S1:

- What is Statistics?
- The field of Statistics consists of two parts:
  - Descriptive statistics: Summarize and describe a set of data.
  - Inferential statistics: a set of techniques which helps decision makers come to rational decisions under uncertainty.
- Hence, statistics is concerned with:
  - Whether data should be gathered.
  - How to gather it.
  - How to analyze it.
- A statistician samples because it is almost the case that obtaining all the data is impossible or very time/money consuming.

  Hence, statisticians sample the population:

  Population: The total collection of data/observations that are of interest.

  Sample: Subset of the population.

- The goal of the statistician is to obtain the desired information at minimum cost (few samples).
- In statistics uncertainty is involved. Hence, the concept of probability is used to measure the amount of confidence one has in various sample results.
- Statistics attempts to infere from samples. This implies that some errors are involved. There are two types of errors:
  - Experimental of sampling errors: Due to large uncontrolled factors. (More sampling, ~~less error~~)
  - Bias: Persistent, ~~consistent~~ systematic sort of error.

Increase sample's size: sampling error ↓, bias stays the same.

Example of bias: Imagine that we want to have an idea of the unemployment in Uppsala. To do so, we perform a sampling by telephone (home). This sampling is done during Wednesdays at noon. We ask them if they are unemployed. If they don't answer the phone, nothing happens. Do you see the bias?

The description of data:

There are a lot of tools and ways to describe the data. Here we will present some.

Suppose that we have a population under study (e.g. Height of the Students at a High School) and we take a sample (200 data). Then we will have a list like

1 → 180 cm
2 → 220 cm
3 → 165 cm
⋮
32 → 152 cm
⋮
199 → 172 cm.

And we want to extract some information. First, what we can do is to sort this list.

1 → 152 cm
2 → 154 cm
3 → 154 cm
⋮
199 → 201 cm
200 → 220 cm.

Sorting the data we get its range: [minimum, maximum]

[152 , 220]

That's nice, it gives us the minimum and maximum values it attains.

- Another quantity that we can get is its mode: the value of the list that appears most often. In our example, let's suppose that it appears 3 times the value 163cm.
- Another is its median: a numerical value that splits the data in the 50% lower half and the 50% upper half.

152, 154, 154, ..., 185, | 187, ..., 201, 220

(50% | 50%)

(In this case, since the number of datum is even

$$\text{median} = \frac{187 - 185}{2}$$

In odd case median = value at the middle).

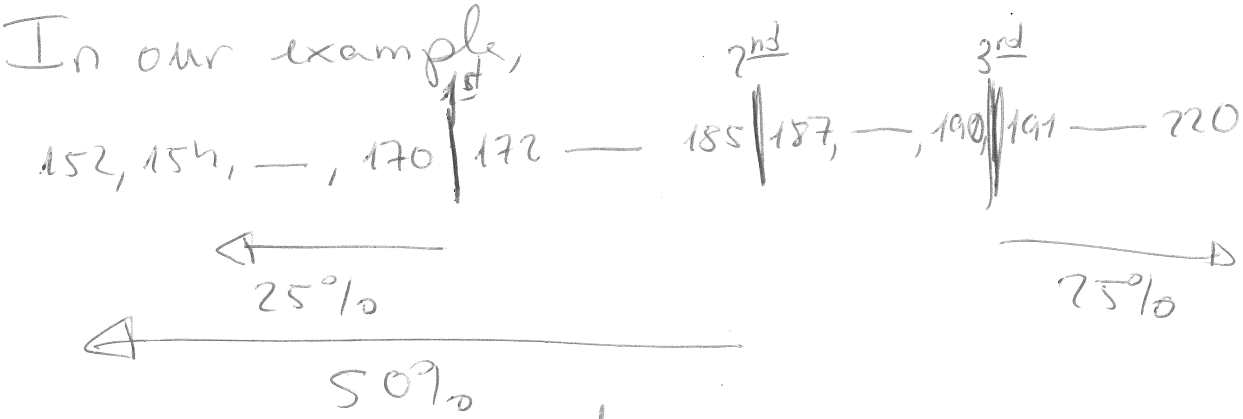
- A refinment of the median are the quartiles:

1st quartile: lower 75%.

2nd quartile: median.

3rd quartile: upper 25%.

In our example,



- Another is the sample mean: It is the sum of all datum divided by the number of data.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right).$$

In our example,

$$\bar{x} = \frac{1}{200}\left(152 + 154 + 154 + \cdots + 201 + 220\right).$$

The mean is the "average" value.

- Another is the sample variance (s^2) and sample standard deviation (s):

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

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

The sample standard deviation is an indicator of how the data spreads:

$s \uparrow \Rightarrow$ more spreading

$s \downarrow \Rightarrow$ less spreading.

Graphing the data:

Graphing the data helps us to visualize how it is distributed.

One way of graphing is by drawing its frequency charts

1) First, we regroup the data in subsets. These are called class intervals. This regruping is helpful because we decrease the number of data and groupp them in more meaningful sets.

In our example, we could do.

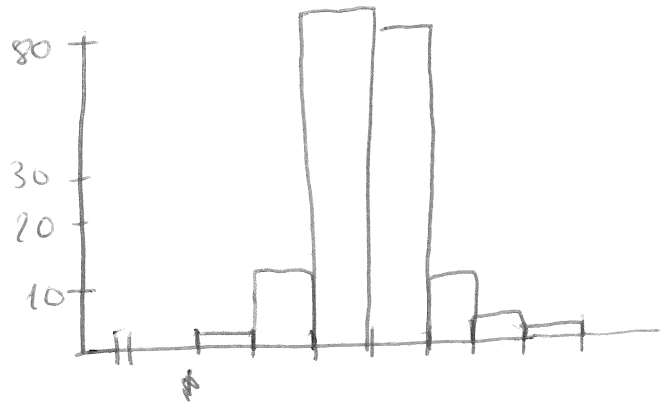
150-160	3
160-170	15
170-180	33
180-190	30
190-200	14
200-210	4
210-210	1

} Frequency distribution

2) Once we have the frequency distribution, we can draw it by means of a

~~Bar chart~~

Histogram:



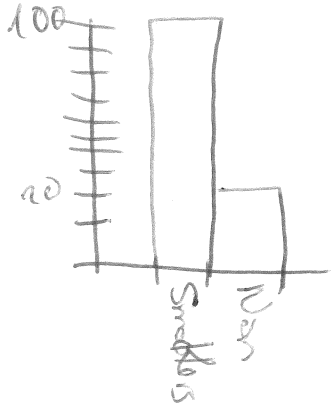
Sometimes the data is not given by quantities (height) but by ~~weight~~ qualities.

Ex:

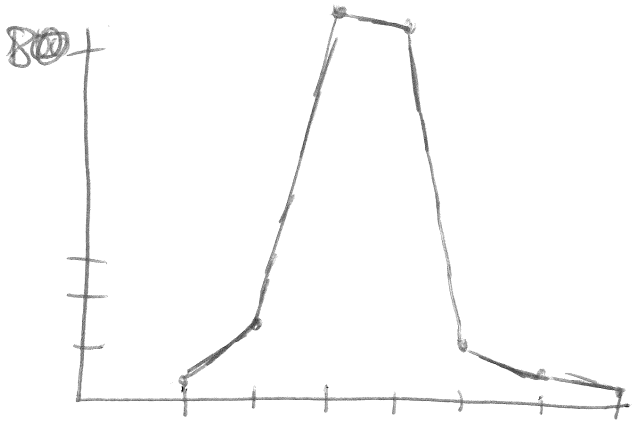
Smokers	100
Non-smokers	22

Then we can draw this data ~~by~~ with a ~~bar~~

bar chart:

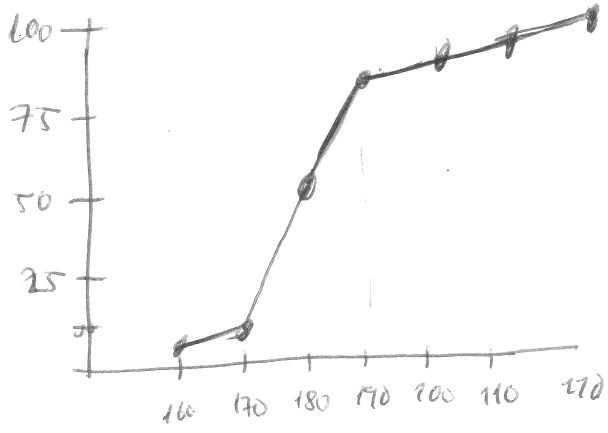


Returning to our previous example (heights), another graph that we can draw is a line diagram:



Another graph that is useful is the cumulative graph: It represents the cumulant across the data.

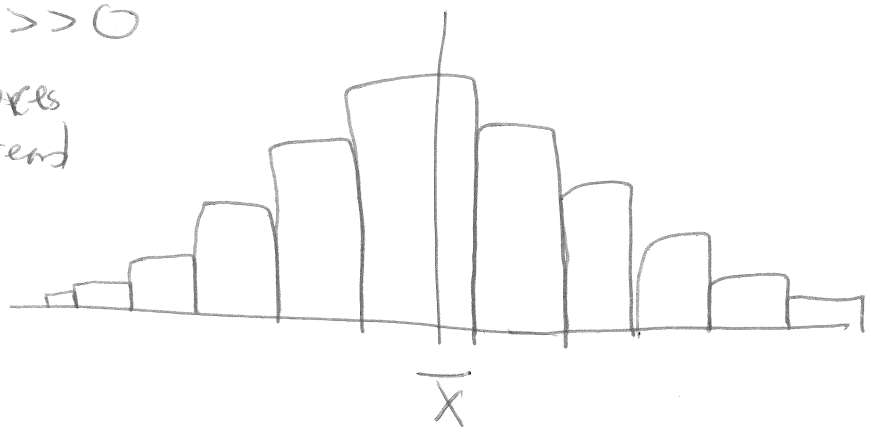
≤ 160	3	1.5%
≤ 170	15	9%
≤ 180	83	50.5%
≤ 190	80	90.5%
≤ 200	14	97.5%
≤ 210	4	99.5%
≤ 220	1	100%



Relation between s and the frequency graph

- $s >> 0$

moves spread



- $s << 1$

less spread.