MAT 2377
Probability and Statistics for Engineers

Chapter 4 (Sections 4.2,4.8)
Descriptive Statistics

Iraj Yadegari (uOttawa)

Fall 2021

University of Ottawa

Contents

Data Descriptions

In a sense, the underlying reason for statistical analysis is to reach an **understanding of the data**.

Studies and experiments give rise to **statistical units**.

These units are typically described with **variables** (and measurements).

Variables are either **qualitative** (categorical) or **quantitative** (numerical).

Categorical variables take values (**levels**) from a finite set of **categories** (or classes).

Numerical variables take values from a (potentially infinite) set of **quantities**.

Examples:

1. Age is a numerical variable, measured in years, although is is often reported to the nearest year integer, or in an age range of years, in which case it is an **ordinal** variable (mixture of qualitative or quantitative).

2. Typical numerical variables include distance in m, volume in cm^3, etc.

3. Disease diagnosis is a categorical variable with (at least) 2 categories (positive/negative).

4. Compliance with a standard is a categorical variable: there could be 2 levels (compliant/non-compliant) or more (compliance, minor non-compliance issues, major non-compliance issues).

5. Count variables are numerical variables.

Statistical Summaries

A variable can be described with two type of measures: **centrality**, **spread**.

- **Centrality** measures: **median**, **mean**, (mode, less frequent).
- **Spread** (variation or dispersion) measures: **variance, standard deviation** (sd), **inter-quartile range** (IQR), range (less frequent), (**skew** and **kurtosis** are also used sometimes).

The median, range and the quartiles are easily calculated from an **ordered** list of the data.

Centrality measures

Median

Mean

Quartiles

Outliers

(Sample) Median

The **median** med$(x_1, \dots, x_n)$ of a sample of size n is a numerical value which splits the ordered data into 2 equal subsets: half the observations are below the median, **and** half above it.

- If n is **odd**, then the **position** of the median is $(n+1)/2$, that is to say, the median observation is the $\frac{n+1}{2}^{\text{th}}$ ordered observation.

- If n is **even**, then the median is the average of the $\frac{n}{2}^{\text{th}}$ and the $(\frac{n}{2}+1)^{\text{th}}$ ordered observations.

The procedure is simple: Order the data, and follow the even/odd rules.

Examples:

1. med$(4, 6, 1, 3, 7)$ = med$(1, 3, 4, 6, 7) = x_{(5+1)/2} = x_3 = 4$. There are 2 observations below 4 $(1, 3)$, and 2 observations above 4 $(6, 7)$.

2. med$(1, 3, 4, 6, 7, 23) = \frac{x_{6/2} + x_{6/2+1}}{2} = \frac{x_3 + x_4}{2} = \frac{4+6}{2} = 5$. There are 3 observations below 5 $(1, 3, 4)$, and 3 observations above 4 $(6, 7, 23)$.

3. med$(1, 3, 3, 6, 7) = x_{(5+1)/2} = x_3 = 3$. There seems to be only 1 observation below 3 (1), but 2 observations above 3 $(6, 7)$.

   This is not quite the correct interpretation of the median: **above** and **below** in the definition should be interpreted as **after** and **before**, respectively. In this example, there are 2 observations $(x_1 = 1, x_2 = 3)$ before the median $(x_3 = 3)$, and 2 after $(x_4 = 6, x_5 = 7)$.

(Sample) Mean

The **mean** of a sample is simply the arithmetic average of its observations. For observations $x_1, x_2, \ldots, x_n$, the sample mean is

$$\text{AM}(x_1, \ldots, x_n) = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Other means exist, such as the **harmonic** mean and the **geometric** mean:

$$\text{HM}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \quad \text{and} \quad \text{GM}(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}.$$

Examples:

1. $\text{AM}(4,6,1,3,7) = \dfrac{4+6+1+3+7}{5} = \dfrac{21}{5} = 4.2 \approx 4 = \text{med}(4,6,1,3,7)$.

2. $\text{AM}(1,3,4,6,7,23) = \dfrac{1+3+4+6+7+23}{6} = \dfrac{44}{6} \approx 7.3$, which is not nearly as close to $\text{med}(1,3,4,6,7,23) = 5$.

3. $\text{HM}(4,6,1,3,7) = \dfrac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{1}+\frac{1}{3}+\frac{1}{7}} = \dfrac{5}{53/28} = \dfrac{140}{53} \approx 2.64$.

4. $\text{GM}(4,6,1,3,7) = \sqrt[5]{4\cdot 6\cdot 1\cdot 3\cdot 7} \approx \sqrt[5]{(504)} \approx 3.47$.

If $x = (x_1,\ldots,x_n)$ and $x_i > 0$ for all i,

$$\min(x) \leq \text{HM}(x) \leq \text{GM}(x) \leq \text{AM}(x) \leq \max(x).$$

Mean or Median?

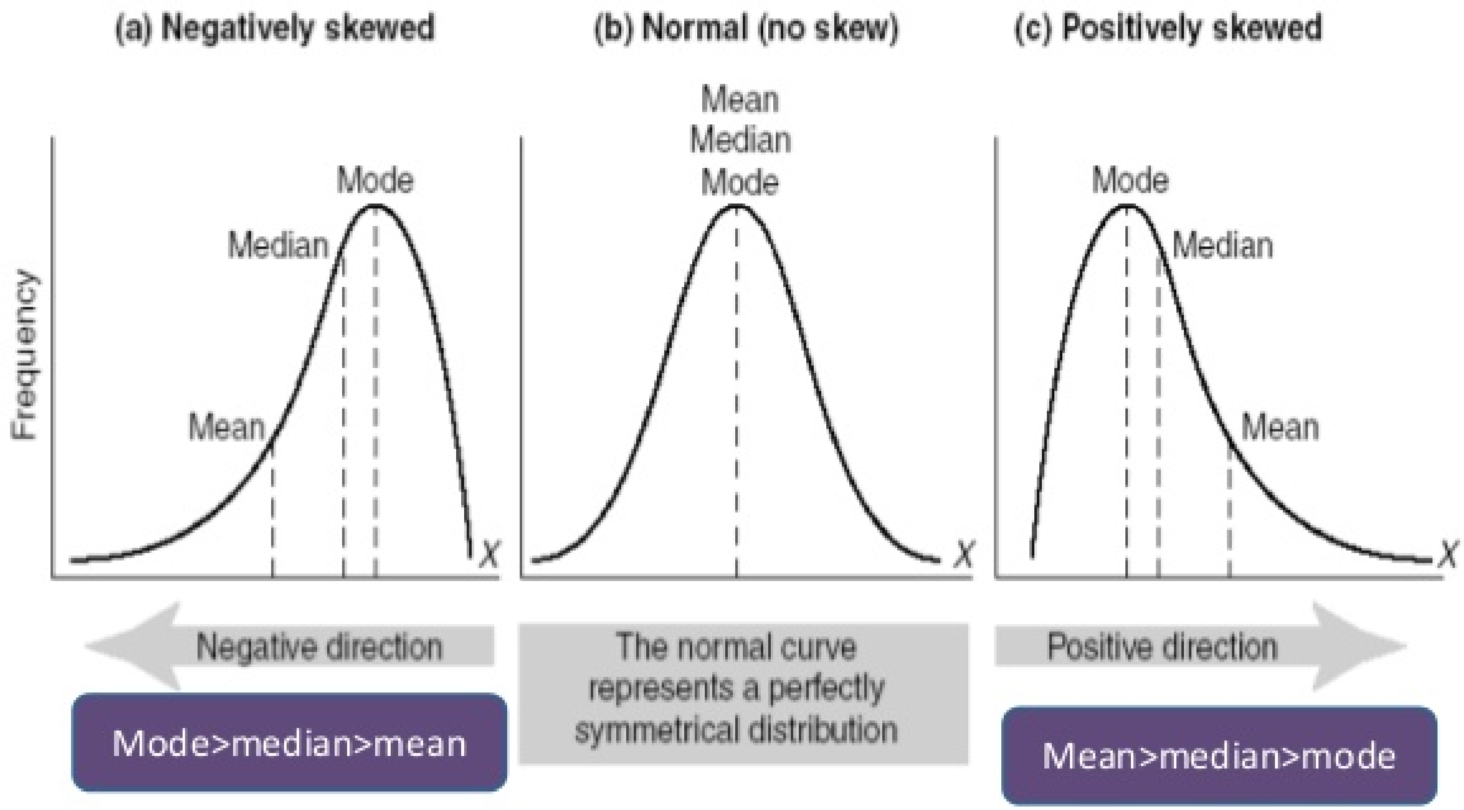
Which measure of centrality should be used to report on the data?

1. The mean is **theoretically supported** (see Central Limit Theorem).

2. If the data distribution is roughly symmetric then both values will be near one another.

3. If the data distribution is **skewed** then the mean is pulled toward the long tail and as a result gives a distorted view of the centre. Consequently, medians are generally used for house prices, incomes etc.

4. The median is **robust** against extreme values, but mean is affected by extremes.



Measures of Dispersion

A) The **sample standard deviation** s and **sample variance** s^2 are estimates of the underlying distribution's σ and σ^2.

For observations $x_1, x_2, \ldots, x_n$, we have

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right).$$

B) The **sample range** is

$$\text{range}(x_1, \ldots, x_n) = \max\{x_i\} - \min\{x_i\} = y_n - y_1,$$

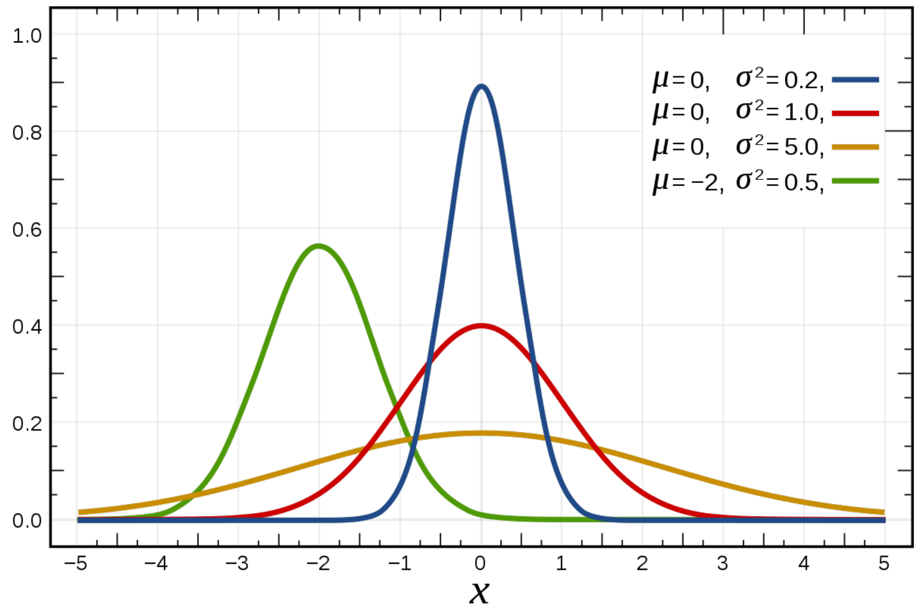
where $y_1 \leq \ldots \leq y_n$ is the ranked data.

C) The **inter-quartile range** is $\text{IQR} = Q_3 - Q_1$.

Standard Deviation

The mean, the median, and the mode provide an idea as to where some of the distribution's "mass" is located.

The standard deviation provides some notion of its spread.



Quartiles

Another way to provide information about the spread of the data is with the help of **quartiles**.

The **lower quartile** $Q_1(x_1, \ldots, x_n)$ of a sample of size n, or Q_1, is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations are below Q_1, **and** 75% of the observations are above Q_1.

Similarly, the **upper quartile** Q_3 splits the ordered data into 75% of the observations below Q_3, **and** 25% of the observations above Q_3.

The median can be interpreted as the **middle quartile**, Q_2: 50% of the observations are below Q_2, **and** 50% of the observations are above Q_2.

How to calculate?

Sort the sample observations $\{x_1, x_2, \dots, x_n\}$ in an **increasing order** as

$$y_1 \leq y_2 \leq \dots \leq y_n.$$

The smallest y_1 has **rank** 1 and the largest y_n has **rank** n.

- The lower quartile Q_1 is computed as the average of ordered observations with ranks $\lfloor \frac{n}{4} \rfloor$ and $\lfloor \frac{n}{4} \rfloor + 1$.
- Similarly, Q_3 is computed as the average of ordered observations with ranks $\lceil \frac{3n}{4} \rceil$ and $\lceil \frac{3n}{4} \rceil + 1$.
- The median can be interpreted as the **middle quartile**, Q_2.

Example:

$$Q_1(1,3,4,6,7,10,12,23) = 3.5, \quad Q_3(1,3,4,6,7,10,12,23) = 11.$$

Example: a dataset describes the daily number of accidents in Sydney:

```
> accident
 6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2,
 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7,
 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25
> sort(accident)
 1  2  2  2  2  3  3  3  3  4  6  6  7  7  7  7  7  7  8  9
 9  9  9 10 11 12 13 14 14 15 17 17 18 21 21 22 23 24 25 31
> summary(accident)
  Min.  1st quartile  Median  Mean  3rd quartile   Max.
  1.00     5.50          9.00  10.78    15.50        31.00
> var(accident) 58.7
```

Now, replace the 31 with 130. The new mean is 13.28 and the new variance is 412.4, but the median is the same.

Outliers

An outlier is an observation that lies outside the overall pattern in a distribution.

Let x be an observation in the sample. It is a **suspected outlier** if

$$x < Q_1 - 1.5\,\text{IQR} \quad \text{or} \quad x > Q_3 + 1.5\,\text{IQR},$$

where $\text{IQR} = Q_3 - Q_1$ it the **inter-quartile range** $Q_3 - Q_1$.

This definition only applies with certainty to **normally distributed** data, although it is often used as a first outlier analysis method.

Exercise: Consider a sample of $n = 10$ observations displayed in ascending order.

$$15, 16, 18, 18, 20, 20, 21, 22, 23, 75.$$

1. Verify that the standard deviation for this sample is $s = 17.81884$.

2. Verify that $Q_1 = 17.5$ and $Q_3 = 22.25$.

3. Are there any likely outliers in the sample? If so, indicate their values.

Visual Summaries

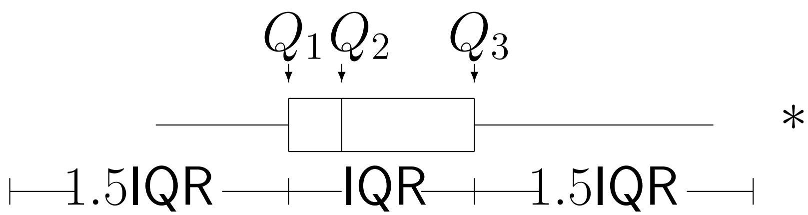
Box plot

Histogram

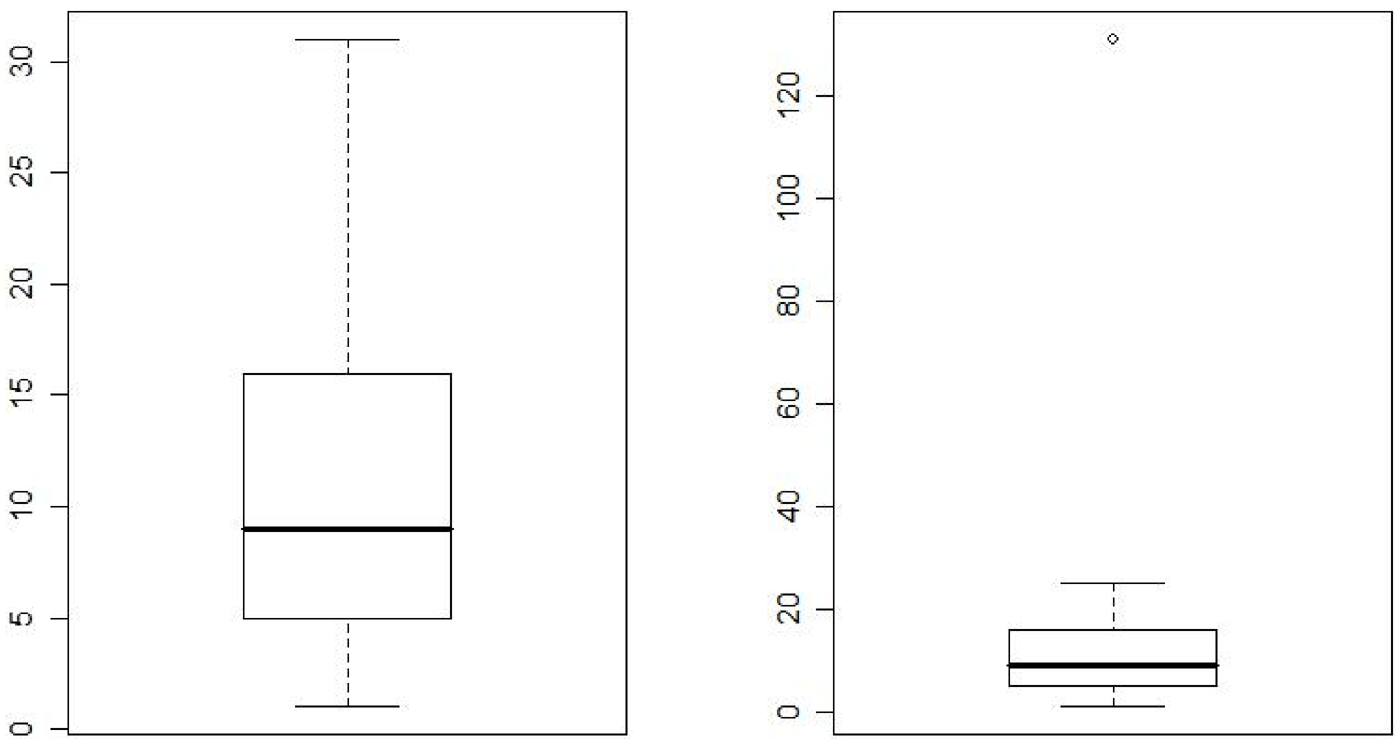
Skewness

Box plot

The **boxplot** is a quick and easy way to present a graphical summary of a univariate distribution.



- The main part is a box, with endpoints at the lower and upper quartiles, and with a "belt" at the median.
- A line is extending from Q_1 to the smallest value less than 1.5IQR to the left of Q_1.
- A line is extending from Q_3 to the largest value less than 1.5IQR to the right of Q_3.
- Suspected outliers are represented by *.

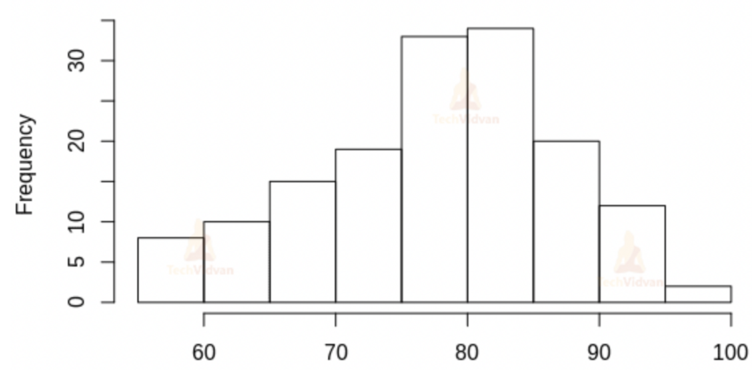

0
5
10
15
20
25
30
0
20
40
60
80
100
120

Histogram

Histograms also provide an indication of the distribution of the sample.

Histograms should contain the following information:

- the range of the histogram is $r = \max\{x_i\} - \min\{x_i\}$;
- the number of bins should approach $k = \sqrt{n}$, where n is the sample size;
- the bin width should approach r/k,
- and the frequency of observations in each bin should be added to the chart.

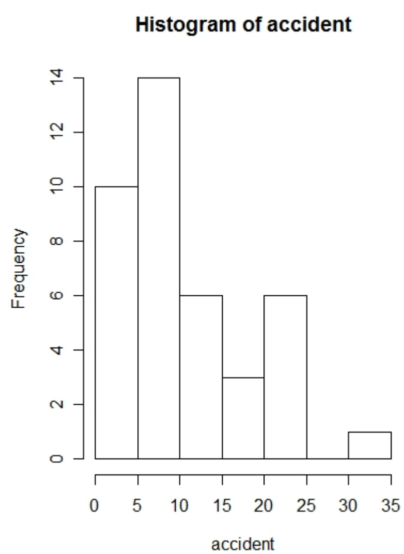


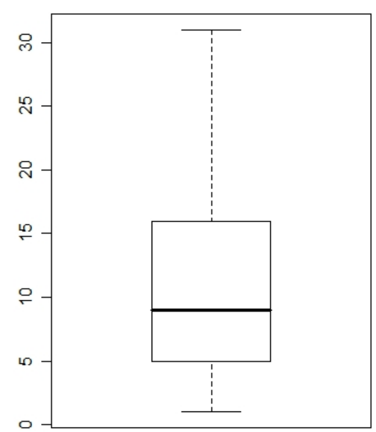
Skewness

Boxplots give an easy graphical means of getting an impression of the shape of the data set. The shape is used to suggest a mathematical model for the situation of interest.

The data set is **right skewed** if the boxplot is stretched to the right.

Similar observations can be inferred from the histogram.





If the data distribution is symmetric then the (population) median and mean are equal and the first and third (population) quartiles are equidistant from the median.

If data is stretched to the right or left, then distribution of data is Asymmetric (skewed).

If $Q_3 - Q_2 > Q_2 - Q_1$ then the data distribution is **skewed to the right**.

If $Q_3 - Q_2 < Q_2 - Q_1$ then the data distribution is **skewed to left**.

Example: the grades for the midterm exam of a course are shown below. Discuss the results.

```
> grades<-c(80,73,83,60,49,96,87,87,60,53,66,83,32,80,66,90,72,55,76,46,48,69,45,48,77,
52,59,97,76,89,73,73,48,59,55,76,87,55,80,90,83,66,80,97,80,55,94,73,49,32,76,57,42,94,
80,90,90,62,85,87,97,50,73,77,66,35,66,76,90,73,80,70,73,94,59,52,81,90,55,73,76,90,46,
66,76,69,76,80,42,66,83,80,46,55,80,76,94,69,57,55,66,46,87,83,49,82,93,47,59,68,65,66,
69,76,38,99,61,46,73,90,66,100,83,48,97,69,62,80,66,55,28,83,59,48,61,87,72,46,94,48,59,
69,97,83,80,66,76,25,55,69,76,38,21,87,52,90,62,73,73,89,25,94,27,66,66,76,90,83,52,52,
83,66,48,62,80,35,59,72,97,69,62,90,48,83,55,58,66,100,82,78,62,73,55,84,83,66,49,76,73,
54,55,87,50,73,54,52,62,36,87,80,80)

> hist(grades)

> # function to calculate mode
> fun.mode<-function(x){as.numeric(names(sort(-table(x)))[1])}
```

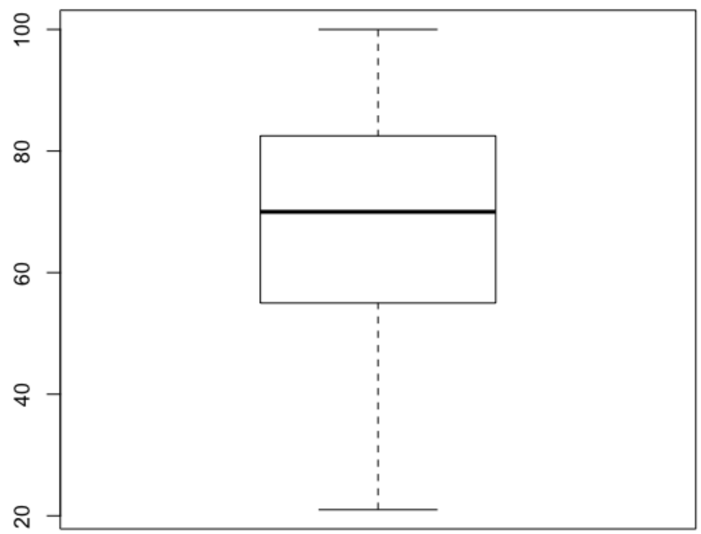
```
> library(ggplot2)
> ggplot(data=data.frame(grades), aes(grades)) + geom_histogram(aes(y =..density..),
                breaks=seq(20, 100, by = 10),
                col="black",
```

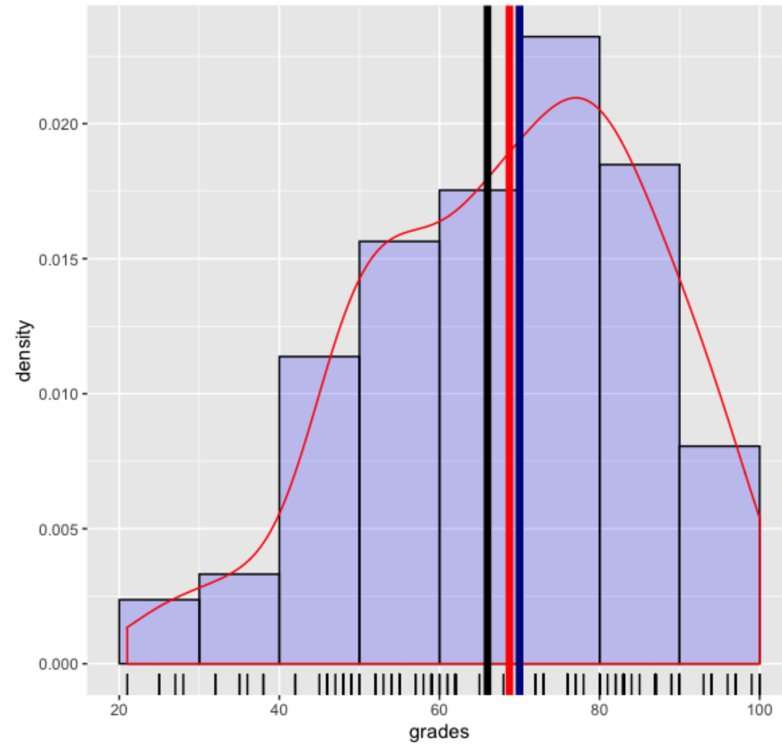
```
                 fill="blue",
                 alpha=.2) +
    geom_density(col=2) + geom_rug(aes(grades)) +
    geom_vline(aes(xintercept = mean(grades)),col='red',size=2) +
    geom_vline(aes(xintercept = median(grades)),col='darkblue',size=2) +
    geom_vline(aes(xintercept = fun.mode(grades)),col='black',size=2)

> boxplot(grades)

> summary(grades)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.00   55.00   70.00   68.74   82.50  100.00

> library(psych)
> describe(grades)
n   mean    sd median trimmed   mad min max range  skew kurtosis  se
211 68.74 17.37     70   69.43 19.27  21 100    79 -0.37    -0.46 1.2
```


100
80
60
40
20


density
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
grades