Descriptive Statistics

Anna Helga Jónsdóttir
Sigrún Helga Lund

Prepared at the University of Iceland

July 2020

Main topics:

1. Graphical representation of discrete variables
2. Graphical representation of continuous variables
3. Shapes of histograms and outliers
4. Statistic
5. Statistics that describe central tendency
6. Statistics that describe spread
7. Five-number summary and boxplots

Graphical representation

- The first step in any statistical analysis should be to look at the data visually.
- The essence of statistical analysis is to understand the nature of the measurements that are investigated.
- The variability of the data is a key issue - how are the measurements distributed?
    - How much difference do we notice in the outcomes of our subjects?
    - How are the outcomes distributed?
- Graphical representation is the best way to understand the nature of the distribution of the measurements.
- The graphs that we will see in this lecture will only show the measurements of one variable at a time and we make a distinction whether the variable is continuous or discrete.

Graphical representation of discrete variables

- The most common form of graphs for discrete variables are the **bar chart** and the **pie chart**.
- Pie charts are commonly used in business and media but are rare in journals and books in natural sciences.
- Bar charts are commonly seen and they are better suited then pie charts to show the measurements of discrete variables. than pie charts.

Bar chart

Bar chart

A bar chart consists of two or more bars. The number of bars is determined by the number of categories/values that the discrete variable takes. Every bar represents one category/value and they may not lie close to each other. The height of the bars shows the frequency of the corresponding category. The bars shall be ordered in an informative way, often by size.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable and how many subjects belong to each category.

Pie chart

Pie chart

When making a pie chart, it is important that all categories/values of the variable under investigation are pictured on the chart. The number of slices in the pie chart is determined by the number of categories/values of the variable. The size of the slice is determined by the proportional number of subjects in the corresponding category compared to the whole sample. Watch out that the ratios add up to 100 %.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable, how many subjects belong to each category and the corresponding percentage of subjects in that category as a proportion of the whole sample.

PIE CHARTS ARE NOT DRAWN BY HAND!

Graphical representation of continuous variables

- The most common method to visualize continuous variables are **histograms**.
- A **Box plot** is also a good method to visualize continuous variables and they will be introduced in the lecture about descriptive statistics (Lecture 40).
- **Scatter plots** will be introduced in the lecture about linear regression (Lecture170) but they are used to explore the relationship between two continuous variables.

Histograms

- Histograms are similar to bar charts but the main difference in their appearance is that there is no gap between the bars in a histogram.
- It is slightly more difficult to make a histogram then a bar chart as continuous variables do not contain real categories or groups.
- First one has to define groups (intervals) before one counts how many measurements belong to each group.
- When the groups have been made it is often useful to make a table that contains the groups and how many measurements lie within each group.

Histograms

Histogram

A histogram consists of bars that are lined continuously one by another. The number of bars is determined by the number of groups (intervals) that the continuous variable is split up into. When the groups are made is is often good to keep in mind the following criteria:

- Lower and upper limits should be simple and easily understood.
- The intervals may not overlap and must cover all measurements.
- The intervals should be equally wide.
- The number of intervals should be appropriate. A rule of thumb is that the number of intervals should be approx. 5 times the logarithm of the total number of measurements.

When the intervals have been made, one bar is drawn for each interval and the height of the bar is determined by the number of measurements within the corresponding interval.

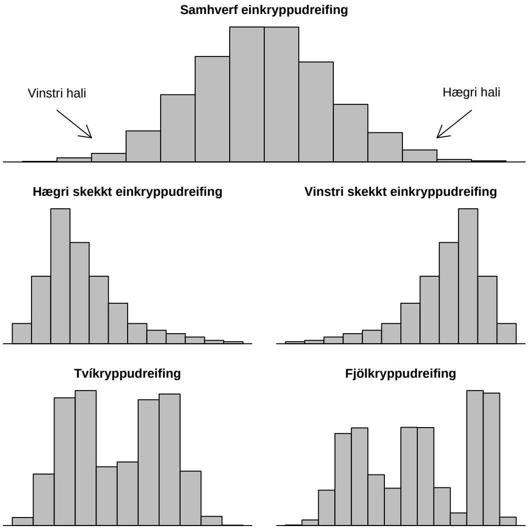
Shapes of distributions

Shapes of distributions

The following concepts are used to describe the distribution of measurements.

- The distribution of the smallest measurements are called the *left-tail* of the distribution. The distribution of the largest measurements is called the *right-tail* of the distribution.
- A distribution is *symmetric* if its right-tail is distributed as the mirror image of the left-tail.
- A distribution that is not symmetric is *skewed*. A distribution is *skewed to the right* if its right-tail is longer then the left-tail and *skewed to the left* if the left one is longer then the right one.
- If a distribution has one peak it is referred to as *unimodal*.
- If a distribution has two peaks it is referred to as *bimodal*.
- If a distribution has more then two peaks it is referred to as *multimodal*.

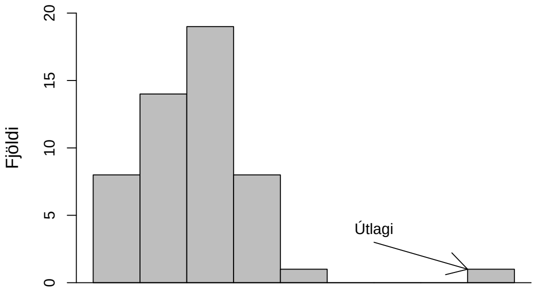
Shapes of distributions



Outliers

Outliers

Outliers are measurements that differ greatly from other measurements in the sample. There can be various reasons for outliers and it is important to look at them specifically and consider their cause.



Statistic

- Descriptive statistics are used to describe certain properties of our measurements.
- A **statistic** is a good way to do so, but it is a number that describes a certain property of our measurements.

Statistic

A **statistic** is a number that is calculated in a some particular way from our measurements.

- There are many different types of statistics that describe different properties of measurements.
- Let us look closer at two types of statistics:
  - Statistics that describe the **central tendency** of the measurements.
  - Statistics that describe the **spread** of the measurements.
- The two most common statistics are the mean and the standard deviation.

Statistics that describe central tendency

Now look at five different statistics that all describe the central tendency of measurements

1. Mid range
2. Mode
3. Median
4. Mean (arithmetic mean)
5. Weighted mean

Mid range

Mid range

Assume that we have n measurements $x_1, x_2, ...x_n$. Let x_{min} denote the smallest one and x_{max} denote the largest one. The **Mid range** is calculated with

$$\text{Mid range} = \frac{x_{\text{min}} + x_{\text{max}}}{2}.$$

Mode

Mode

Assume that we have n measurements, $x_1, x_2, ...x_n$. The **mode** is the most frequent outcome among the measurements. It is the only statistic that describes central tendency that can be used for categorical data. It is on the other hand inappropriate to use the mode to describe continuous variables.

Median

Median

Assume that we have n measurements, $x_1, x_2, ...x_n$. Arrange the measurements by order from the smallest measurement to the largest. Then calculate

$$\text{index number} = 0.5 \cdot (n+1).$$

The **median** is often denoted by M. It depends on whether n is an odd or an even number how we calculate the median.

- If n is an odd number, then the median is the measurement with the index number $0.5 \cdot (n+1)$.
- If n is an even number then the median is the average of the two numbers that have index numbers next to $0.5 \cdot (n+1)$

CAUTION: $0.5 \cdot (n+1)$ is the index number for the measurement, not the median itself!

Mean

Mean (arithmetic mean)

Assume that we have n measurements, $x_1, x_2, ...x_n$. The **mean** is calculated by adding all of the measurements together and divide by their number.

$$\bar{x} = \frac{x_1 + x_2 + ...x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Weighted mean

When the mean is calculated as described in the previous slide, all measurements have the same weight. In some cases we want to give different weights to the measurements. Then we calculate **weighted mean**.

Weighted mean

Assume that we have n measurements, $x_1, x_2, ...x_n$ and their weights $w_1, w_2, ...w_n$. The weighted mean is calculated as

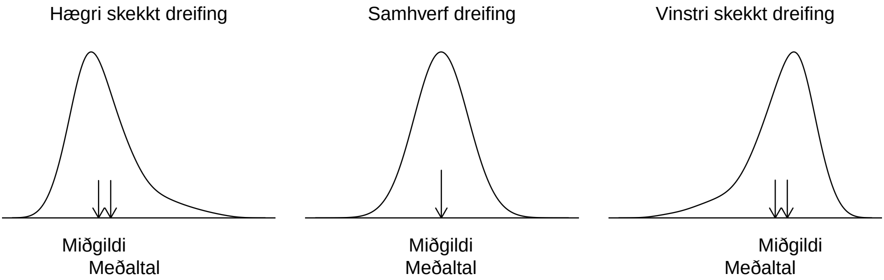
$$\bar{x}_w = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + ...w_n \cdot x_n}{w_1 + w_2 + ... + w_n} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Comparison of statistics that describe central tendency

- All of the statistics previously mentioned are interesting and good to calculate when looking at new data
- Before deciding which statistic is appropriate to use it is good to look at graphical representations of the data in order to get a picture of the distribution of the measurements
- If the distribution is skewed bimodal or multimodal the median shall be used rather then the mean
- The median should also be preferred if there are outliers in the measurements

Comparison of the median and the meani

- If the distribution is skewed to the right, the mean is higher then the median.
- If the distribution is symmetric, the median and the mean are the same.
- If the distribution is skewed to the left, the mean is less then the median.



Statistics that describe spread

Spread

The **spread** of measurements describes how spread out the measurements are.

We will discuss 6 statistics that describe the spread of measurements

1. Range
2. Quartiles
3. Interquartile range
4. Percentiles
5. Variance
6. Standard deviation
7. Coefficient of variation

Range

Range

Assume that we have n measurements, $x_1, x_2, ...x_n$ and let x_{min} denote the smallest one and x_{max} the largest one. The **range** of the measurements is calculated with

$$\text{Range} = x_{\text{max}} - x_{\text{min}}$$

Quartiles

There are three **quartiles** and they are commonly named Q_1, Q_2 og Q_3. They are often sometimes denoted with $Q_{25\%}, Q_{50\%}$ og $Q_{75\%}$. The former notation will be used.

-

Q_1: The first quartile is such that 25% of the measurements are lower then Q_1. Q_1 is therfore the median of the lower half of the measurements, excluding the median.

Q_1: The second quartile is such that 50% of the measurements are lower then Q_2. Q_2 is therefore the median, $Q_2 = M$.

Q_1: The third quartile is such that 75% of the measurements are lower then Q_3. Q_3 is therfore the median of the upper half of the measurements, excluding the median.

Quartiles

Quartiles

Assume that we have n measurements, $x_1, x_2, ...x_n$. Arrange the measurements by order from the smallest measurement to the largest. Then calculate

$$Q_1 \text{ - index number: } = 0.25 \cdot (n+1)$$
$$Q_2 \text{ - index number: } = 0.50 \cdot (n+1)$$
$$Q_3 \text{ - index number: } = 0.75 \cdot (n+1)$$

- Q_1 is the measurement with index number $0.25 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.25 \cdot (n+1)$.
- Q_2 is the measurement with index number $0.50 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.50 \cdot (n+1)$.
- Q_3 is the measurement with index number $0.75 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.75 \cdot (n+1)$.

Interquartile range

Interquartile range

Assume that we have n measurements, $x_1, x_2, ...x_n$ and let Q_1 denote the first quartile and Q_3 denote the third quartile. The **Interquartile range** of the measurements is denoted with IQR and calculated with

$$IQR = Q_3 - Q_1.$$

Percentiles

The idea behind **percentiles** is similar to that of quartiles, but instead of only looking at the borders at 25%, 50% or 75% of the measurements any proportion at all can be used.

Percentiles

The $a\%$ percentile is the numbers that has the property that $a\%$ of the measurements have values less then that number.

As with the quartiles, there are several different ways to calculate percentiles, and is almost never done "by hand"but a statistical software used to do so.

Variance

Variance

Assume that we have n measurements, $x_1, x_2, ...x_n$. The **variance** of the measurements is denoted with s^2 and calculated with

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

$s^2 = 0$ if and only if all of the measurements are equal, if not, s^2 is always greater then 0. The further the measurements lie from the mean, the higher s^2 will be.

Standard deviation

Standard deviation

Assume that we have n measurements, $x_1, x_2, ...x_n$. The **standard deviation** of the measurements is denoted with s and calculated with

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

$s = 0$ if and only if all of the measurements are equal, if not, s is always greater then 0. The further the measurements lie from the mean, the higher s will be.

Coefficient of variation

- One should be careful when comparing the standard deviation of measurements that have different units or have very different means.
- In these cases the **coefficient of variation** is used to compare the spread of two or more groups. It is denoted with CV.

Coefficient of variation

The **coefficient of variation** is calculated with

$$CV = \frac{s}{\bar{x}}$$

As CV is higher, the more spread are the measurements.

Comparison of statistics for spread

- All of the previously noticed statistics that describe spread are interesting and good to calculated when looking at new measurements.
- Variance and standard deviation are used to describe the spread of measurements around the mean and should therefore only be used when the mean is used to describe the central tendency.
- Standard deviation is normally preferred to variance as the unit of the standard deviation is the same as for the measurements.
- Standard deviation is sensitive to skewness and outliers. Only few outliers can increase the standard deviation greatly.
- If the measurements are skewed or if there are outliers in the dataset, a **five-number summary** and a **box plot** are the best descriptives for the spread of the measurements.

Five-number summary

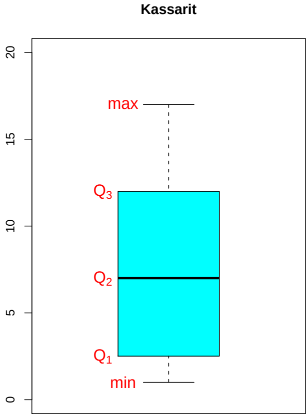
Five-number summary

A **Five-number summary** consists of the smallest value (min), the interquartile ranges and the largest value (max), that is

$$\text{min},\ Q_1,\ Q_2,\ Q_3,\ \text{max}$$

Boxplot

- **Boxplot** is used to look at the central tendency and spread of the data.
- They reflect the measurements well and show clearly whether the distribution is symmetric or skewed.
- If the distribution is skewed, bimodal or multimodal, the five-number summary is the best method to describe central tendency and spread.
- There are a few different versions. The one seen here is the simples one.



Boxplot

Boxplot

- A **boxplot** consists of a box and two lines that go out of the ends of the box. These lines are often called whiskers.
- The box may lie (horizontal) or stand (vertical), we let it stand in our explanation. Then the y-axis shall cover both the smallest and the largest measurements.
- The lower end of the box shall be at Q_1 and the upper end at Q_3 a line should be drawn through the box in Q_2.
- The lower whisker should touch the lowest measurement (min) and the upper whisker should touch the tallest measurement (max).

1.5 · IQR rule for outliers

- Outliers are measurements that differ greatly from other measurements and are therefore important to identify.
- One method of determining whether a measurement is an outlier is to compare the distance of its value to the next quartile (Q_1 or Q_3).

1,5 · IQR rule for outliers

- First we calculate the distance of the potentially outlying measurement to the next quartile (Q_1 or Q_3).
- This distance is then compared to the IQR. If the distance of the measurement to the next quartile is more then $1.5 \cdot IQR$ the measurement is considered an outlier.

IQR rule

- Many statistical programs use the $1.5 \cdot \text{IQR}$ rule when drawing boxplots and these boxplots are often called **modified boxplots**.
- If the highest and/or lowest value reaches further then one and half of the length of the box, the lines that go out of the box, the whiskers, stop at that value, but do not reach all the way to the highest and/or lowest value.
- The measurements outside the whiskers are outliers and labelled with a circle.