Descriptive Statistics

Statistics provide us with a standardised way of analysing and measuring a variable of interest based on observations obtained from a sample of the population.

But that all sounds very abstract!

Example 0.1. Chemotherapy is often used to treat Hodgkin's lymphoma. For stage 1 and 2 more than 90% will survive 5 years or more after diagnosis. For stages 3 and 4 between 75% and 90% will survive for 5 years or more after diagnosis (Figures obtained from cancerresearchuk.org).

If a new Chemotherapy drug is developed …

- how do we measure its effectiveness against controls?
- how do we know if our new drug is more/less effective than other drugs?

Statistics!

Definition 0.2. Below a few terms you need to be familiar with are defined.

- **Random variable:** Usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon.
- **Mean:** This is the "average" of the data. It's obtained by adding up all the values in your data and dividing by the total number of values in your data.
- **Median:** This is the "middle" value of your data. It's obtained by arranging your values in ascending order then picking the value in the middle of the list.
- **Mode:** The value which occurs most frequently in your data.

Example 0.3. The following table shows how long (in years) ten patients survived after diagnosis with chemotherapy treatment

Patient no.	1	2	3	4	5	6	7	8	9	10
Years after diagnosis	7	8	5	10	1	6	11	7	12	13

In this example:

- the random variable we're studying is the number of years of survival after diagnosis.
- Our population is all patients who have been diagnosed with Hodgkin's lymphoma but it is not practical to measure everyone, so we've taken the sample of the ten patients in the table.

The **mean** survival time is

$$\bar{x} = \frac{7+8+5+10+1+6+11+7+12+13}{10} = \frac{80}{10} = 8.$$

To get the **meadian** first we reorder the list of measurements, so:

$$\{7,8,5,10,1,6,11,7,12,13\}$$
$$\downarrow$$
$$\{1,5,6,7,7,8,10,11,12,13\}.$$

The "middle" entry of this list is the 5.5^{th} entry, so halfway between the 5^{th} and 6^{th} entry.

$$\tilde{x} = \frac{7+8}{2} = \frac{15}{2} = 7.5$$

Finally the **mode** is 7 since this occurs most often in our list of measurements.

Exercise 0.4. Seven patients prescribed a particular drug for insomnia. The number of side effects experienced by each patient is listed in the table below

Patient no.	1	2	3	4	5	6	7
Number of side effects	0	0	1	2	0	1	4

Calculate the mean, median and mode number of side effects experienced by patients.

Standard deviation, denoted by σ (sigma), provides us with a way of measuring how dispersed the data is.

- A low standard deviation would mean that the data points are very close to the mean.
- Whereas a high standard deviation would mean the data points are very dispersed and not very close to the mean.

Definition 0.5. Let $\{x_1, x_2, \ldots, x_n\}$ be a set of random variables with mean equal to $\overline{x}$. Then

$$\sigma = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Definition 0.6. The previous definition is for the population and is said to be **biased** if we only have a sample of the population. Here is the unbiased estimator of the standard deviation

$$S = \sqrt{\frac{1}{n-1} \times \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Example 0.7. Let's calculate the standard deviation of our previous example on Hodgkin's lymphoma.

Patient no.	1	2	3	4	5	6	7	8	9	10
Years after diagnosis	7	8	5	10	1	6	11	7	12	13

Recall that the mean was $\overline{x} = 8$.

$$\sigma = \sqrt{\frac{(-1)^2 + 0^2 + (-3)^2 + 2^2 + (-7)^2 + (-2)^2 + 3^2 + (-1)^2 + 4^2 + 5^2}{10}}$$

$$\sigma = \sqrt{\frac{1 + 0 + 9 + 4 + 49 + 4 + 9 + 1 + 16 + 25}{10}}$$

$$= \sqrt{\frac{118}{10}}$$

$$= \sqrt{11.8}$$

$$\approx 3.435$$

And the unbiased estimator of the standard deviation is

$$S = \sqrt{\frac{118}{9}} \approx 3.621$$