Learning Objectives

Distribution

Skewness

Practise Data Cleaning

DPhi

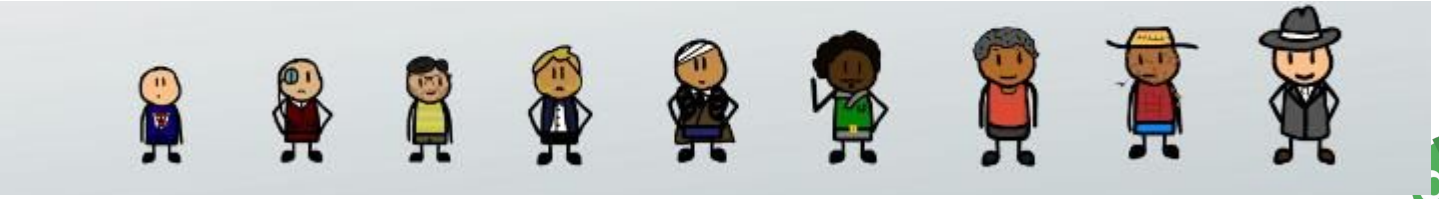
Distribution

- **Probability** - by going with the literal meaning it is the extent to which something is likely to happen or probable. The same concept applies to statistics when we talk about probability distributions.

- A **probability distribution** is a function that describes the likelihood of obtaining the possible values (probability) of a variable.

- **Suppose a teacher** notes down the heights of all students in her class.

- Now, she can draw a probability distribution when she needs to know which outcomes (heights) are most likely, the spread of potential values (range of heights), and the likelihood of different results (probability of each height).

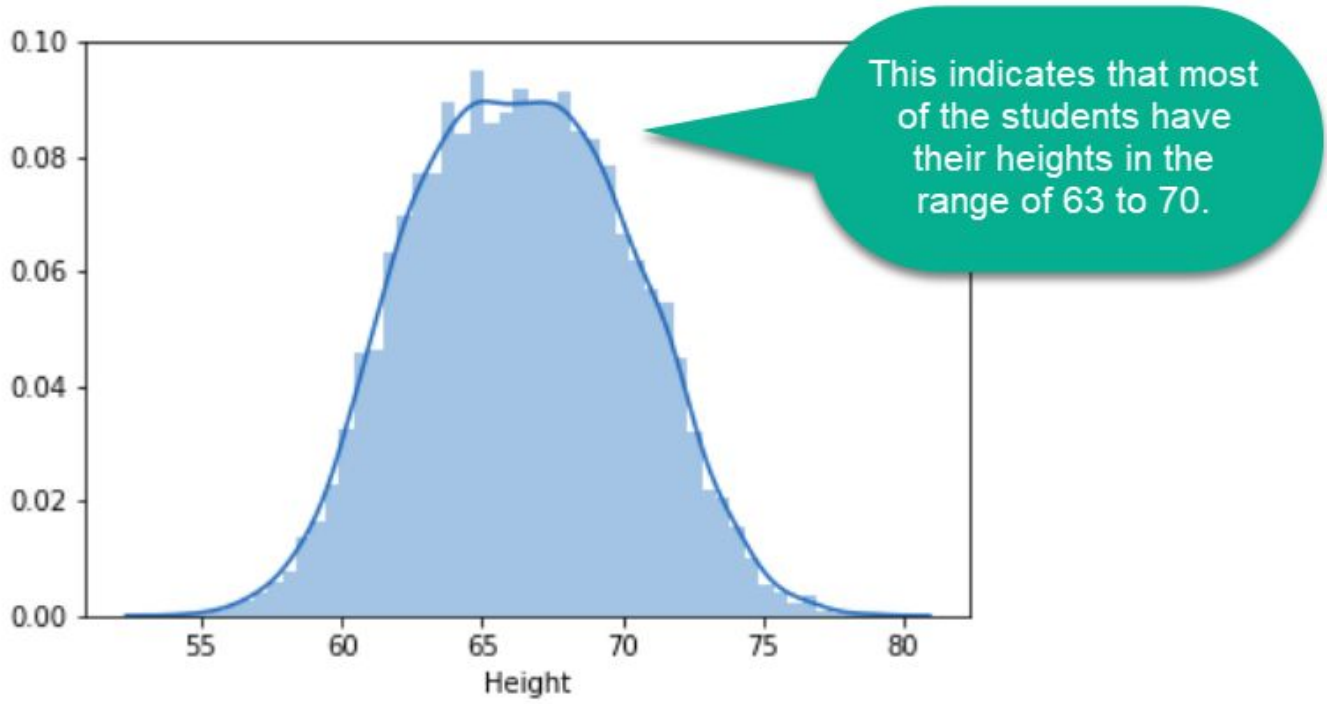


Distribution

- Let's take the heights dataset consisting of student IDs and heights: https://bit.ly/Heights_Data
- The most convenient way to take a quick look at a distribution in seaborn is the distplot() function. On plotting a distplot on this dataset, we'll observe a bell shaped curved like this:

```
sns.distplot(data['Height'])
```

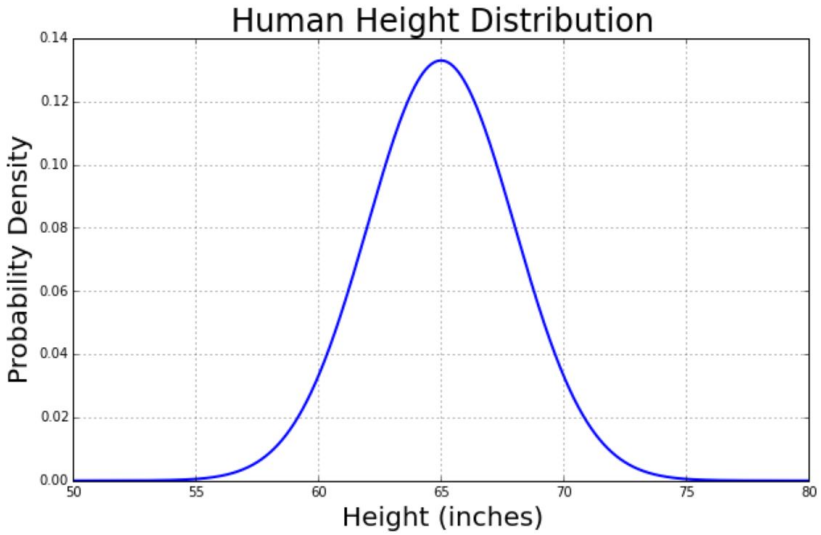
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51ac458438>
```



DPhi

Distribution

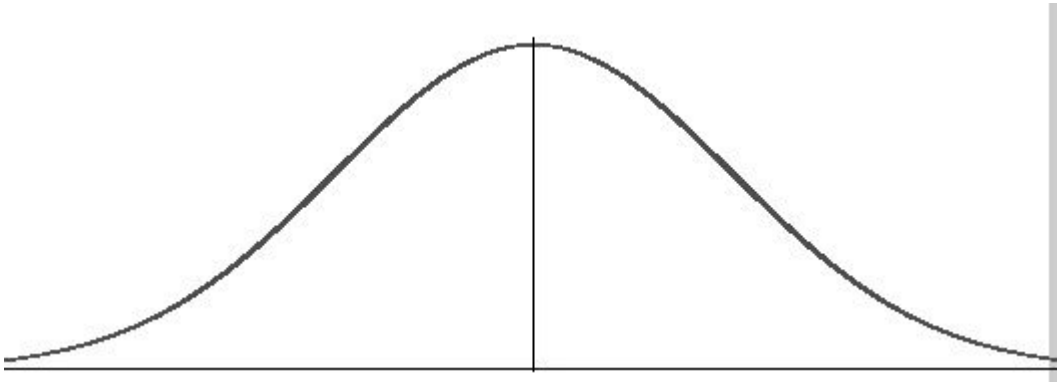
- Probability distributions are usually (but not solely) represented in charts whose abscissa axis (x axis) represents the possible values of the variable and whose ordinal axis (y axis) represents the probability of occurrence (probability density).



- Most statistical models rely on a normal distribution, a distribution that is symmetric and has a characteristic bell shape.

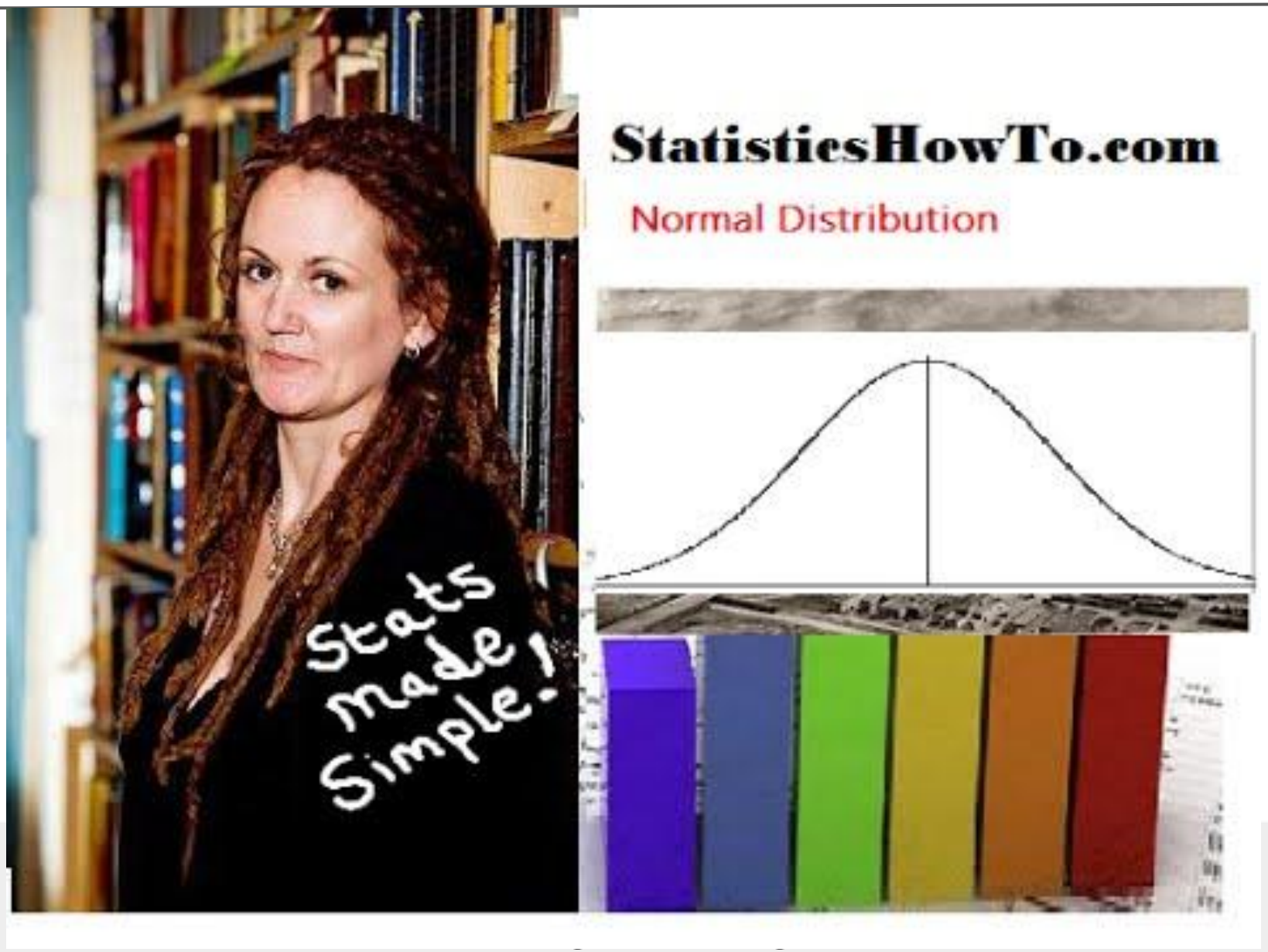
Normal/ Gaussian Distribution

- Also called bell shaped curve.



- It is a distribution of data that occurs naturally in many situations.
- It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

Normal Distribution

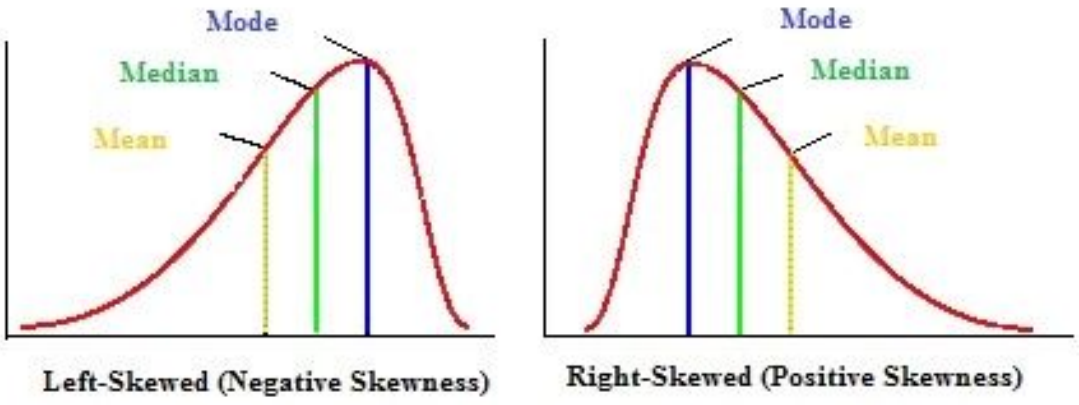


Skewness

- Real life data rarely, if ever, follow a perfect normal distribution.
- The skewness measures the symmetry of a distribution or how different a given distribution is from a normal distribution.
- The normal distribution is symmetric and has a skewness of zero.
- There are two types of Skewness: Positive and Negative

Types of Skewness

- There are two types of Skewness: Positive and Negative



- **<u>Positive/Right Skewness</u>** means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.
- **<u>Negative/Left Skewness</u>** is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.
- **For further reading refer:** https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/

Understanding skewness with an example

- Let us take a very common example of house prices. Suppose we have house values ranging from $100k to $1,000,000 with the average being $500,000.

- If the peak of the distribution was left of the average value, portraying a positive skewness in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k. This could be for many reasons, but we are not going to interpret those reasons here.

- If the peak of the distributed data was right of the average value, that would mean a negative skew. This would mean that the houses were being sold for more than the average value.

Data Cleaning Practice

Time for getting your hands dirty and cleaning a dataset!

Follow the various steps of Data Cleaning in this article to clean the Russian Housing Dataset. It covers everything we've learnt till now. You'll analyse and visualise data, detect outliers, remove irrelevant and inconsistent values and get a structured, clean data at the end.

- **https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d**

DPhi

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.

DPhi