Contents

1 Basic Concepts

1.1 Variables and Data

Definitions and introduction to terminology used in statistics.

Definition: Data are numbers with a context.
Any set of data contains information on individuals.

Individuals(experimental units) in a particular population usually posses many different characteristics:

Example:
Student is individual Characteristics: age, height, weight, sex, hair color, grade in Stat151.

Definition:

- *Individuals* (experimental units) are the objects about which information is desired. Individuals may be people, animals, or things.
- A *variable* is any characteristic of an individual. A variable can take different values for different individuals.
- The entire collection of individuals is called the population of interest.
- A sample is a subset of the population, selected for study in some prescribed manner.

1.2 Two types of Statistics

1. In *descriptive statistics* the purpose is to describe a data set with the help of numerical and graphical methods. Usually the interest is just in the sample presented
2. In *inferential statistics* the goal is to draw conclusions about an entire population based on the findings in a sample. We use the sample to estimate, predict and make decisions about the entire population

Example 1

1. The proportion of students at MacEwan who have a U-pass is of interest. The sales numbers are obtained from all participating outlets selling the passes, in addition the registrars office provides the number of students enrolled at MacEwan and the proportion can be found. This describes a purely descriptive statistical application.
2. Does attendance matter in student success?

   To answer this question for a "random sample" of students enrolled in first-year statistics courses the attendance, midterm, and final exam results are recorded.

   Then this data is used to "test" if the data provide sufficient evidence for the statement "attendance impacts the marks obtained in an exam for students enrolled in first-year courses."

   This describes the setup of a study employing inferential statistics. Data from a sample are used to draw conclusions about all students enrolled in first-year courses (population).

1.3 Observational Study versus Designed Experiment

1. In *observational studies,* researchers simply observe and measure and do not intervene in what they observe and measure.

2. In *designed experiments,* researchers impose treatments or controls and then observe and measure the consequences.

Observational studies only allow to identify associations, whereas designed experiments allow to derive cause and effect relationships.

Example 2

1. Is treatment A better than treatment B for treating kidney stones?

   Patients are randomly assigned to either treatment A or B, then the outcome of the treatment is observed and either classified as success, partial success, or failure. This is a designed experiment since the treatment is imposed on the patients by random assignment.

   If patients after treatment A are found to fare *significantly* better than patients after treatment B it is safe to conclude that the treatment A caused patients to do better.

2. Does living close to a nuclear plant increase your risk of a certain type of cancer?

   Health records of cancer patients and healthy matched individuals (matched by age, gender) are analyzed. Since the distance to the nuclear plant is not assigned but only observed this is an observational study.

   If this sample wold reveal that the distance of homes to nuclear plant of cancer patients is *significantly* lower than for healthy patients, this would only reveal an association but not that the distance is causing cancer.

1.4 Samples

When making decisions based on data it is crucial how the data is selected for the decision to be well-informed and relevant.
We wish for the sample to be *representative of the population*, which means that it reflects the relevant characteristics of the population as closely as possible.

One approach is **Simple Random Sampling**.
This approach avoids bias in the sample but is sometimes inefficient.

Definition:
We speak of simple random sampling if every sample of the population has the same chance of being selected.

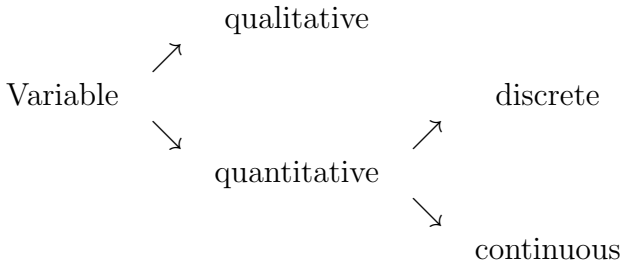
Example:
Four students, Adi(A), Berta (B), Carly (C), and Dominique (D) are in a student union committee, two are invited to attend a president's function. They decide to choose a simple random sample of who might attend the function.
the population: A, B, C, D
possible samples: AB, AC, AD, BC, BD, CD

roll a fair die to choose which of the six samples is going to attend the function.

Example: If the population is larger (all students currently enrolled at MacEwan), one way of choosing a simple random sample uses Random Number Tables, like Table I in the text book. Use Student IDs to identify students (assume they only have 3 digits), choose 4 students starting in column 10 line 03 (discard doubles, and skip students who are not enrolled anymore) moving downward.
We choose: 862, 807, 948, 408 the registrar tells us that student 807 has already graduated, so we choose one more from the list 577.
The sample is 862, 948, 408, 577

How else could we choose a simple random sample?

1.5 Types of Variables



Example:

1. qualitative: sex, hair color, species, scales (very good ..., very bad)
2. quantitative discrete: number of siblings (often result of counting)
3. quantitative continuous: height weight (often result of measurement)

Definition:

1. A *qualitative* variable places an individual into one of several groups or categories.
2. A *quantitative* variable measures a numerical quantity or amount in each individual.
3. The *distribution* of a variable tells us what values it takes and how often it takes these values.

2 Descriptive Statistics

(Exploratory data analysis)

1. Examine each variable by itself
2. Study the relationships among the variables.

Example 3

1. number of apartments in a building

2. average square meters per apartment
3. gym (yes/no)
4. laundry (yes/no)
5. how many one/two bedroom apartments
6. most widely used flooring products

2.1 Displaying Data with Graphs

Different type of graphs for different type of data

2.1.1 Qualitative data

After qualitative data has been sampled it should be summarized to provide the following information:

1. What values have been observed? (red, green, blue, brown, orange, yellow)
2. How often did every value occur?

The distribution of a qualitative variable is given in form of a table giving the following information:

- each possible category
- frequency, or number of individuals who fall into each category
- relative frequency(proportion) or percent of individuals who fall into each category
- percentage of measurements in each category

Definition: The *relative frequency* for a particular category is the fraction or proportion of the frequency that the category appears in the the data set. It is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

$$\text{percent} = 100 \times \text{relative frequency}$$

Example:
sample size=number of observations=n=200

category	frequency	relative frequency	percentage
wood	50	0.25	25%
tiles	10	0.05	5%
linoleum	20	0.1	10%
carpet	120	0.6	60%
total	21	1	100%

Such a table is called the frequency distribution for qualitative data or statistical table.
Once the data is summarized in a frequency distribution table, the data can be displayed in a bar chart or pie chart. The bar chart will effectively show the frequencies or percent in the different categories whereas the pie chart will show the relationship between the parts and the whole.

Bar chart

A bar chart is a graph of the frequency distribution of a qualitative data variable

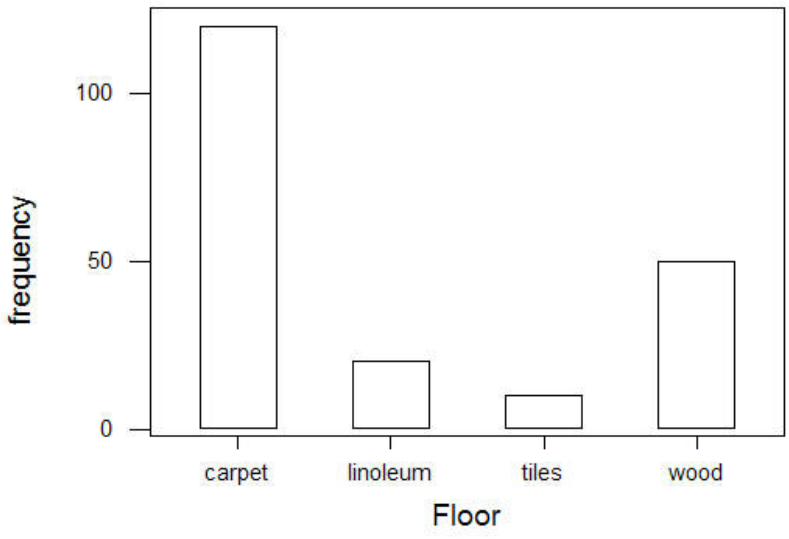
- For every category the x-axis is marked with a tick.
- Each category is represented by a bar, which AREA is proportional to the corresponding frequency (relative frequency).
- label the y-axis.

Remark: The width of each bar should be the same, so the height is proportional to the corresponding frequency.

Example 4

Suppose the frequency distribution of the mainly used flooring products is:

	frequency	relative freq
wood	50	0.25
tiles	10	0.05
linoleum	20	0.1
carpet	120	0.6



Pie charts

Pie charts provide an alternative kind of graph for qualitative data:

In a pie chart a circle is used to represent the sample. The size of the slice representing a particular category is proportional to the corresponding frequency (relative frequency).

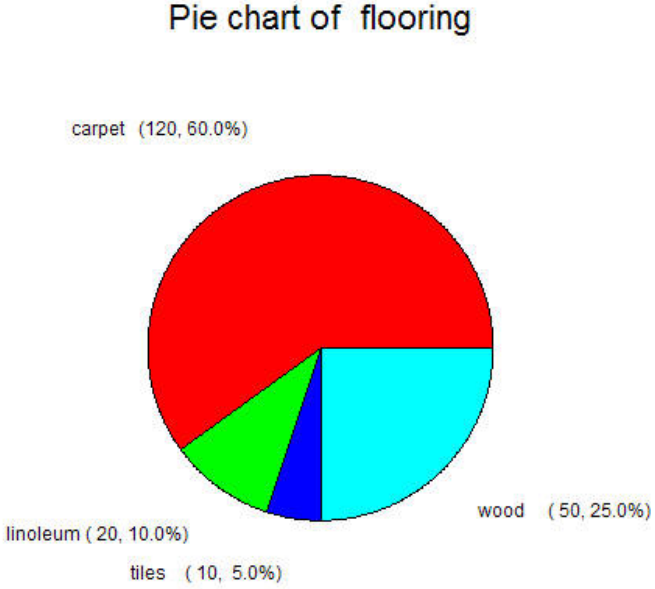
How to create a pie chart:

- Draw a circle
- Calculate the slice size (angle)

  $$\text{slice size}=\text{category relative frequency} \cdot 360$$

  (fraction of the circle for the category)
- use protractor to mark the angles

	frequency	relative freq	angle
wood	50	0.25	90
tiles	10	0.05	18
linoleum	20	0.1	36
carpet	120	0.6	216



M&M's example:
On the M&M's webpage the following information on the distribution of colors in peanut M&M's is provided

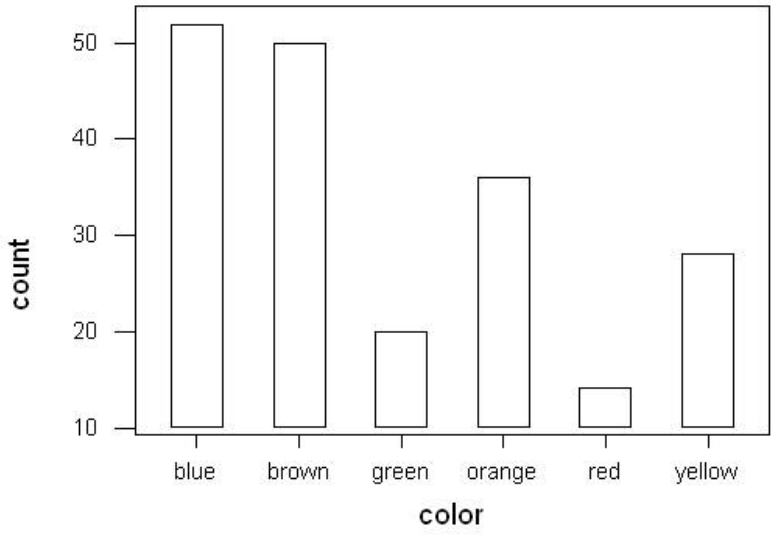
color	brown	yellow	red	blue	orange	green
percent	12%	15%	12%	23%	23%	15

In order if this distribution is a "true" description of what is in a bag, someone bought a bag with 200 peanut M&M's and wants to describe the colors of the contents.

Color is a qualitative variable, so a relative frequency table shall be obtained.

color	count	rel. freq.	percentage
brown	50	0.25	25%
yellow	28	0.14	14%
red	14	0.07	7 %
blue	52	0.26	26%
orange	36	0.18	18%
green	20	0.1	10%
Total	200	1	100%

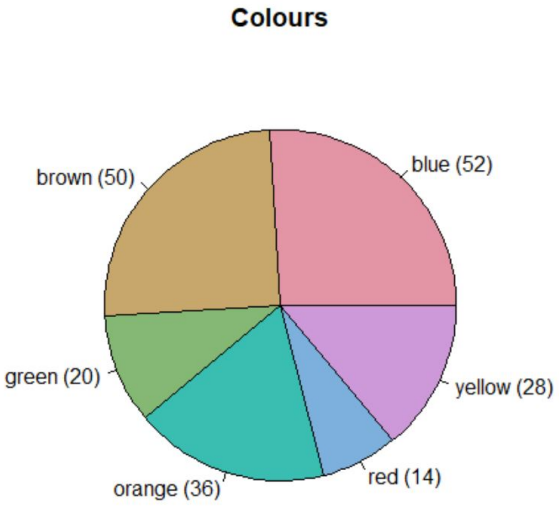
And a bar chart would look like this:



For the pie chart the angles of the slices have to be determined

color	count	rel. freq.	angle
brown	50	0.25	90^o
yellow	28	0.14	50.4^o
red	14	0.07	25.2^o
blue	52	0.26	93.6^o
orange	36	0.18	64.8^o
green	20	0.1	36^o
Total	200	1	360^o

resulting in the pie chart on the right



2.1.2 Quantitative Data

Graphs from this section display the data for a quantitative variable in a fashion so that the distribution of the data becomes apparent.

Stemplots

One way of displaying quantitative data is a stem and leaf plot.
Each observed number is broken into two pieces called the *stem* and the *leaf*.

How to do a stemplot:

1. Divide each measurement into two parts:
   - The first digit(s) of the number are the stems.
   - The last digit(s) of the number are the leaves.
2. List the stems in a column, with a vertical line to their right.
3. For each measurement, record the leaf portion in the same row as its corresponding stem.
4. Order the leaves from lowest to highest in each stem.
5. Provide a key to your stem and leaf coding so that the reader can recreate the actual measurements.

Example 5

Acceptance rates at some business schools:

16.3, 12.0, 25.1, 20.3, 31.9, 20.7, 30.1, 19.5, 36.2, 46.9, 25.8, 36.7, 33.8, 24.2, 21.5, 35.1, 37.6, 23.9, 17.0, 38.4, 31.2, 43.8, 28.9, 31.4, 48.9

Stemplot:

1	20 63 70 95
2	03 07 15 39 42 51 58 89
3	01 14 19 38 51 62 67 76 84
4	38 69 89

It shows: center, range, concentration, nature of distribution (unimodal, bimodal, multimodal), unusual values, skewed to the right/left.
Sometimes the available stem choices result in a plot that contains too few stems and a large number of leaves within each stem. In this situation you can *stretch* the stems by dividing each into several lines.

The two common choices for dividing stems are:

- Into two lines, with leaves 0 to 4 and 5 to 9
- into 5 lines, with leaves 0-1, 2-3, 4-5, 6-7, 8-9

Example:(acceptance rates)

1	20
1	63 70 95
2	03 07 15 39 42
2	51 58
3	01 14 19 38
3	51 62 67 76 84
4	38
4	69 89

You also can use stem and leaf plots for the comparison of the distribution of two groups:

Example 6

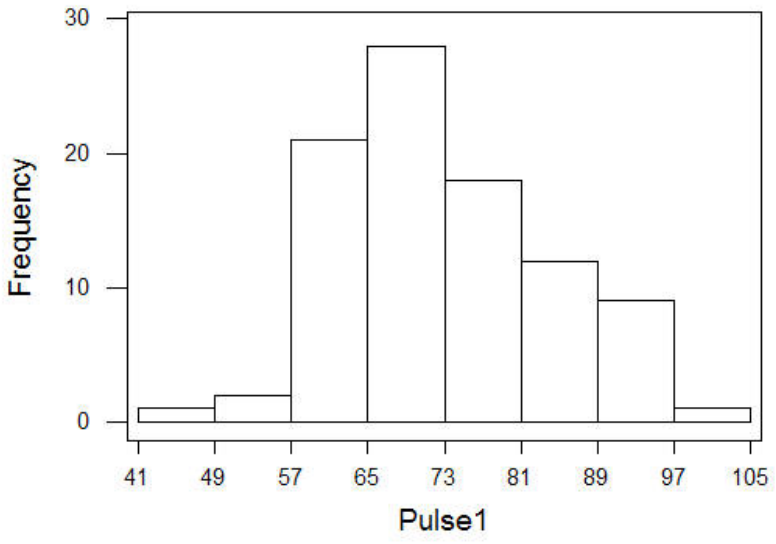
Prices of walking shoes in $:
90 70 70 70 75 70 65 68 60 74 70 95 75 68 85 40 65

Prices of running shoes in $:
49 59 65 79 80 95 105 69 69 60 77 95 78 85 89

Comparative stem and leaf plot:

9	4	0
9	5	
9950	6	05588
987	7	00000455
950	8	5
55	9	05
5	10	

Histograms
The most common graph for describing numerical continuous data is the histogram.
It visualizes the distribution of the underlying variable very well.
How a histogram looks like:



Definition:
A relative frequency histogram for a quantitative variable is a bar graph in which the hight of the bar shows "how often" (measured as a relative frequency) measurements fall in a particular interval. The classes or intervals are plotted along the horizontal axis.

How to do a histogram:

1. Decide which *class intervals* of equal length to use for the histogram.

   The number of *class intervals* used should be approximately the square root of the sample size, but not larger than 12.

   Use sensible interval boundaries: The intervals should have the same width and the boundaries are if possible whole numbers or tenth. The resulting intervals are called *class intervals*.

2. Create a frequency table for the class intervals using the method of **left inclusion**.

3. Mark the boundaries of the class intervals on a horizontal axis.

4. Use the relative frequency on the vertical axis.

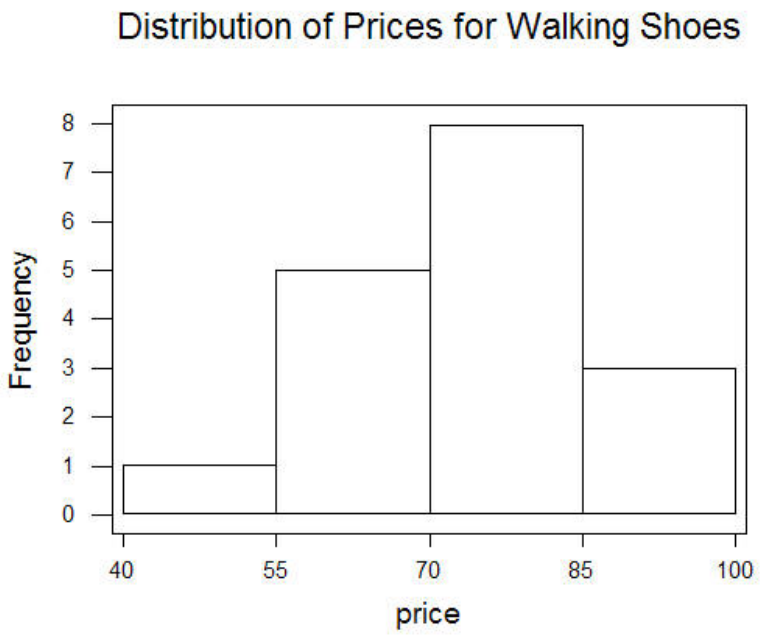
5. Draw a bar for each class interval, with heights according to the relative frequency of the corresponding interval.

Histogram for walking shoes:

1. The sample size is 15, the square root is a little smaller than 4, so use 4 class intervals. the range is 40 - 95, about 60. 60/4=15 $[40, 55), [55, 70), [70, 85), [85, 100)$

2.

class intervals	frequency	relative frequency
$[40, 55)$	1	0.066
$[55, 70)$	5	0.333
$[70, 85)$	6	0.4
$[85, 100)$	3	0.2



3.

It shows: center, range, concentration, nature of distribution (unimodal, bimodal, multimodal), unusual values, skewed to the right/left.

Histogram shapes

1. number of peaks: unimodal, bimodal (often occurs if you have observation from two groups (men, women), multimodal
2. symmetry: if you can draw a vertical line so that the part to the left is a mirror image of the part to the right, then it is symmetric.
3. nonsymmetric graphs are skewed. If the upper tail of the histogram stretches out farther than the lower tail, then is the histogram positively skewed, or skewed to the right.

   Is the lower tail longer than the upper tail the histogram is negatively skewed.

Check a distribution for shape, center, and spread and look for deviations from the overall pattern.

2.2 Describing Data with numbers

Only for quantitative variables!!

2.2.1 Describing the center of a data set

The mean of a set of numerical observation is the familiar arithmetic average. To write the formula for the mean in a mathematical fashion we have to introduce some notation.

Introduction of notation:
x= the variable for which we have sample data
n= sample size = number of observations

x_1=the first sample observation
x_2=the second sample observation
$\vdots$
x_n= the nth sample observation

For example, we might have a sample of n=4 observations on x=battery lifetime(hr):

x_1=5.9, x_2=7.3, x_3=6.6, x_4=5.7,

The sum of x_1, x_2, ..., x_n can be denoted by

$$x_1 + x_2 + \ldots + x_n$$

but this is cumbersome.

The Greek letter Σ is traditionally used in mathematics to denote summation. In particular $\Sigma_{i=1}^{n} x_i$ will denote the sum of $x_1, \cdots, x_n$. Abbreviation Σx is used in the book.

For the example above $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 5.9 + 7.3 + 6.6 + 5.7 = 25.5$

Definition
The *sample mean* of a numerical sample x_1, x_2, ..., x_n denoted by $\bar{x}$ is

$$\bar{x} = \frac{\text{sum of all observations}}{\text{number of observations}} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\Sigma_{i=1}^{n} x_i}{n}$$

The mean battery life is

$$\bar{x} = \frac{5.9 + 7.3 + 6.6 + 5.7}{4} = \frac{25.5}{4} = 6.375$$

If a set of observations include an outlier (an unusual value) the sample mean will be drawn into the direction of this outlier an in result will not describe the true center of the distribution. The mean is not resistant to outliers.
For this reason we are looking for a different value to measure the center of a distribution.

Another number to describe the center of a sample is the median. The median is the value that divides the *ordered* sample in two sets of the same size.

Definition

The sample median is determined by first ordering the n observation from smallest to largest. Then

$$M = \text{sample median} = \begin{cases} \text{the single middle value} & \text{if } n \text{ is odd} \\ \text{the average of the middle two values} & \text{if } n \text{ is even} \end{cases}$$

Example
Suppose you have the following ordered sample of size 6:
2 5 6 7 9 11
The median would be in this case the mean of the fifth and sixth observation $M = (6+7)/2 = 6.5$ and the sample mean $\bar{x} = 40/6 = 6.666$.

The median of the sample 6 7 8 8 8 is the third observation which is 8 and the mean is $\bar{x} = 7.4$.

Comparing mean and median.

The mean is the balance point of the distribution. If you would try to balance a histogram on a pin, you would have to position the pin at the mean in order to succeed.

The median is the point where the distribution is cut into two parts of the same area.

In a symmetric distribution mean and median are equal.
In a positively skewed distribution the mean is greater than the median.
In a negatively skewed distribution the mean is smaller than the median.

2.2.2 Describing the variability (spread) of a distribution

It is **not** enough just to report a number that describes the center of a sample. The spread, the variability, in a sample is also an important characteristic of a sample.

Definition The range of a sample is the difference between the largest and the smallest value in the sample.

Usually the greater the range the larger the variability. However, variability depends on more than just the distance between the two most extreme values. It is a characteristic of the whole data set and every observation contributes to it.

```
Sample 1:  *   *   *   *   *   o   *   *   *   *   *
           ---------------------------------------
Sample 2:  *               ****O****               *
```

Definition
The n deviations from the sample mean are the differences

$$x_1 - \bar{x}, x_2 - \bar{x}, \cdots x_n - \bar{x}$$

A specific deviation is greater than zero if the value is greater than $\bar{x}$ and negative if it is less than $\bar{x}$.
The set of deviations describes the variability of the data set, but $\sum_{i=1}^{n}(x_i - \bar{x})$=0.

Square every deviation before summing them up and you will receive a number that characterizes the variability in the data set.

Definition
The sample variance, denoted by s^2, is the sum of squared deviations from the mean divided by $n-1$. That is

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

The sample standard deviation is the positive square root of the sample variance and is denoted by s.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

For calculating the sample variance the following formula can help:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

then s^2 can be calculated by

$$s^2 = \frac{S_{xx}}{n-1}$$

Calculate the standard deviation of the 4 battery lives.

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
1	5.9	-0.475	0.226	34.81
2	7.3	0.925	0.856	53.29
3	6.6	0.225	0.051	43.56
4	5.7	-0.675	0.456	32.49
Σ	25.5	0	1.589	164.15

The sample variance is $s^2 = 1.589/3 = 0.530$ and the sample standard deviation is $s = \sqrt{0.530} = 0.728$.

Using S_{xx}, first calculate

$$\begin{aligned} S_{xx} &= \textstyle\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \\ &= 164.15 - \frac{25.5^2}{4} \\ &= 164.15 - \frac{650.5}{4} \\ &= 1.59 \end{aligned}$$

With this we get $s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{1.59}{3}} = 0.73$.

Whenever we measure the center of a sample by the mean, we measure the spread by the standard deviation. As well as the mean the standard deviation is no resistant to outliers, so we are in need of a value that measures the spread and is resistant to outliers.

2.2.3 Interpreting the standard deviation

Chebychev rule One way to evaluate the distance to the mean measured in standard deviations is given by the Chebyshev's Rule.

Chebyshef's Rule:
Consider any number $k \geq 1$,

$\Rightarrow$ the percentage of observations that are within k standard deviations of the mean is at least $100(1 - 1/k^2)\%$.

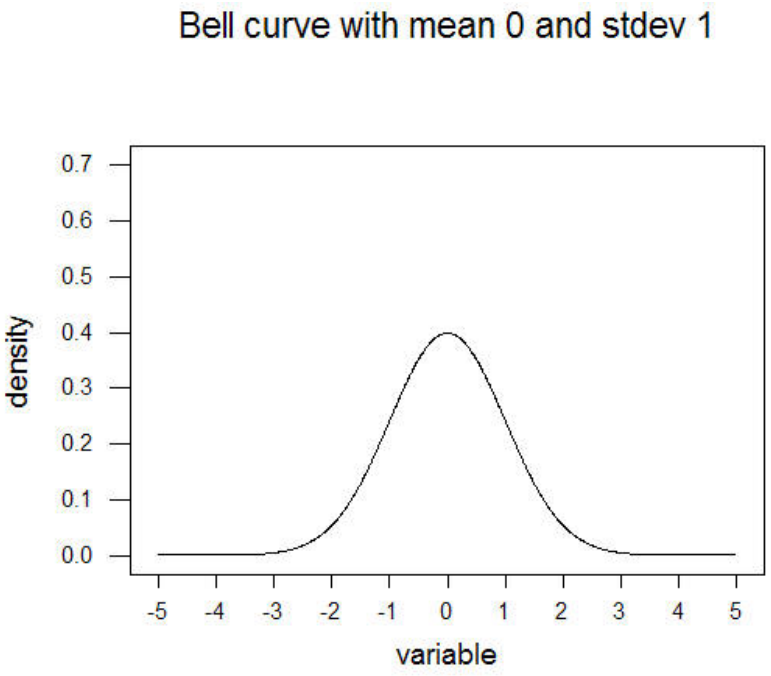
The following table illustrates the conclusion to this rule.

k	$1 - \frac{1}{k^2}$	percentage within k std dev of the mean
2	1-1/4=3/4=0.75	at least 75%
3	1-1/9=8/9=0.89	at least 89%
4	1-1/16=0.94	at least 94 %

One can conclude that at least 75% of the observations will be in the interval [mean - 2 std dev, mean + 2 std dev].
Because this rule applies to all possible data sets it is very conservative. Most data sets, will have much more than 75% observations in the interval $[\bar{x} - 2s, \bar{x} + 2s]$.

The Empirical Rule

The Empirical Rule applies to data which histograms are approximately mound shaped.



Empirical means deriving from practical experience, which has shown that the normal curve provides a good model for a lot of different kind of variables. Physical measurement can usually be well described by a normal curve.

Empirical Rule

If the histogram of values in a data set is mound shaped, then

approximately 68% of the observation are within 1 standard deviation of the mean.
approximately 95% of the observation are within 2 standard deviation of the mean.
approximately 99.7% of the observation are within 3 standard deviation of the mean.

The Empirical Rule gives *approximate* numbers, where the Chebyshev Rule gives the *minimal* numbers.

z–Score
The z–score is a measure of relative standing.
If you obtained a score in an achievement test you might want to know your standing in relation to other people who have taken the test. Are you below or above the mean, how far above or below in relation to the other people.

The z–score helps finding the position of a particular observation in data set.

Definition:

The z–score corresponding to a particular observation in a data set is

$$\text{z} - \text{score} = \frac{\text{observation} \ - \ \text{mean}}{\text{standard deviation}} = \frac{\text{x} - \bar{\text{x}}}{\text{s}}$$

The z–score tells how many standard deviations the observation is from the mean. This is also called standardization. The z–score is a standardized score.

The z–score is especially useful for distributions of observation that are approximately normal. In this case, a z–score outside the interval from -2 to 2 only occurs in 5% of the cases and a z–score outside the interval of -3 to 3 only in 0.3% of the time.

Example:
Two graduate students:
accounting major gets job offer for $ 35000
advertising major gets job offer for $ 33000

accounting: $\bar{x} = 36000$ and $s = 1500$
advertising: $\bar{x} = 32.500$ and $s = 1000$

acc $z - score = \frac{35000-36000}{1500} = -0.67$
adv $z - score = \frac{33000-32500}{1000} = 0.5$

The advertising major can be happier about the job offer than the accounting major.
The distance to the mean of a specific observation measured in standard deviations gives information about the location of this observation in relation to the other observations in the sample.

Example: Consider a data set of resting pulses, with mean $\bar{x} = 72$ and std dev $s = 11$.
A pulse of 61=72-11 is one std dev below the mean and 83 is one std dev above the mean.
A pulse of $50 = 72 - 2 \cdot 11$ is two std dev below the mean and 94 is two std dev above the mean.

2.2.4 Percentiles

Definition:
A set of n measurements on the variable x has been arranged in order of magnitude. Let p be a number between 0 and 100. The pth *percentile* is the value x_p that exceeds p% of the measurements and is less than the remaining (100-p)%.

- The median is the 50^{th} percentile,

Definition:
The *lower quartile* Q_1 is the 25th percentile, that is, it separates the bottom 25% of the measurements from the top 75%.
The *upper quartile* Q_3 is the 75th percentile, that is, it separates the top 25% of the measurements from the bottom 75%.

The middle 50% of the measurements fall between the lower and the upper quartile.

How to calculate the quartiles?

- Order the measurements in order of magnitude.
- Q_1 is the median of those measurements that fall below the overall median.
- Q_3 is the median of those measurements that fall above the overall median.

Example:
Find the upper and lower quartiles of the following 6 numbers: 4 11 15 17 25 33

The median of these numbers is (15+17)/2=16.

Then Q_1 is the median of 4 11 15, which is 11.

And Q_3 is the median of 17 25 33, which is 25.

Definition:
The interquartile range (IQR) for a set of measurements is the difference between the upper and the lower quartile:
IQR=$Q_3 - Q_1$.

The IQR for the example above equals IQR=25-11=14.

2.2.5 The Five-Number-Summary and Boxplots

Definition:
The five-number-summary of a set of measurements consists of the
minimum, the first quartile, the median, the third quartile, and the maximum.

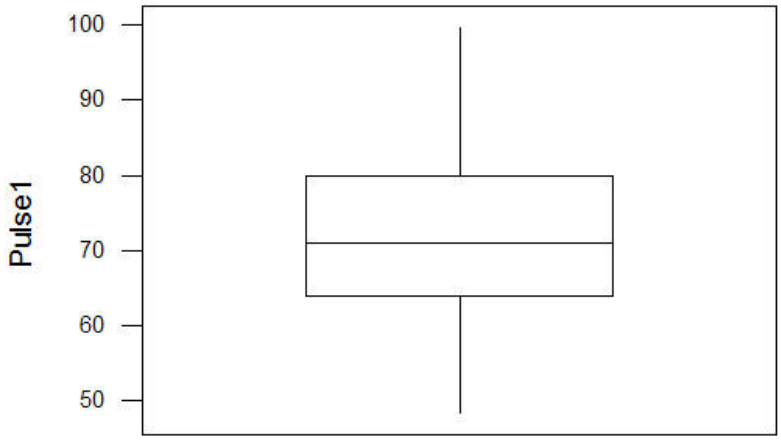
These numbers give a good summary of a distribution of quantitative observations.

The boxplot is a powerful graphical tool for summarizing data It shows the center, the spread, and the symmetry or the skewness at the same time. It is based on the five-number-summary.

Construction of a boxplot

1. Draw a horizontal or vertical measurement scale.
2. Draw a rectangular box, whose lower edge is at the lower quartile and whose upper edge is at the upper quartile.
3. Draw a line segment inside the box at the location of the median.
4. Add line segments from each end of the box to the smallest and largest observation in the data set.

Example: Sample of resting pulse of size 92. median= 71.0, min=48, max=100, Q_1=64, Q_3=80



A boxplot can be supplied with even more information. Sometimes a star '*' is added for the mean. This will help to give a visual comparison between mean and median. In addition outliers may me identified in the boxplot. In order to do this, we first have to define, what numbers we will consider to be outliers.

Definition: An observation is an outlier if it is falls more than 1.5 IQR above the third quartile or below the first quartile.

In order to determine outliers, calculate an upper and a lower fence:

- upper fence=$Q_3 + 1.5 \cdot$IQR, every measurement **above** the upper fence is an outlier.
- lower fence=$Q_1 - 1.5 \cdot$IQR, every measurement **beneath** the lower fence is an outlier.

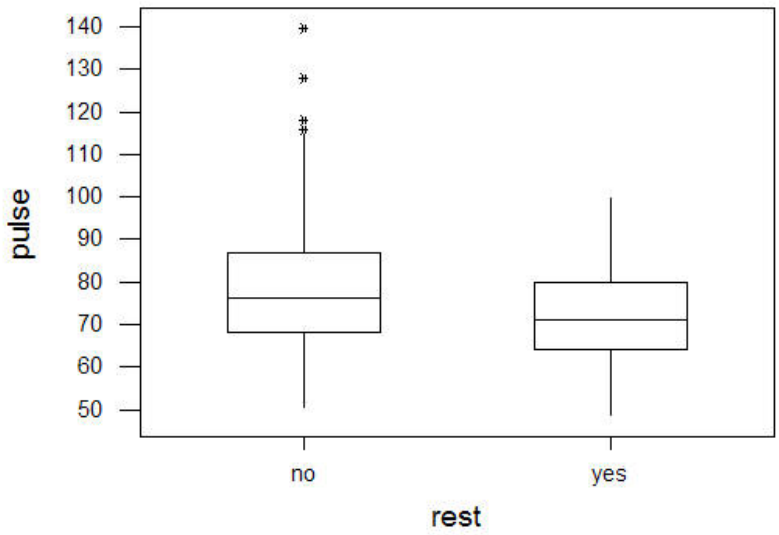
Example: The IQR in the example is 80-64=16. (1.5 *16)=24.

- upper fence = 80+24=104. Since the maximum equals 100, there is no upper outlier.
- lower fence =64-24=40. Since the minimum equals 42, there is no lower outlier.

Outliers may be marked by a circle or a star in a box plot. In this case the whiskers only extend to the smallest and largest non outliers.

One can create comparative boxplots by drawing several boxes in one graph. This is a good tool for comparing continuous variables in different categories.

Example: Resting pulse and pulse after exercise boxplots in one graph.



Choosing a summary

The five number summary is usually better than the mean and the standard deviation for describing skewed distributions.
Only use the mean and the standard deviation for reasonably symmetric distributions.

3 Looking at data – Relationships

3.1 Two categorical variables

Most of the time graphs are used for the visual comparison of two or more groups in regard to a specific characteristic. We can use bar charts for displaying bivariate categorical data.
For this purpose use a clustered or stacked bar graph. This kind of graph is created by drawing two or more bar charts in one set of horizontal and vertical axes.
The following *two way table* displays the bivariate data for two qualitative variables.

Example 7

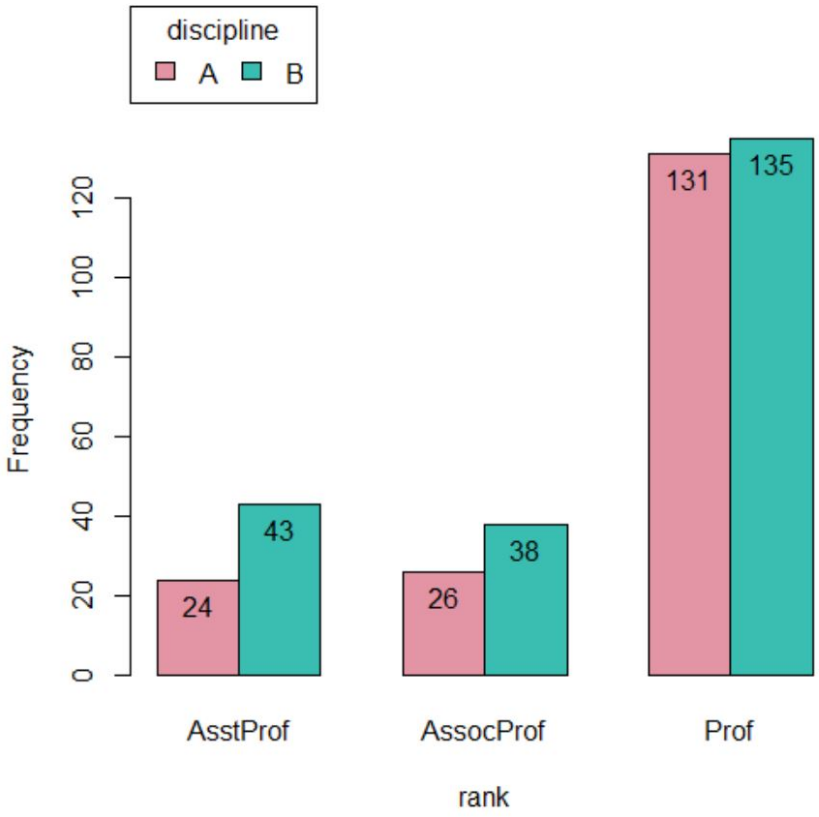
Floor	gym	
	yes	no
wood	30	90
linoleum	1	19
carpet	40	80
tiles	4	6

Remark: Pie charts may also be used. In putting several pie charts beside each other you can obtain a graphical representation of two categorical data variables.

Example 8
We are using a data set ("Salaries" in the package "carData") that is available within R. It includes the 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.
The data set includes the following variables:

- rank (Assistant Prof, Associate Prof, Prof)
- discipline (A= theoretical department, B= applied department)
- yrs.since.phd
- yrs.service
- sex (female, male)
- salary (nine-month salary, in dollars US)

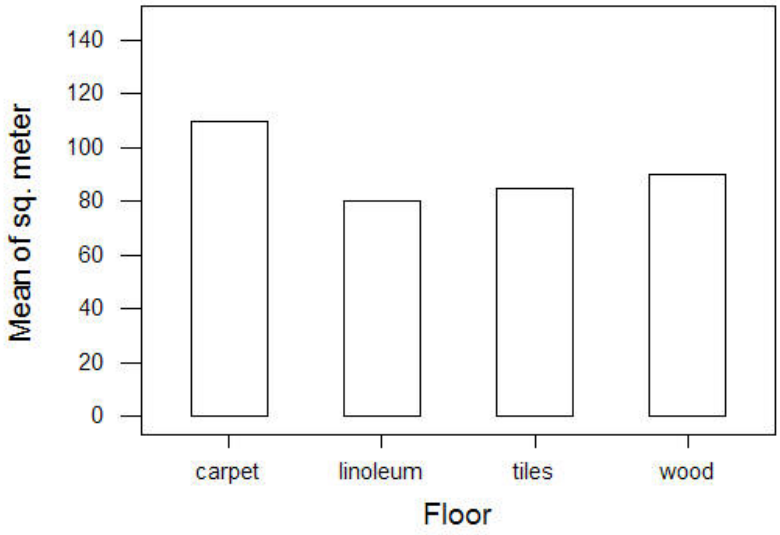
Consider the relationship between discipline and rank



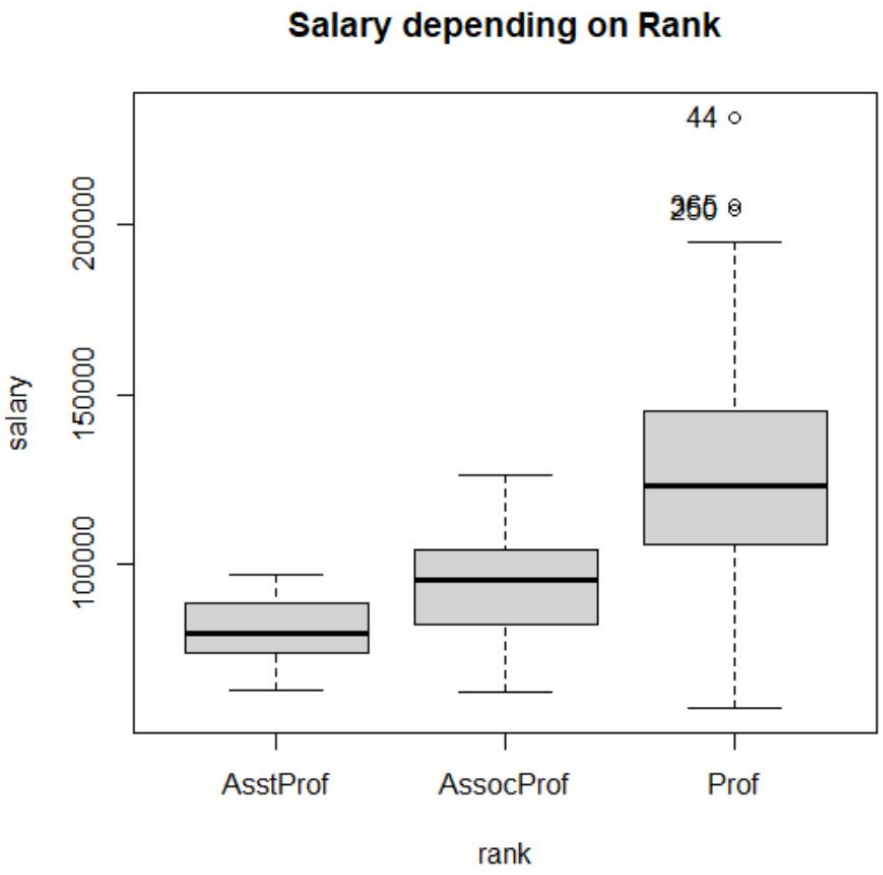
It seems as if in the applied disciplines ("B") there are relatively more Assistant and Associate Professors than in the theoretical disciplines.

3.2 One categorical, one numerical variable

If summarizing one numerical variable depending on one categorical variable you will calculate mean, standard deviation, median, range, interquartile range, etc. for the numerical variable for all observations that fall into one category. That is the data set will be broken apart according to the categorical variable and then for all those subsets a numerical summary of the numerical variable would be calculated.
For a graphical display bar charts can also be used. They are used to visualize the mean of a numerical variable in different subgroups. In this case the height of the bars are proportional to the mean in the corresponding subgroup of the data.



Example 9
How is salary related to the rank of faculty members?
rank is categorical, and salary numerical. The "best" graph is a side by side box-plot.



We can clearly see that salary goes up as the rank increases, but some profs have salaries similar to Assistant and Associate Profs. There are three outliers at the Professor level (these could

be upper administration(deans, provost)).

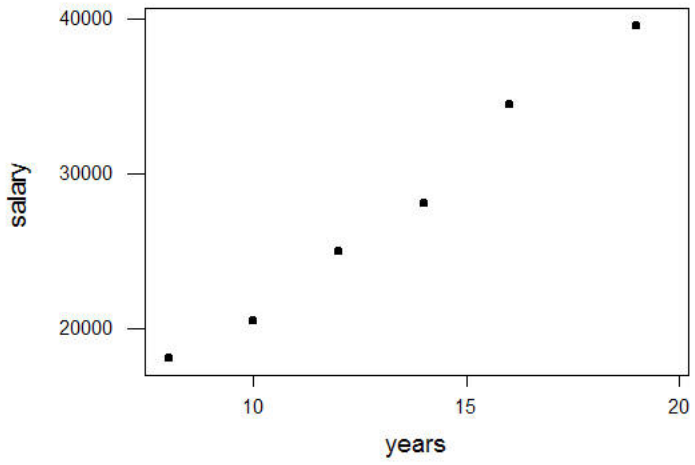
3.3 Two quantitative variables

Create a scatterplots to visualize the relationship between two quantitative variables and evaluate the **quality** of the relationship.

Example:
Does the number of years invested in schooling pay off in the job market?
Apparently so – the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed.

x=Years of Schooling	y=Salary (dollars)
8	18,000
10	20,500
12	25,000
14	28,100
16	34,500
19	39,700

Start of with creating a scatterplot for X and Y.



The scatterplot shows a strong, positive, linear association between years and salary.

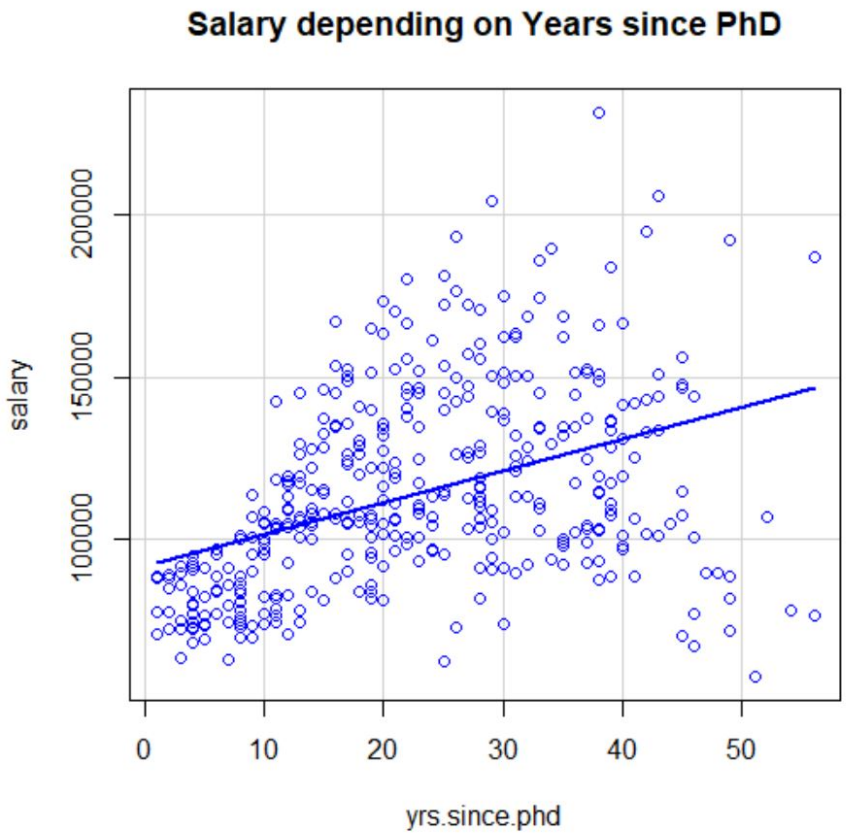
Questions to be answered with the help of the scatterplot:

1. Does a relationship exist that can be described by a straight line (which means is there a linear relationship)?

2. Is there a relationship, that is not linear?

3. If the scatterplot of the variables look like a cloud there is no relationship between both variables and one would stop at this point.

Example 10
Back to the Salaries example.
Rank was not everything that seemed to explain the salary. How about the time since a faculty member completed their PhD?



Is the line a good representation of the data?