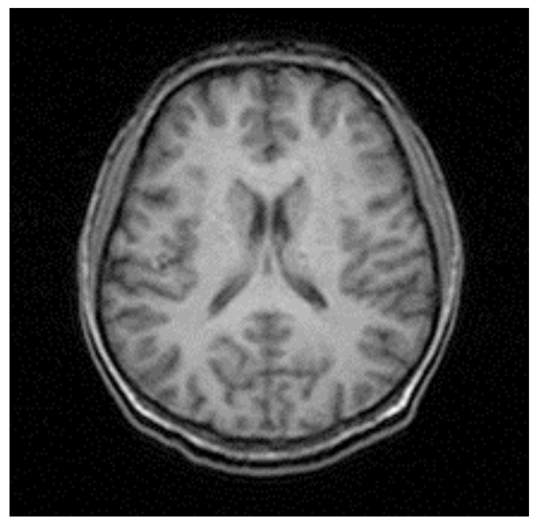
Applications of probability & statistics

- Cell phone transmission over a noisy medium has a probability of error. Number of extra bits sent depend on statistics of the noise.
- Weather prediction based on past observations makes use of probabilities.
- Speech recognition is based on determining the most likely (highest probability) spoken words based on statistics of past speech. The statistics may be speaker specific.
- Image compression standards like *jpeg* make use of unequal probabilities of pixel intensities.

Example: Testing the effect of a new drug

- Problem: A neurologist wants to determine if a certain drug slows down the progress of Alzheimer’s disease.
- It is known that Alzheimer’s disease results in abnormal enlargement of the ventricles (a compartment of the brain) as it progresses.

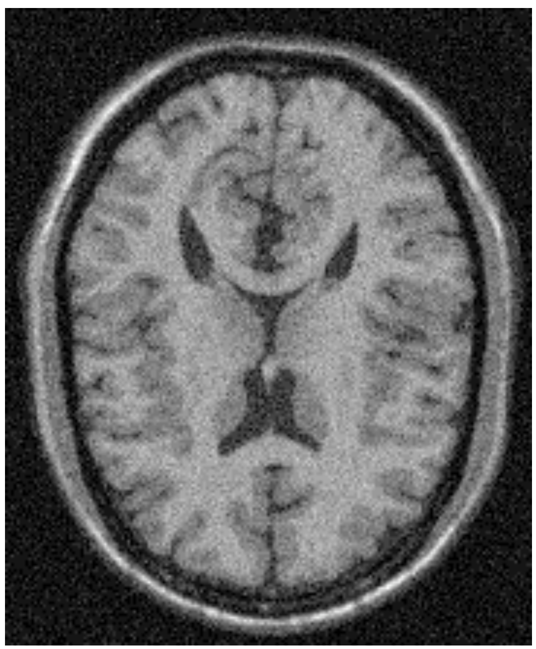
Alzheimer’s patient



Normal brain

Measuring ventricular volume

- The neurologist works together with an engineer who specializes in digital image processing to develop a computer program that automatically measures ventricular volume from Magnetic Resonance Images (MRI)



MRI

Partitioned image

The clinical trial

- The neurologist then recruits 20 Alzheimer's patients into a clinical trial.
- He randomly assigns the 20 patients into 2 groups of 10. One group will take the drug that is being tested for 6 months while the other group will take the placebo (no drug) for 6 months.
- At the beginning of the 6 months all patients have a MRI taken and their ventricular volume is measured.
- This is repeated at the end of the 6 months which allows us to compute the change in ventricular volume for each patient.

The data

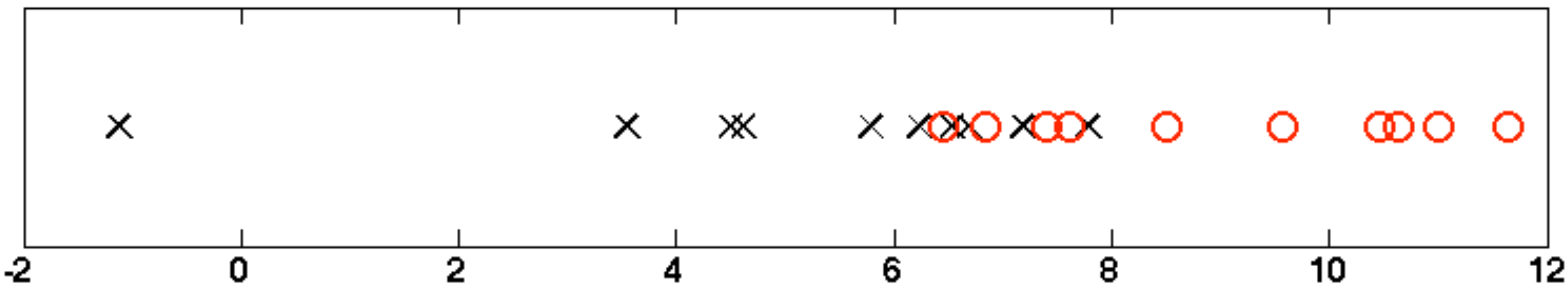
- The following ventricular volume changes (in units of ml) are recorded

Drug	4.5	3.5	7.8	-1.1	5.8	7.2	6.7	6.2	4.6	6.5
Placebo	10.5	9.6	7.4	7.6	10.6	6.4	11.6	11.0	6.8	8.5

- Do you think the drug was effective in slowing down Alzheimer's disease?
- How can we start to look at this data…
  - in a more visual way?
  - in a more quantitative way?

Looking at the data visually

- Lets make a plot of the data:



  - Crosses correspond to “drug” sample
  - Circles correspond to “placebo” sample
- Does this help in forming an opinion?
- Does it directly allow us to make a formal decision on whether the drug is effective?

Towards more quantitative analysis

- If you had to summarize both samples with a single number what would you choose?
  - Answer: **Sample mean**
- Lets take the “drug” sample:

Drug	4.5	3.5	7.8	-1.1	5.8	7.2	6.7	6.2	4.6	6.5

  - What is the sample mean? **5.17**
- How about the “placebo” sample:

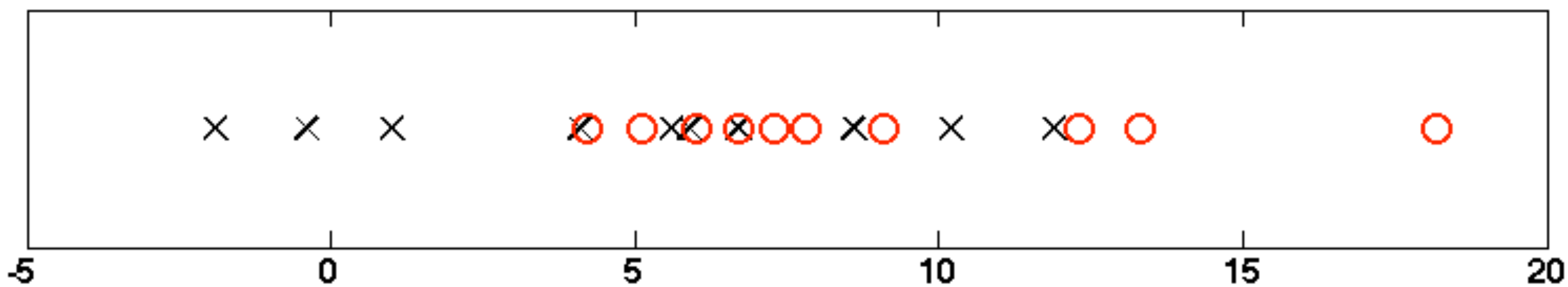
Placebo	10.5	9.6	7.4	7.6	10.6	6.4	11.6	11.0	6.8	8.5

  - What is the sample mean? **9.0**

Lets change our data a bit…

- Assume the following data was recorded instead:

Drug	8.6	5.6	-0.4	6.7	5.9	11.9	10.2	1.0	-1.9	4.1
Placebo	7.3	6.7	12.3	18.2	6.0	4.2	7.8	13.3	9.1	5.1



- Do you think the drug was effective in slowing down Alzheimer's disease in this case?

Compute the sample means

- "Drug" sample:

Drug	8.6	5.6	-0.4	6.7	5.9	11.9	10.2	1.0	-1.9	4.1

  – Sample mean = **5.17**

- "Placebo" sample

Placebo	7.3	6.7	12.3	18.2	6.0	4.2	7.8	13.3	9.1	5.1

  – Sample mean = **9.0**

- <u>Observation</u>: These are the exact same sample means as before!
- <u>Conclusion</u>: Sample mean by itself is not sufficient to describe the data.
- <u>Question</u>: So what changed?

Measuring variability

- Sample variance is a measure of variability.
- Sample variances in our examples:

Units are in ml^2

	Drug	***Placebo***
Dataset 1	6.61	3.59
Dataset 2	20.55	18.99

- Sample standard deviations:

Units are in ml

	Drug	***Placebo***
Dataset 1	2.57	1.89
Dataset 2	4.53	4.36

There is greater variability in the second dataset.

Large variability can hide the difference between two samples.

So even though the sample means remained the same, we are less sure about what to conclude about dataset 2.

Descriptive and inferential statistics

- *Mean*, *variance* and *standard deviation* are what we call descriptive statistics. There are many more such as *median*, *range*, etc.
  - Descriptive statistics provide a summary of our data.
- When we make higher-level decisions about our data this is called statistics inference.
  - For instance: We decide that the drug under testing is not effective in slowing down Alzheimer’s disease.
  - Descriptive statistics become important tools in making statistical inferences.

Probability and Statistics

- For a statistical problem, the sample along with inferential statistics allows us to draw conclusions about the population using elements of probability.
- Problems in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population based on known features of the population.

Some other things…

- Observational study vs. experimental design
- Discrete vs. continuous data
- Random sampling, sample size
- Biased vs. unbiased sample