Statistics using GeoGebra

Statistics helps you understand the past, make sense of the present and infer about the future.

Let’s consider the ages of a group of people **I**.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This is the number of people who are of a particular age or frequency:

54	55	56	57	58	60
3	1	1	2	2	2

There are more people who are 54 years old than any other age. Hence, the **mode** of this distribution is 54 years.

But there may be 1 mode or no mode at all. So the **median** (the middle value) may be better to look at such distributions.

The above group **I** is already arranged in increasing order. As there are 11 values, the exact middle would be the middle (6th) value, i.e., 57. If there had been only 10 values in this group, then the median would be the mean of the 5th an 6th values. But the median cannot be used when you have categorical nominal data.

The **mean** or the **arithmetic average** of group **I** is the sum of all the values in group **I** divided by the number of values.

So mean = 623/11 = 56.6 years

Population mean is indicated by **μ (mu)** while **sample mean** is indicated by $\boldsymbol{\overline{X}}$ **(X bar)**.

When the mode, median and mean are in the middle of the distribution, it is said to be **symmetrical**. When the mean is pulled away from the middle where the median lies, such a distribution is said to be **skewed**. If the tail on the right side is longer, it is said to be **positively** or **right skewed**. If the tail on the left side is longer, it is said to be **negatively** or **left skewed**.

Outliers are extreme data values that are very different from the rest of the data. They alter the results of your analysis, especially, the mean. In group **I**, let us replace the last value of 60 with 81. 81 would be very different from the other values and hence, would be an outlier. It would change the mean but the median would still be the same.

Can you draw a box plot for group **I**?

Standard deviation (s.d.) is a measure of dispersement or how spread out your data are around the mean. Population s.d. is indicated by **σ (sigma)** while sample s.d. is indicated by **s**.

If you had to calculate the σ for group **I**,

i) Add up all the numbers: 623

ii) Square 623 and divide by n : $(623)^2/11 = 35284.45$

iii) Square all values in **I** individually and then add them all up: 35335

iv) Subtract ii from iii: 35335 – 35284.45 = 50.54545

v) Subtract 1 from n: 11-1 = 10

vi) Divide iv by v: 50.54545/10 = 5.054545

vii) Take the squareroot of vi: 2.248232. This is your standard deviation. This means that your data is ± 2.248232 from the mean.

The whole point of collecting data is to make sense of it and see if there is any formula that will explain or "model" any relationship between the different data points. This process of coming up with an equation that will explain our data is called "**regression**". To know how "good" any given regression is, *i.e.*, whether the equation explains our data well, we use the **coefficient of determination**, $\mathbf{R^2}$ or $\mathbf{r^2}$. The closer this coefficient is to 1 or -1, the better is our equation in explaining or modelling our data. If the equation is a linear one of the form, **y = mx + b**, where **m** is the slope and **b** is the y-intercept, then this is called **linear regression**. But sometimes, the data won't fit a straight line. You may have to see if **quadratic** or **exponential** or **cubic** or **quartic** works better.

In **simple linear regression**, the **best fit line** that best describes the data is also called the **least squares regression line**. The measure by which the data point vertically misses the regression line or parabola etc is called the **residual** value. A **residual plot** has residual values on the y-axis and the independent variable on the x-axis.