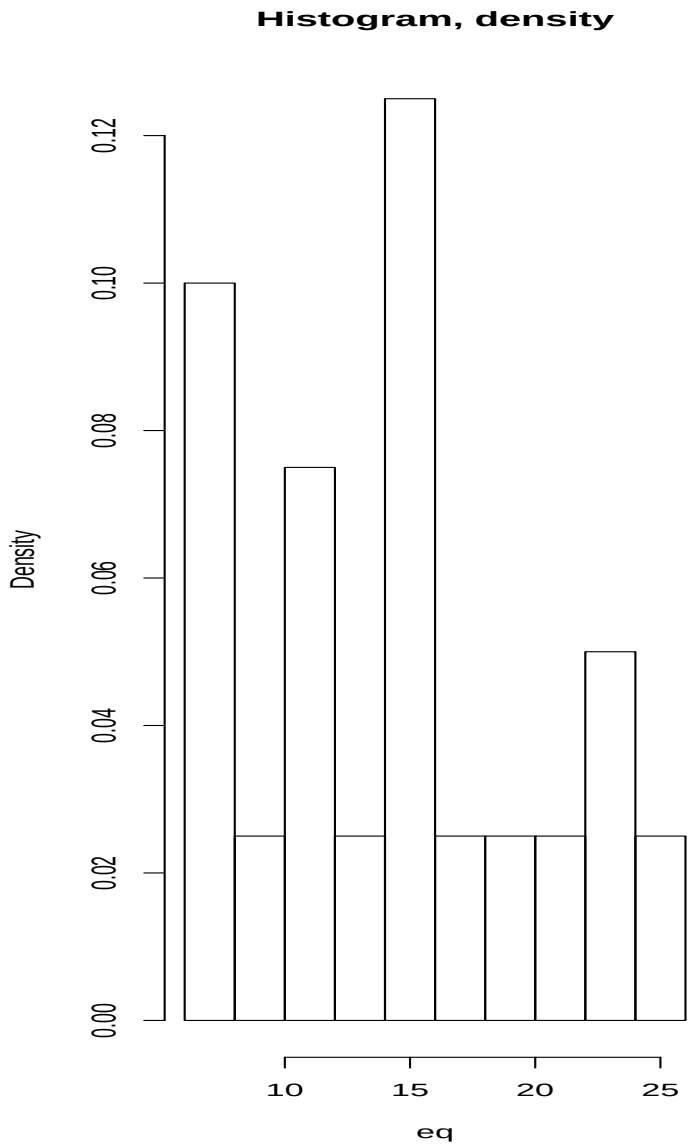
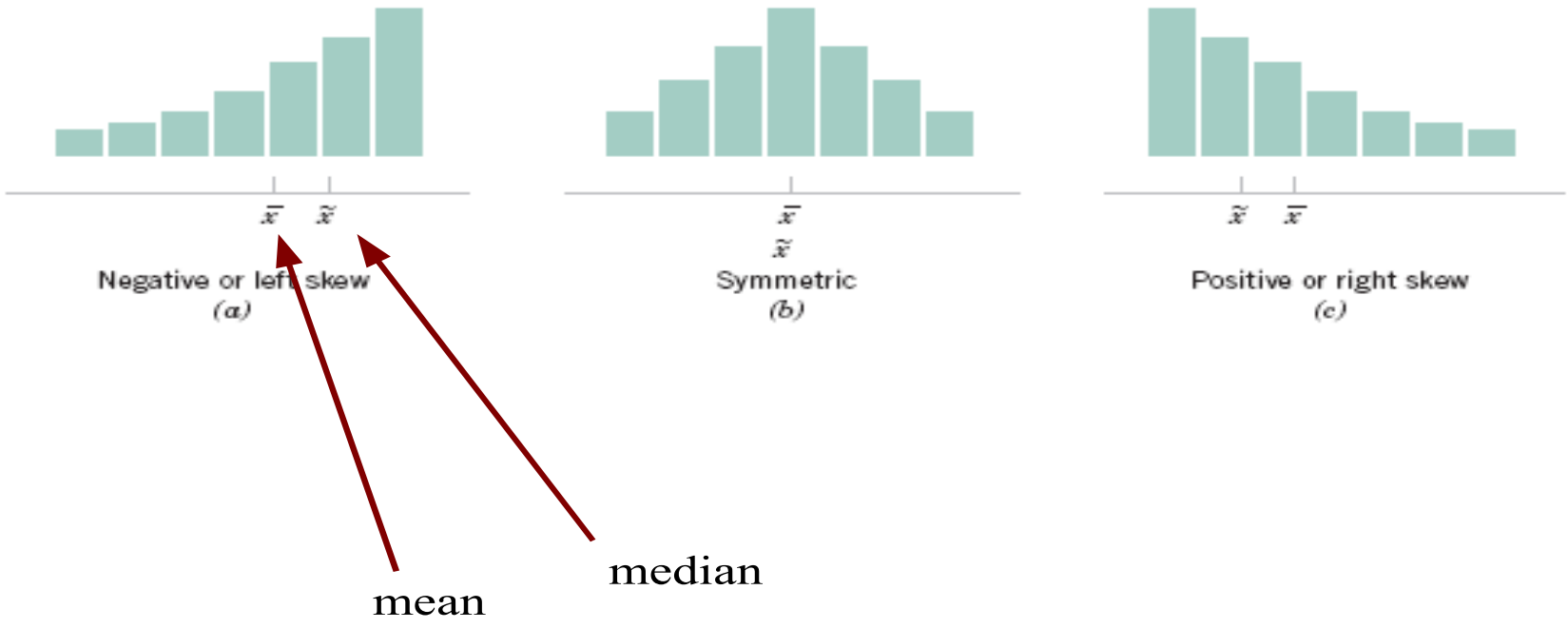

Histogram, density
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
10
15
20
25
eq

**Figure 6-11** Histograms for symmetric and skewed distributions.

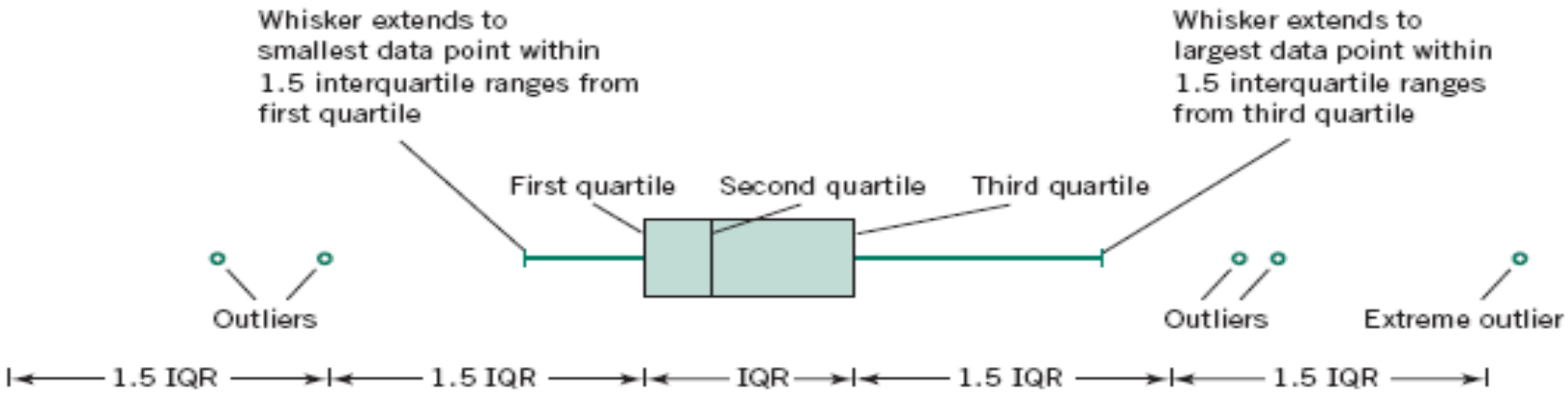


## 6-4 Box Plots

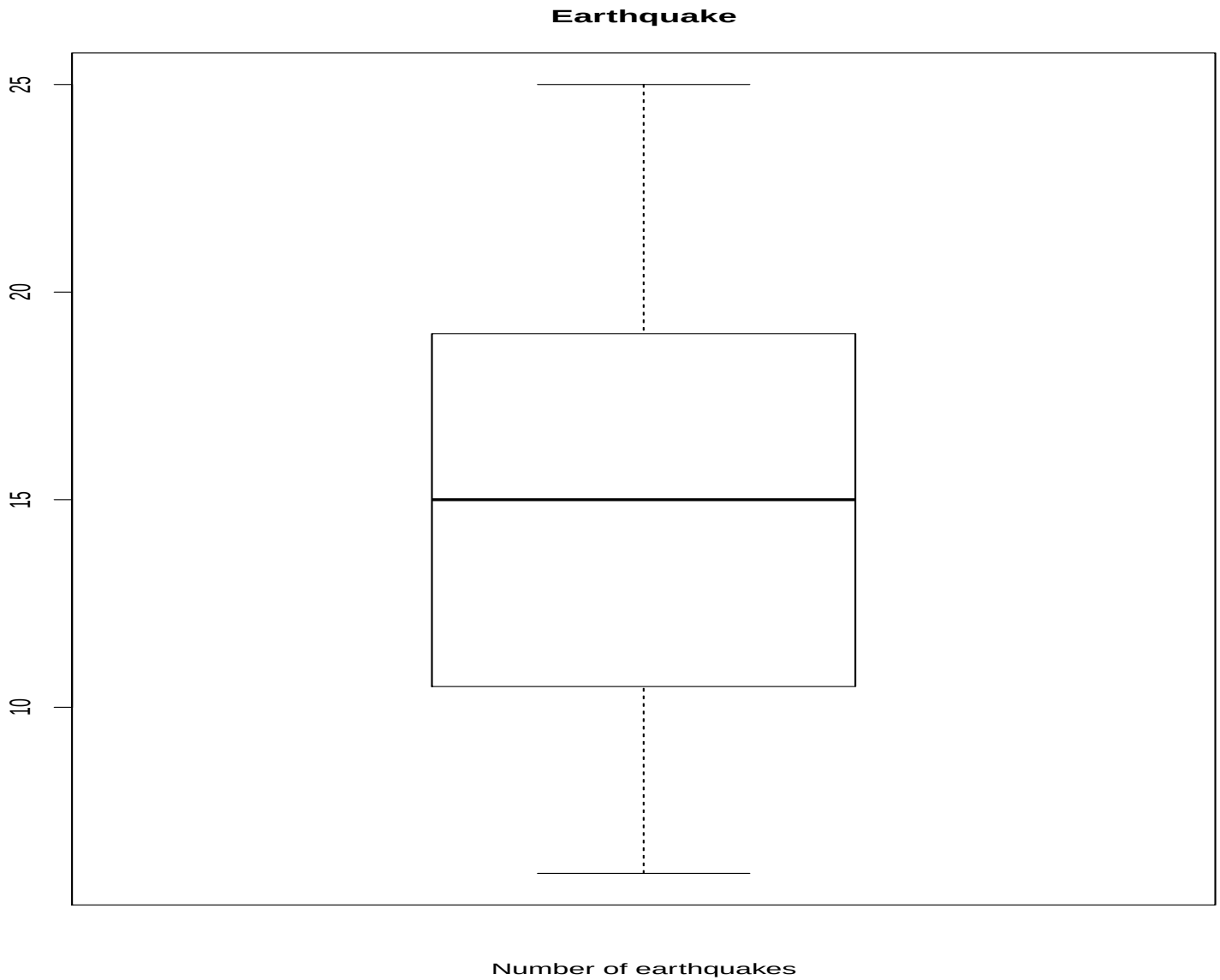
- The **box plot** is a graphical display that **simultaneously** describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data.

- **Whisker**
- **Outlier**
- **Extreme outlier**

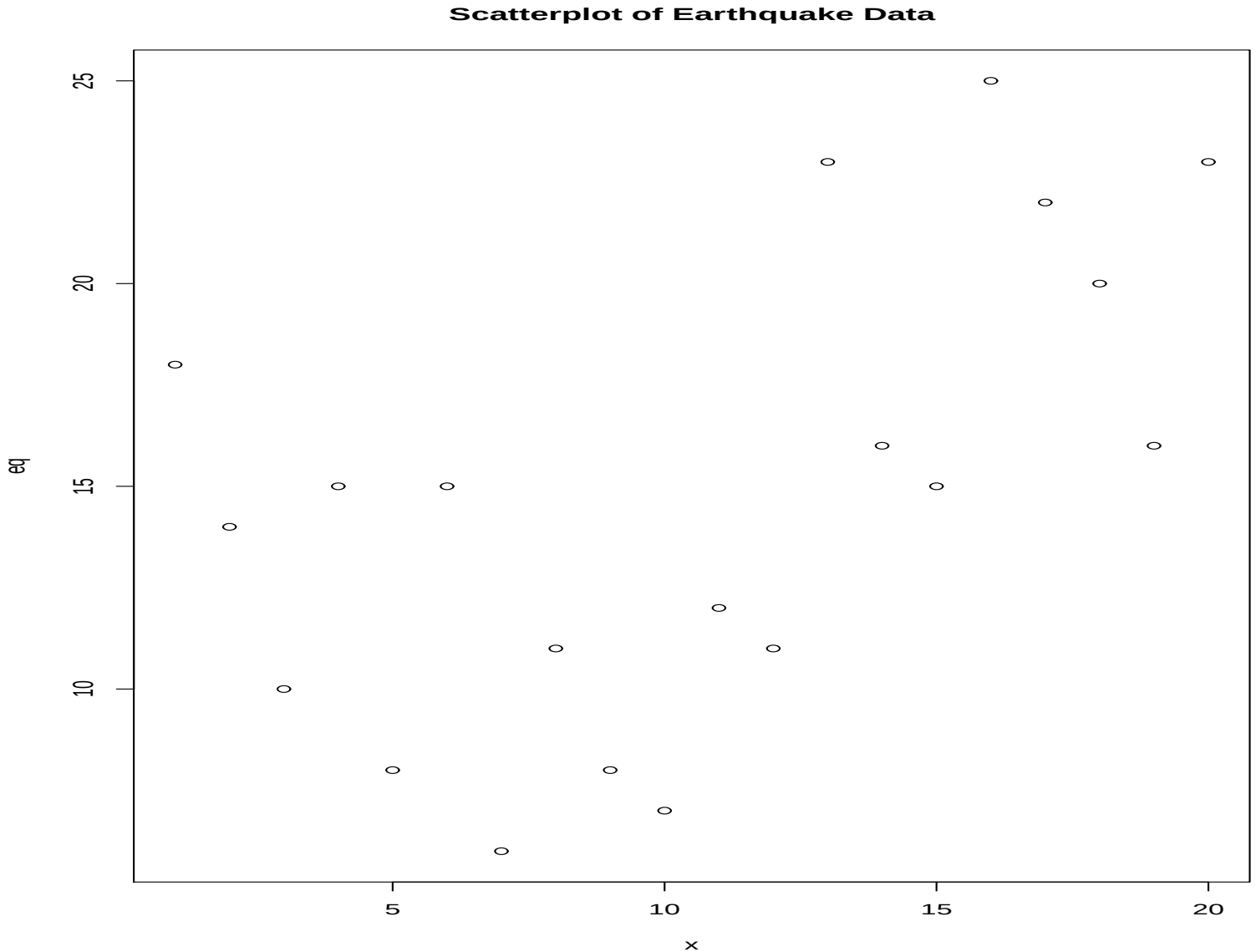
# Box Plots


Whisker extends to
smallest data point within
1.5 interquartile ranges from
first quartile
Whisker extends to
largest data point within
1.5 interquartile ranges
from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

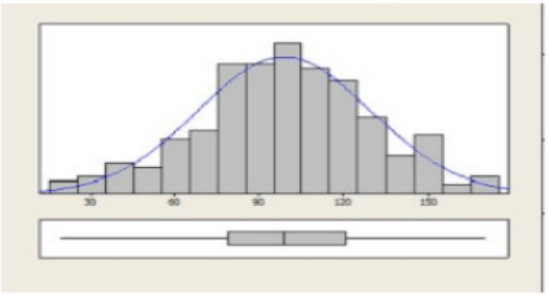
```
#### Box-plot
par(mfrow=c(1,1))
boxplot(eq, horizontal=FALSE, main="Earthquake",
xlab="Number of earthquakes")
```


Earthquake
25
20
15
10
Number of earthquakes

```
##plot data
x<-seq(1:20)
plot(x,eq,main="Scatterplot of Earthquake Data")
```

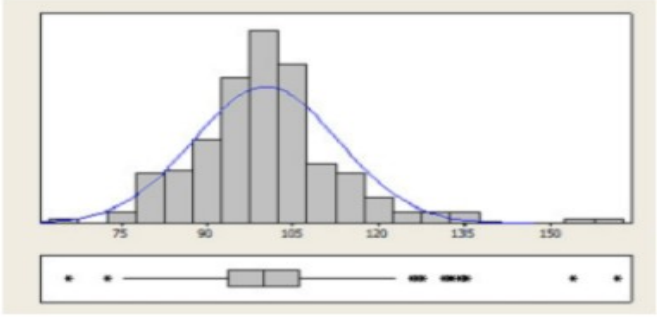

Scatterplot of Earthquake Data
eq
x
25
20
15
10
5
10
15
20

# Distributional shapes

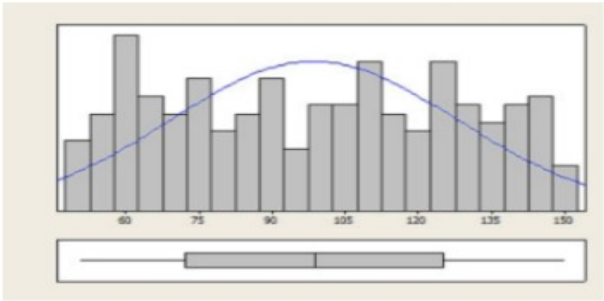



The distribution is **unimodal, symmetric and bell-shaped ("normal").**

The distribution is **unimodal, symmetric and heavy-tailed**

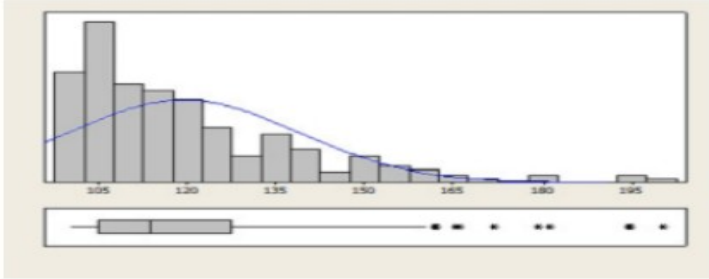


# Distributional shapes

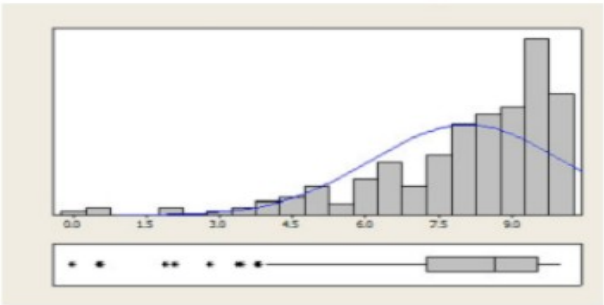



The distribution is **symmetric, but not bell-shaped.**
The boxplot shows symmetry, but the tails of the distribution are shorter(lighter) than in the normal distribution.

The distribution is **skewed to the right,** because the right tail is much longer than the left tail.



# Distributional shapes




The distribution is **skewed to the left,** because the left tail is much longer than the right tail.

The distribution is

**bimodal**.