Statistics

MUS4218 - Metodologisk emne: Kognitiv musikkvitenskap

23 March 2017

What is statistics?

Variables

Something that varies:

- Word length
- Night/day
- Musical tempo
- Perceived expressivity in a musical performance
- Number of students attending a lecture
- Most popular child's name
- Pitch

Independent variables Manipulated by the experimenter

Dependent variables Measured by the experimenter

Scales of measurement

- Nominal
- Ordinal
- Interval
- Ratio

Types of data:

Categorical data			Numerical data (= numbers)	
Two categories	Multiple categories	Ordinal categories	Counting data	Continuous data
Male Female	Violin Piano Drums Guitar Organ	Nothing A little A lot All the time	Number of children	Age
Boolean	Nominal	Ordinal	Discrete	Continuous

Scales of measurement and musical parameters:

- Nominal:
  - Categorical data that cannot be ordered along a continuum.
  - Music example: *Timbre* (klangfarge)
- Ordinal:
  - Data that can be ordered, but without saying anything about the distance of the ordering
  - Music example: *Loudness*. A sound played at 70 dB is louder than the same sound played at 69 dB. But the difference between 69-70 dB and 70-71 dB are the same. How to compare $pp \rightarrow p$ with $f \rightarrow ff$?
- Interval:
  - Can be ordered and the distances between values can be compared.
  - Music example: *Pitch*. The distance between C' and D' is the same as the distance between F' and G'. But there is no 0 tone, so ratios between tones cannot be determined. (What is "twice" of C?)
- Ratio:
  - Interval data with a natural zero point:
  - Music example: *Duration*. A 6 second tone is twice as long as a 3 second tone.

Kvifte (1989): *Instruments and the Electronic Age* for details on the music examples.

Research question

Research question

- How many hours per week do Norwegian students work on their studies?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:
  - with regards to the colors of their clothes?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:
  - with regards to the colors of their clothes?
  - shoe size?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:
  - with regards to the colors of their clothes?
  - shoe size?
  - mood?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:
  - with regards to the colors of their clothes?
  - shoe size?
  - mood?
  - dancing skills?

Research question

- How many hours per week do Norwegian students work on their studies?
- How many people would wote Miljøpartiet de grønne if there was an election tomorrow?
- Does the likelyhood of voting for Pensjonistpartiet increase with age?
- Is there a difference between people who listen to hard rock and funk:
  - with regards to the colors of their clothes?
  - shoe size?
  - mood?
  - dancing skills?

Some of these are complicated (what is mood?)
Some are not that complicated (which party would you vote for?)
Others requre special equipment (e.g. motion capture system).

What do we use statistics for?

Descriptive statistics:

Summarizing and describing an entire data set.

- What is the average number of study hours per week for Norwegian students?
- What is the maximum number of study hours per week for any Norwegian student?

Inferential statistics:

Predicting or inferring something about a large group (population), given only a subset of data (sample)

- How many people would vote Pensjonistpartiet if there was an election tomorrow?
- What is someone's lifetime expectency, given their medical record?

Describing a data set:

Categorical data			Numerical data (= numbers)	
Two categories	Multiple categories	Ordinal categories	Counting data	Continuous data
Male Female	Violin Piano Drums Guitar Organ	Nothing A little A lot All the time	Number of children	Age
Boolean	Nominal	Ordinal	Discrete	Continuous

Describing a data set:

Categorical data			Numerical data (= numbers)	
Two categories	Multiple categories	Ordinal categories	Counting data	Continuous data
Male Female	Violin Piano Drums Guitar Organ	Nothing A little A lot All the time	Number of children	Age
Boolean	Nominal	Ordinal	Discrete	Continuous

Bar charts
Tables/proportions

Histogram / Boxplot
Symmetric (mean and SD)
or
Skewed (median and quartiles)

Describing a categorical data set

Names of children in a kindergarten:
Lena, Kari, Per, Emilie, Kari, Knut, Ole, Lise, Per, Lars, Kari, Anette, Ole

Gender	Count
F	7
M	5

Name	Count
Lena	1
Kari	3
Per	2
Emilie	1
Knut	1
Lise	1
Ole	2
Lars	1
Anette	1
Sum	12

- 7 out of 12 are female ~~(58.3 %)~~
- 3 out of 12 are named Kari ~~(25 %)~~
- 3 out of 7 females are named Kari ~~(42.9 %)~~

Describing a categorical data set

Names of children in a kindergarten:
Lena, Kari, Per, Emilie, Kari, Knut, Ole, Lise, Per, Lars, Kari, Anette, Ole

Gender	Count
F	7
M	5

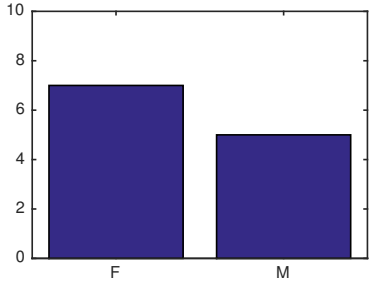
Name	Count
Lena	1
Kari	3
Per	2
Emilie	1
Knut	1
Lise	1
Ole	2
Lars	1
Anette	1
Sum	12

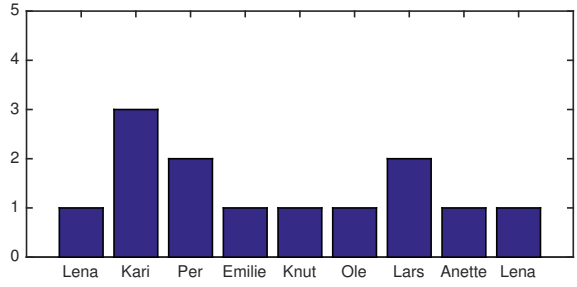
- 7 out of 12 are female ~~(58.3 %)~~
- 3 out of 12 are named Kari ~~(25 %)~~
- 3 out of 7 females are named Kari ~~(42.9 %)~~

(do not use percentages when you have limited data material)

Describing a categorical data set

Visualization with bar charts:





Describing the average value of a numerical data set

Mean (gjennomsnitt/middelverdi) The sum of all the data divided by the number of values in the data
e.g. mean(6 3 20 4 5 1 3) = 42/7 = **6**

Median The mid point of an ordered set of data.
e.g. 1 3 3 **4** 5 6 20

Mode (typetall/modalverdi) The most frequent value in the data set
e.g. 1 **3** **3** 4 5 6 20

Example: Hair length of music students

Hair lengths of all the 14 music students in a class (in cm).

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

Example: Hair length of music students

Hair lengths of all the 14 music students in a class (in cm).

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

$$\begin{aligned} \text{Mean} &\approx 15.18 \\ \text{Mode} &= 5 \\ \text{Median} &= 12.5 \end{aligned}$$

Example: Hair length of music students

Hair lengths of all the 14 music students in a class (in cm).

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

$$\begin{aligned} \text{Mean} &\approx 15.18 \\ \text{Mode} &= 5 \\ \text{Median} &= 12.5 \end{aligned}$$

To find the median, we sort the values:
0.1 0.4 3 5 5 10 11 14 15 20 21 24 34 50

Example: Hair length of music students

Hair lengths of all the 14 music students in a class (in cm).

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

$$\begin{aligned} \text{Mean} &\approx 15.18 \\ \text{Mode} &= 5 \\ \text{Median} &= 12.5 \end{aligned}$$

To find the median, we sort the values:
0.1 0.4 3 5 5 10 11 14 15 20 21 24 34 50
When there is no single mid point, we calculate the mean of the two mid-points

Example: Hair length of music students

What if we want to compare this year's hair length to the hair langth from the previous year?

This year:

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

mean = 15.18

Previous year:

15	14	13	16	12	17	18	17	15	15	16	15	19	10

mean = 15.14

Example: Hair length of music students

What if we want to compare this year's hair length to the hair langth from the previous year?

This year:

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

mean = 15.18

Previous year:

15	14	13	16	12	17	18	17	15	15	16	15	19	10

mean = 15.14

Conclusion:
The hair lengths this year and last year are almost the same.

Example: Hair length of music students

What if we want to compare this year's hair length to the hair langth from the previous year?

This year:

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

mean = 15.18

Previous year:

15	14	13	16	12	17	18	17	15	15	16	15	19	10

mean = 15.14

Conclusion:
The hair lengths this year and last year are almost the same.

OR.....?

Describing the spread of a data set

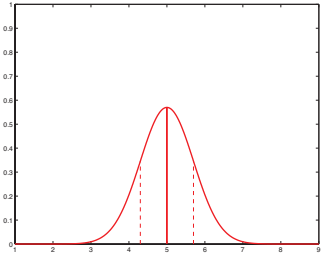
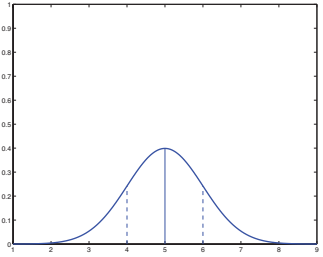
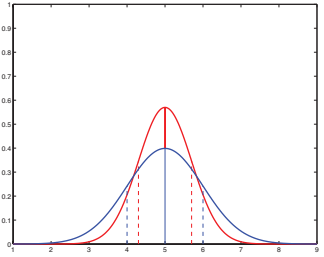
Standard deviation (standardavvik) Mean distance from the mean

Variance (varians) Mean squared distance from the mean

Maximum The highest value in the data set

Minimum The lowest value in the data set

Range (variasjonsbredde) maximum - minimum

Back to the hairy music students

Let us take one more look at this year's hair lengths:

20	3	5	0.1	50	24	5	14	10	11	15	21	34	0.4

Results from LibreOffice/OpenOffice/Excel

20			
3			
5	**What?**	**result**	**function**
0.1	SUM	212.5	sum()
50	MEAN	15.1786	average()
24	SD	14.002	stdev()
5	MEDIAN	12.5	median()
14	MODE	5	mode()
10	VARIANCE	196.056	var()
11	MAXIMUM	50	max()
15	MINIMUM	0.1	min()
21			
34			
0.4			

Equivalent MATLAB functions:

- sum
- mean
- std
- median
- mode
- var
- max
- min

The reason for looking at more than mean values

This year:

20		
3		
5	**What?**	**result**
0.1	SUM	212.5
50	MEAN	15.1786
24	SD	14.002
5	MEDIAN	12.5
14	MODE	5
10	VARIANCE	196.056
11	MAXIMUM	50
15	MINIMUM	0.1
21		
34		
0.4		

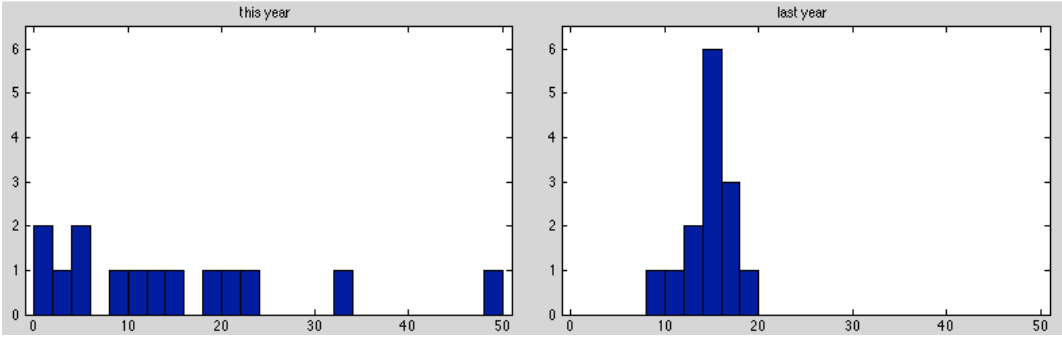
Last year:

15		
14		
13	**What?**	**result**
16	SUM	212
12	MEAN	15.1429
17	SD	2.38125
18	MEDIAN	15
17	MODE	15
15	VARIANCE	5.67033
15	MAXIMUM	19
16	MINIMUM	10
15		
19		
10		

Even though the mean values are roughly the same, the set of hair lengths is completely different. This is why we use *distributions* when describing data sets.

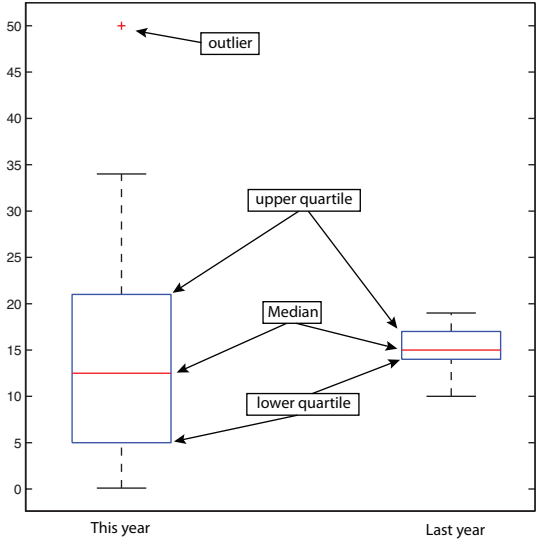
Visualising distributed data

Histograms: visualising a distribution



- Order the data set into *bins*, and count the number of instances in each bin.
- Here the bin size is 2
  (e.g. counting instances between 0 and 2, 2 and 4, 4 and 6, ...)
- Use the matlab function *hist*:
  a = [20 3 5 0.1 50 24 5 14 10 11 15 21 34 0.4];
  hist(a,[1:2:50])
- Easily tells you wether your data set is skewed or symmetric

Boxplots: an easy way to visualise a distribution



- Use the matlab function: boxplot: `boxplot(b)`
- b contains data from this year in column 1 and data from last year in column 2.
- Quartile: 25% of the data is between the median and the quartile

Recap: Describing a data set

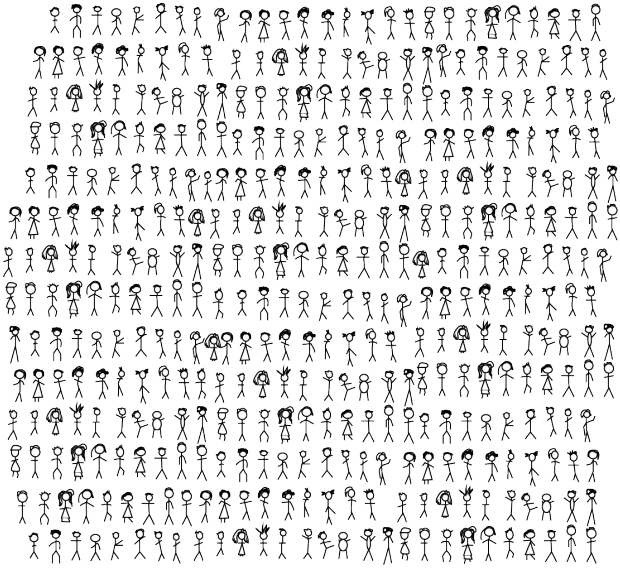
Categorical data			Numerical data (= numbers)	
Two categories	Multiple categories	Ordinal categories	Counting data	Continuous data
Male Female	Violin Piano Drums Guitar Organ	Nothing A little A lot All the time	Number of children	Age
Boolean	Nominal	Ordinal	Discrete	Continuous

Bar charts
Tables/proportions

Histogram / Boxplot
Symmetric (mean and SD)
or
Skewed (median and quartiles)

Population and samples

Inferrential statistics: Population and sample



Sometimes, it is impossible to measure an entire *population* (imagine measuring the hair lengths of all students in the world)

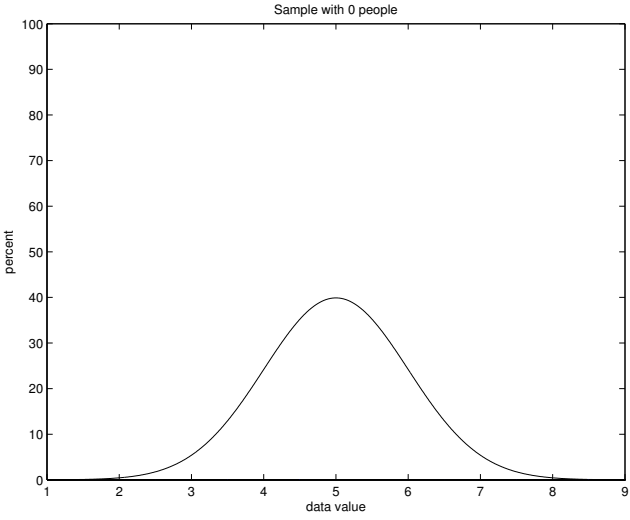
Inferential statistics: Population and sample

Sometimes, it is impossible to measure an entire *population* (imagine measuring the hair lengths of all students in the world)

That is why we use a *sample*, i.e. a subset, of the population, as a basis for our analysis

Sample size and reliability

Increased sample size provides a more accurate distribution:

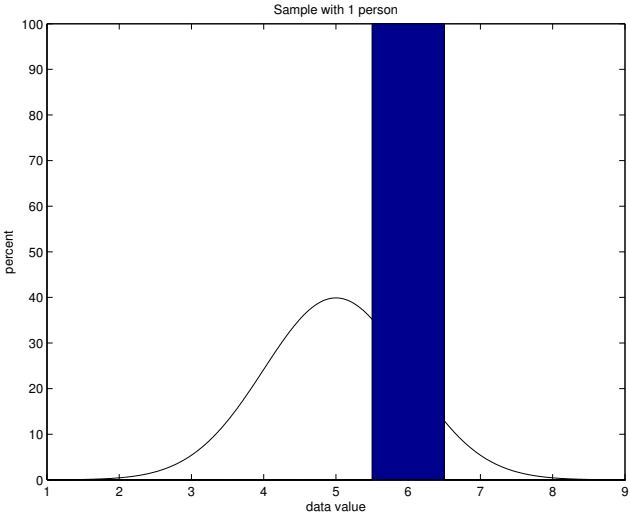


Example: How many slices of bread does a Norwegian eat in a day?

Let's assume that the underlying distribution is
Mean = 5
SD = 1
We cannot know this distribution when we start our experiment

Sample size and reliability

Increased sample size provides a more accurate distribution:

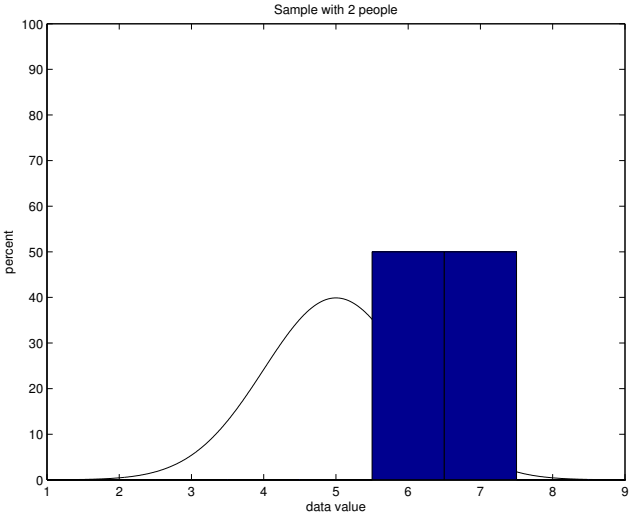


Example: How many slices of bread does a Norwegian eat in a day?

We ask one person who claims to eat 6 slices of bread per day, and so 100% of our answers is 6.

Sample size and reliability

Increased sample size provides a more accurate distribution:

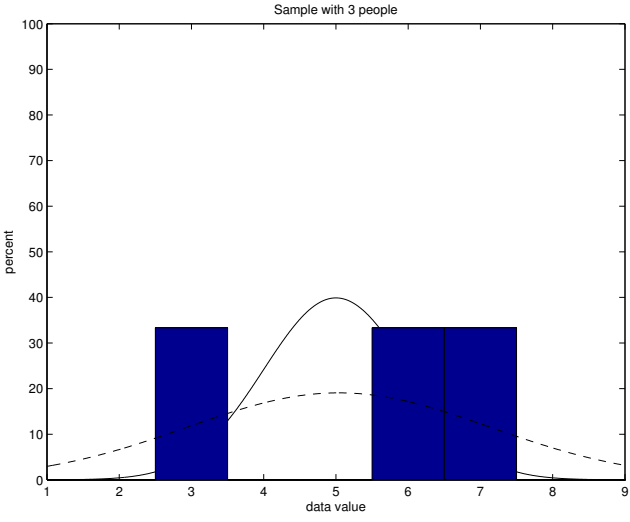


Example: How many slices of bread does a Norwegian eat in a day?

We ask another who answers 7. So far the data suggests that Norwegians eat 6-7 slices of bread per day.

Sample size and reliability

Increased sample size provides a more accurate distribution:

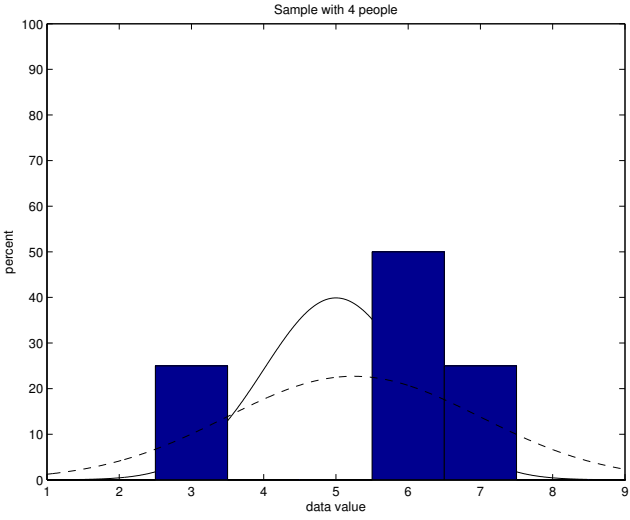


Example: How many slices of bread does a Norwegian eat in a day?

After asking 3 people, we can use mean and SD to estimate the unknown underlying distribution.

Sample size and reliability

Increased sample size provides a more accurate distribution:

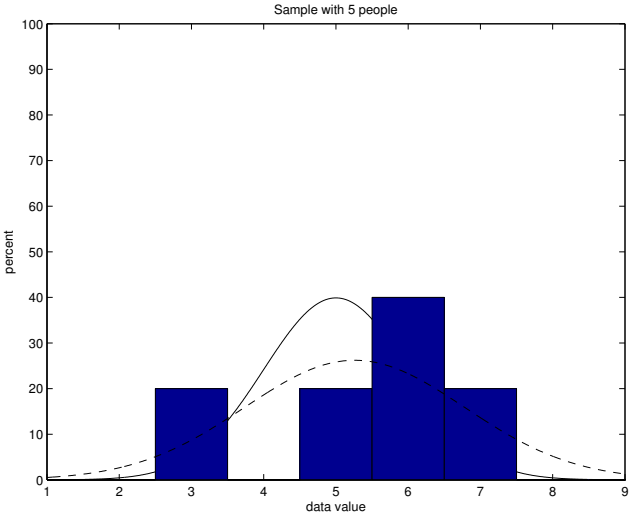


Example: How many slices of bread does a Norwegian eat in a day?

As we keep asking, the distribution comes closer to the true picture.

Sample size and reliability

Increased sample size provides a more accurate distribution:

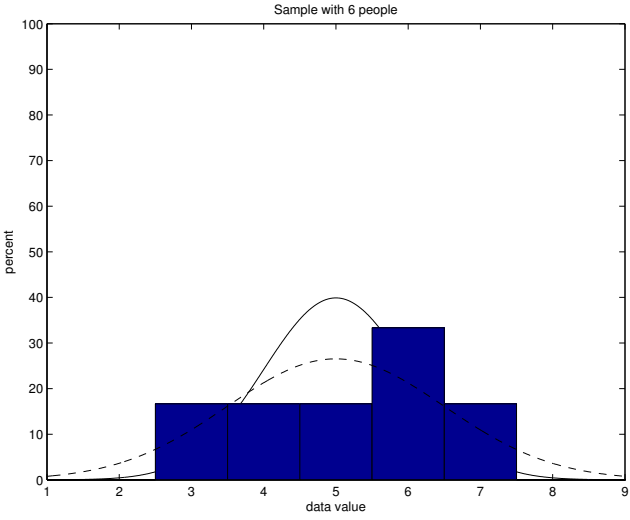


Example: How many slices of bread does a Norwegian eat in a day?

As we keep asking, the distribution comes closer to the true picture.

Sample size and reliability

Increased sample size provides a more accurate distribution:

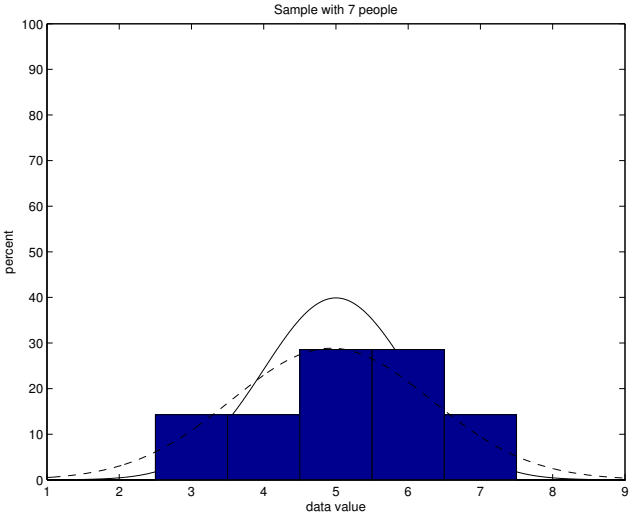


Example: How many slices of bread does a Norwegian eat in a day?

As we keep asking, the distribution comes closer to the true picture.

Sample size and reliability

Increased sample size provides a more accurate distribution:

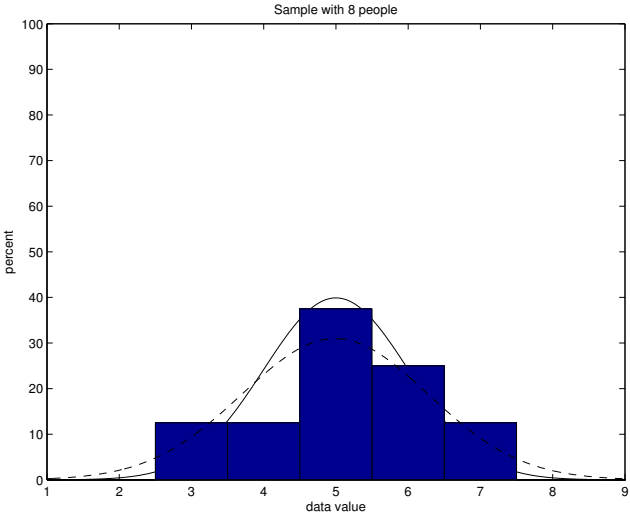


Example: How many slices of bread does a Norwegian eat in a day?

As we keep asking, the distribution comes closer to the true picture.

Sample size and reliability

Increased sample size provides a more accurate distribution:

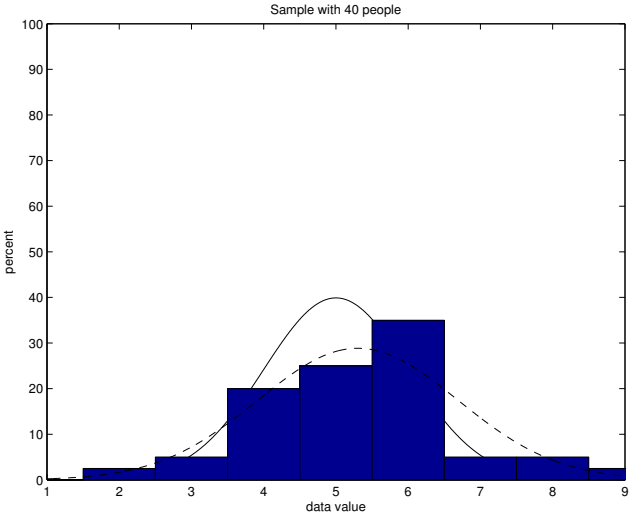


Example: How many slices of bread does a Norwegian eat in a day?

8 people now. Our sample has the following stats:
Mean = 5.13
SD = 1.25

Sample size and reliability

Increased sample size provides a more accurate distribution:

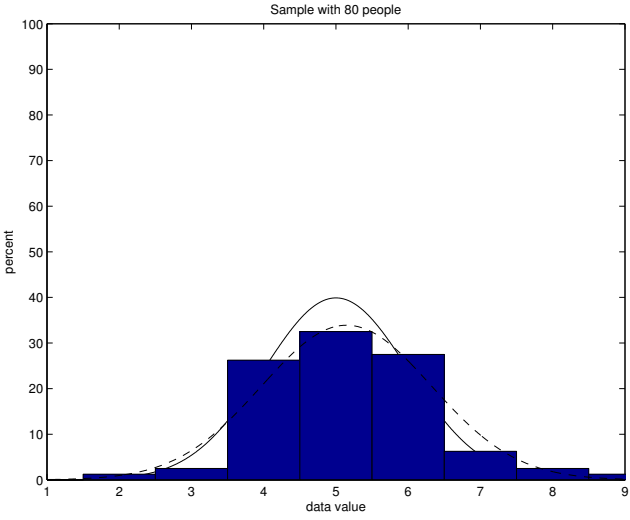


Example: How many slices of bread does a Norwegian eat in a day?

40 people.
Mean = 5.30
SD = 1.38

Sample size and reliability

Increased sample size provides a more accurate distribution:

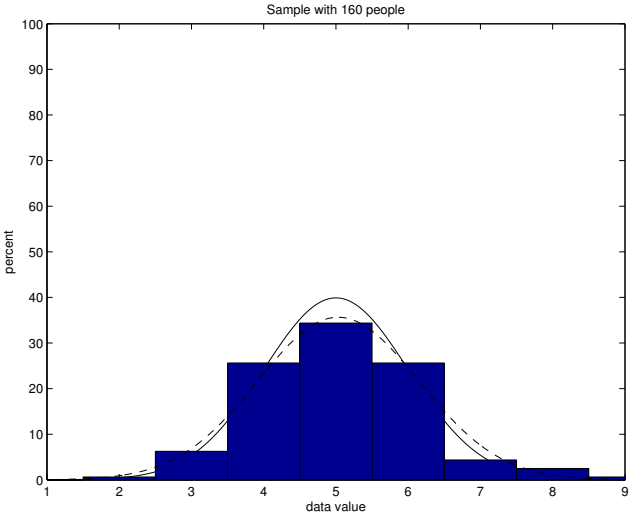


Example: How many slices of bread does a Norwegian eat in a day?

80 people.
Mean = 5.15
SD = 1.18

Sample size and reliability

Increased sample size provides a more accurate distribution:

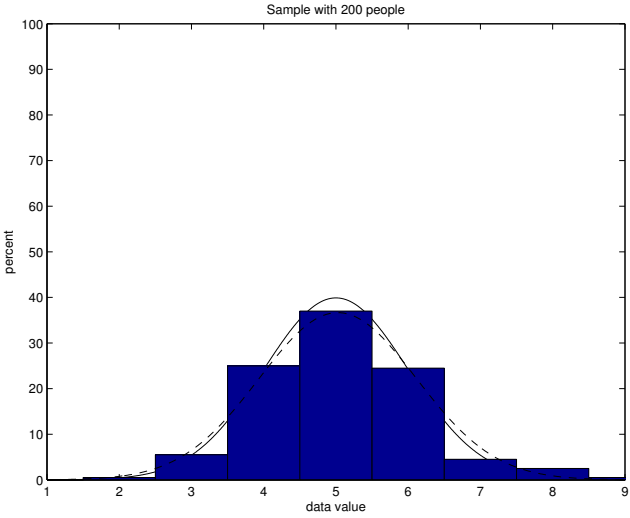


Example: How many slices of bread does a Norwegian eat in a day?

160 people.
Mean = 5.03
SD = 1.12

Sample size and reliability

Increased sample size provides a more accurate distribution:

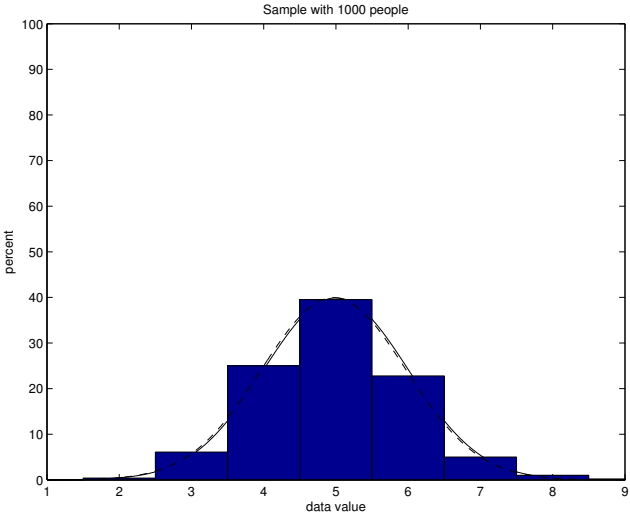


Example: How many slices of bread does a Norwegian eat in a day?

200 people.
Mean = 5.02
SD = 1.09

Sample size and reliability

Increased sample size provides a more accurate distribution:



Example: How many slices of bread does a Norwegian eat in a day?

1000 people. Our sample has the following stats:
Mean = 4.96
SD = 0.999

Example: Questionnaire

Play a music example, and ask people to fill in this questionnaire:

"How sad do you think this song is?"

1) Not sad at all ☐
2) Just a little bit ☐
3) Somewhat sad ☐
4) Quite sad ☐
5) Extremely sad ☐

How can we create descriptive statistics for the gathered data?

Changing the questionnaire

"How sad do you think this song is?"

	1	2	3	4	5	
Not sad at all	☐	☐	☐	☐	☐	Extremely sad

What about the descriptive statistics now?

Be careful when claiming something based on data!

- Make sure that you have data that can be analysed. For example:
  - Don't try to calculate the mean of nominal data
  - Don't use analysis methods intended for ratio data on ordinal data.
- Don't make false claims about your results
  - e.g. If you have an average satisfaction score of 2.0 for group A, and 4.0 for group B, you can claim that group B is more satisfied than A, but NOT that group B is twice as satisfied as group A.

Statistical tests

Statistical tests

- The visualisations and statistic methods we have seen so far are very useful to get an overview of a dataset.
- Many scientific publications also stop there, giving statements like: “As you can see in this plot, student’s hair lengths vary with time.” or “Method A was more effective than Method B”.
- This may be ok, as boxplots may suffice to prove your point
- But if you want to have a more solid fundation for your claims, and be able to say how confident you are in your claims, you should perform a test of *statistical significance*.

H_0: The null-hypothesis

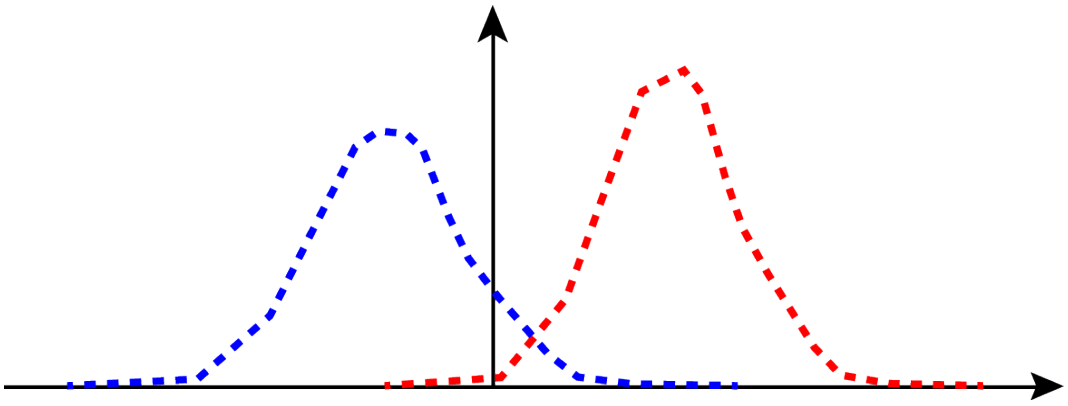
We cannot say for certain that a hypothesis is correct, but if we observe something that contradicts the hypothesis, we can discard (falsify) it.

- Example: Black swans
- Hypothesis: There are no black swans.
  - Cannot be confirmed, but will be faslified if a black swan is observed

In statistics we use the null-hypothesis and use a test of statistical significance to decide whether or not H_0 can be discarded

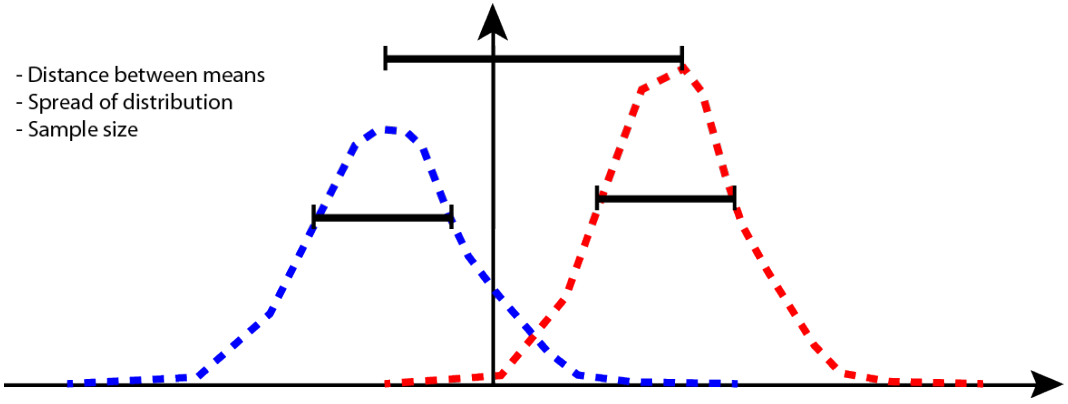
- Hypothesis: There is a difference between A and B
- Null-hypothesis: There is no difference between A and B

Statistical tests



Statistical tests

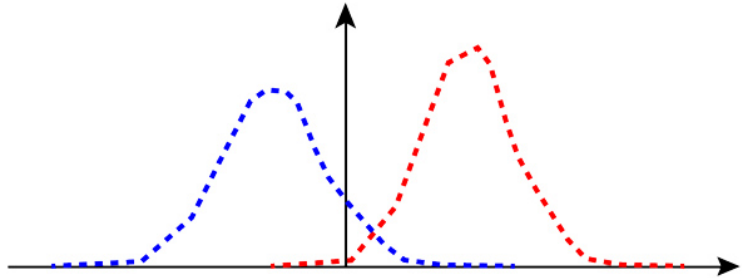
- Distance between means
- Spread of distribution
- Sample size



What statistical results in research may look like

Motion feature	**Comparison**	***df***	***t***	***p***
OnsetAcceleration	Impulsive vs non-impulsive sounds	526	13.65	< 0.01
VerticalVelocityMean	Rising vs falling sounds	284	18.89	< 0.01
AbsAccelerationMean	Pitched vs noise-based sounds	179	5.53	< 0.01

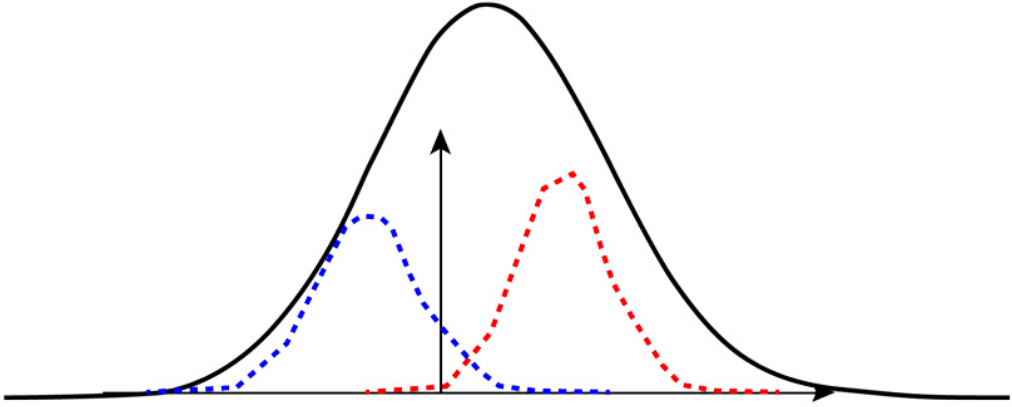
The meaning of the p-value



p: the probability of that we are rejecting H_0 when H_0 is in fact true

α: threshold for *p* to claim statistical significance. Typical values of α: 0.05, 0.01, 0.001.

The meaning of the p-value



p: the probability of that we are rejecting H_0 when H_0 is in fact true

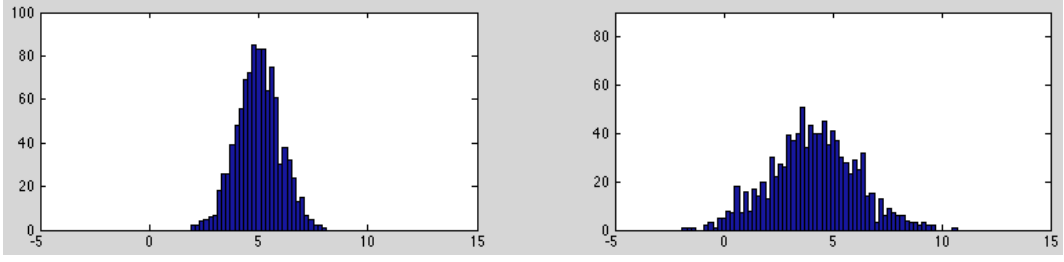
α: threshold for *p* to claim statistical significance. Typical values of α: 0.05, 0.01, 0.001.

Statistical test, an example: t-test

- The *t*-test was invented by W. Gosset. He worked at the Guinness brewery and needed a way to tell the difference between brew batches.
- Gosset was prevented by his employer from publishing under his own name, and therefore published this technique under the pseudonym "Student", hence the name *Student's t-test*.
- We will do a step-by-step t-test in Matlab.

Statistical test, an example: t-test

- Before we start, we need some data. For now, we let Matlab generate this for us:
  - `A = randn(1000,1) * 1 + 5;`
  - `B = randn(1000,1) * 2 + 4;`
- A contains 1000 data points with a mean = 5, and SD = 1
- B contains 1000 data points with a mean = 4, and SD = 2
- Histograms of the two data sets look like this:



Statistical test, an example: t-test

- First step: defining a *null-hypothesis*, usually stating the opposite of what your real hypothesis is.
  - I believe that the two data sets are different, so my null hypothesis is that the two data sets are *not* different.
- Luckily, this is exactly what the Matlab function *ttest2* is made for.
- FIRST TASK: Let's pretend that we have the full set of data available, all 1000 values in A and all 1000 in B.
  - `[h, p] = ttest2( A , B )`
  - result:
    - $h = 1$, means that the null hypothesis is discarded
    - $p = 8.3356 * 10^{-35}$, is an extremely low number, saying that the probability of these two data sets stemming from the same original distribution is extremely low.

However, it is not often we have as many as 1000 data points from each set...

Statistical test, an example: t-test

- NEXT TASK: What happens when we compare 5 data points from set A with 5 data points from set b?
  - Since the numbers in A and B are randomly generated, we can simply select the first 5 from each of them: A(1:5) and B(1:5)
    - A(1:5) contain the values 4.95 4.22 6.20 4.15 6.43
    - B(1:5) contain the values 3.33 10.65 3.44 3.09 3.86
  - We run the two subsets through the *ttest2* function:
  - `[h, p] = ttest2( A(1:5) , B(1:5) )`
  - result:
    - $h = 0$, means that the null hypothesis is not discarded — the two data sets may very well stem from the same underlying distribution.
    - $p = 0.842$

Statistical test, an example: t-test

- A more concrete example: drummers and pianists
  - For instance, let us say that you have asked 50 pianists and 50 drummers to transcribe a melody, they get more points for using less time and having less errors.
  - Further, let us say that the scores for the 50 pianists are in A(1:50), and the drummers are in B(1:50).
  - We run the two subsets through the *ttest2* function:
  - `[h, p] = ttest2( A(1:50) , B(1:50) )`
  - result:
    - $h = 1$, means that the null hypothesis is discarded — it is unlikely that the two data sets stem from the same underlying distribution.
    - $p = 0.032$, so we could claim that our result is statistically significant, with a significance level $\alpha = 0.05$.

Statistical test, an example: t-test

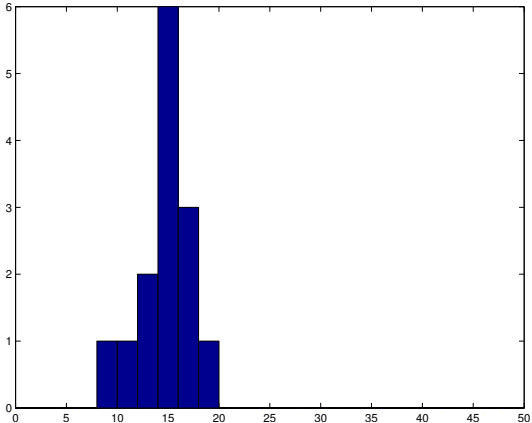
Command Window

```
>> A = randn(1000,1)+5;
>> B = randn(1000,1)*2+4;
>> hist(A)
>> hist(B)
>> boxplot(A)
>> boxplot(B)
>> boxplot([A,B])
>> [h,p]=ttest2(A,B)

h =

     1


p =

   8.3356e-35

>> [h,p]=ttest2(A(1:50),B(1:50))

h =

     1


p =
```

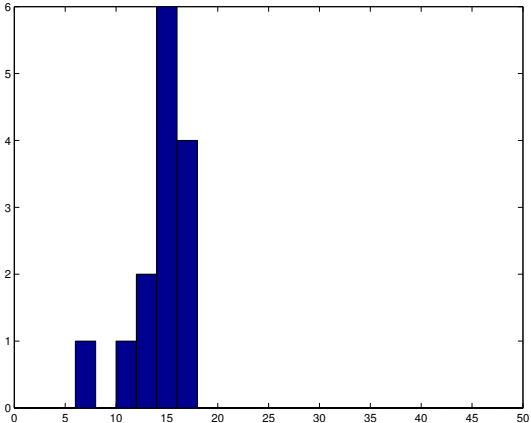
- Here's the matlab command window for what was shown on the previous slides
- Note that if you repeat this, you will get slightly different results because the randn functions generate random numbers.
- There are also ttest functions in Excel and OpenOffice which are slightly more cumbersome to use than the one in Matlab.

A note on skewed distributions

The distribution from last year's students hair length seems to be normally distributed:



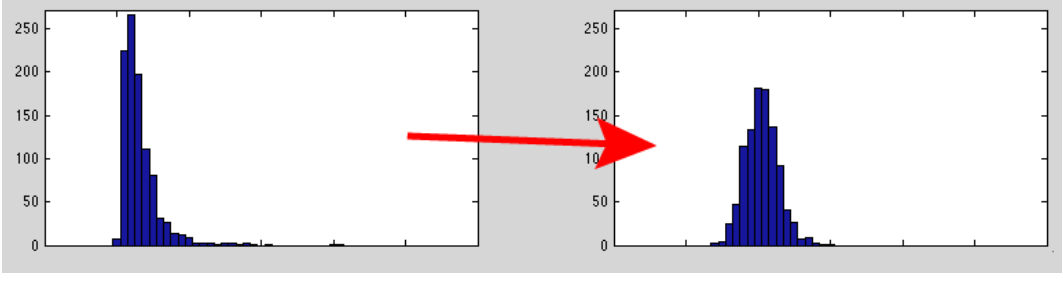
But not all data is. This is an example of data that is slightly skewed:



Most statistical tests assume that the data is normally distributed!

What to do about skewed distributions

- Statistical inferrences can still be made within a single data set, as long as the sample size is large enough (Moore and McCabe 2006).
- Skewed distributions can sometimes be made normal by applying some mathematical function, for instance log-transformation:



Use and misuse of statistics

There is an important difference between *correlation* and *causation*.

- Example: risk of coronary heart disease.
- A number of studies showed that that women who were following a program called combined hormone replacement therapy (HRT) had a lower-than-average incidence of coronary heart disease (CHD).
- This lead doctors to assume that HRT was protective against CHD.
- Later tests showed quite the opposite: That HRT caused a small but statistically significant increase in risk of CHD.
- Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups, with better-than-average diet and exercise regimens.
- The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause.

  Source: `http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation`

Statistics in your master's thesis?

- Research quesiton
- Independent and dependent variables
- Collecting data: Interview / Motion capture / nettskjema.uio.no
  `https://nettskjema.uio.no/answer/71184.html`
- Tools: Excel / Matlab / by hand?