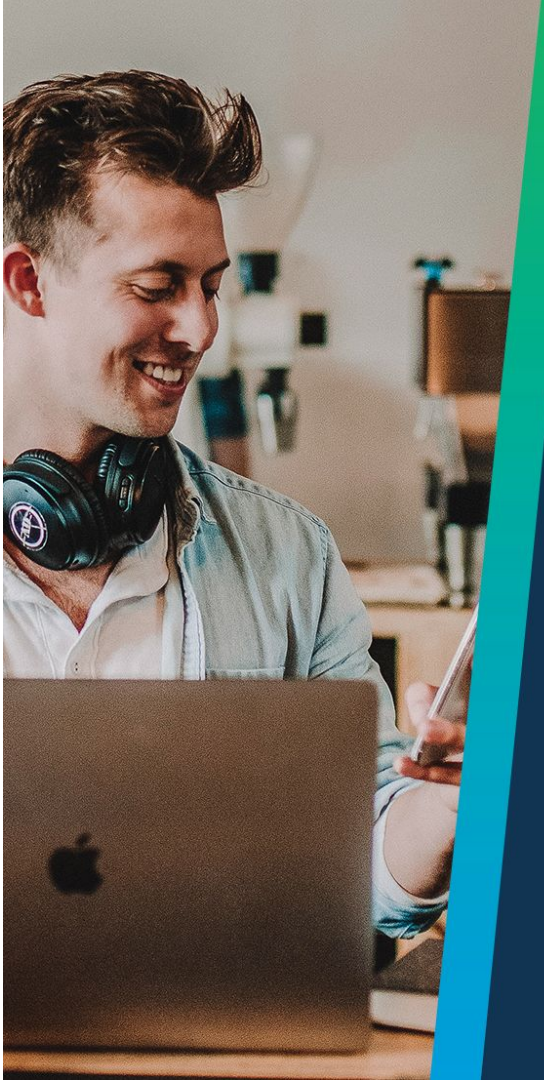


DS Workshop

Hyperiondev

Probability for Data Scientists

Welcome

Your Lecturer for this session

Sanana Mwanawina

Workshop – Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment - please engage accordingly.
- No question is daft or silly - **ask them!**
- There are Q/A sessions midway and at the end of the session, should you wish to ask any follow-up questions.
- You can also submit questions here: www.hyperiondev.com/support
- For all non-academic questions, please submit a query: www.hyperiondev.com/support
- Report a safeguarding incident: hyperiondev.com/safeguardreporting
- We would love your feedback on lectures and workshops: https://hyperionde.wufoo.com/forms/zsgv4m40ui4i0g/

GitHub repo

Go to: github.com/HyperionDevBootcamps

Then click on the "**C4_DS_lecture_examples**" repository, do view or download the code.

Objectives

1. Learn about descriptive statistics
2. Use descriptive statistics to make inferences on data
3. An introduction to random variables and probability distributions

Descriptive Statistics

Before you dive into tackling a situation or problem, it is wise to step back and take a glimpse at the “the big picture”. We can use descriptive statistics to do so. We have two classes of descriptive statistics:

1. Measures of central tendency
2. Measures of spread

Measures of Central Tendency

1. Mean or average: the sum of all values divided by the number of values.

$$M = \frac{\sum_i^n X_i}{n}$$

2. Median: the middle-most value in a sequence of numbers.

$$X_{\frac{N+1}{2}} \text{ if } \mathbf{N} \text{ is odd and } \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2} \text{ if } \mathbf{N} \text{ is even}$$

Measures of Central Tendency

3. Mode: the number that appears the most frequently in the data. Like the Median, this is a good measure for the typical value of something.

   To get the mode, you must count the number of occurrences of each type of category.

Measures of Spread

1. Variance: an absolute measure of how “spread out” the data is.

$$S^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

Let’s say you are managing stock portfolios. An investor wants to find a low-risk stock in the tech industry. You consider two stocks:

Company A: last few stock prices were [128, 146, 112, 153] with mean 134.75
Company B: last few stock prices were [163, 52, 208, 128] with mean 137.75

While company B has a higher mean, we need to consider the variance of these prices since out investor is interested in low-risk stocks.

Measures of Spread

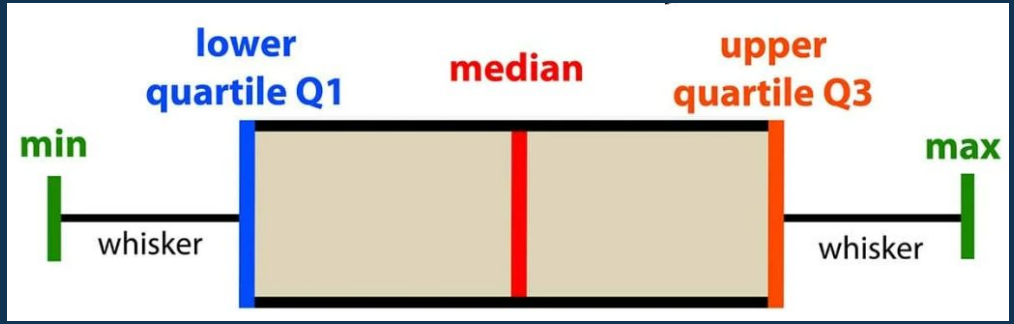
Company A variance = 340.92
Company B variance = 4340.25

The decision is obvious now. Company A has stocks that Tend to vary less from the mean.

2. Standard deviation: the square root of the variance. Commonly used when defining outliers. When data points lie x standard deviations from the mean, they would be classified as an outlier.

Inferences on data

There are many ways we could go about this. We will specifically look at the box-and-whisker plot, or simply the box plot.



Inferences on data

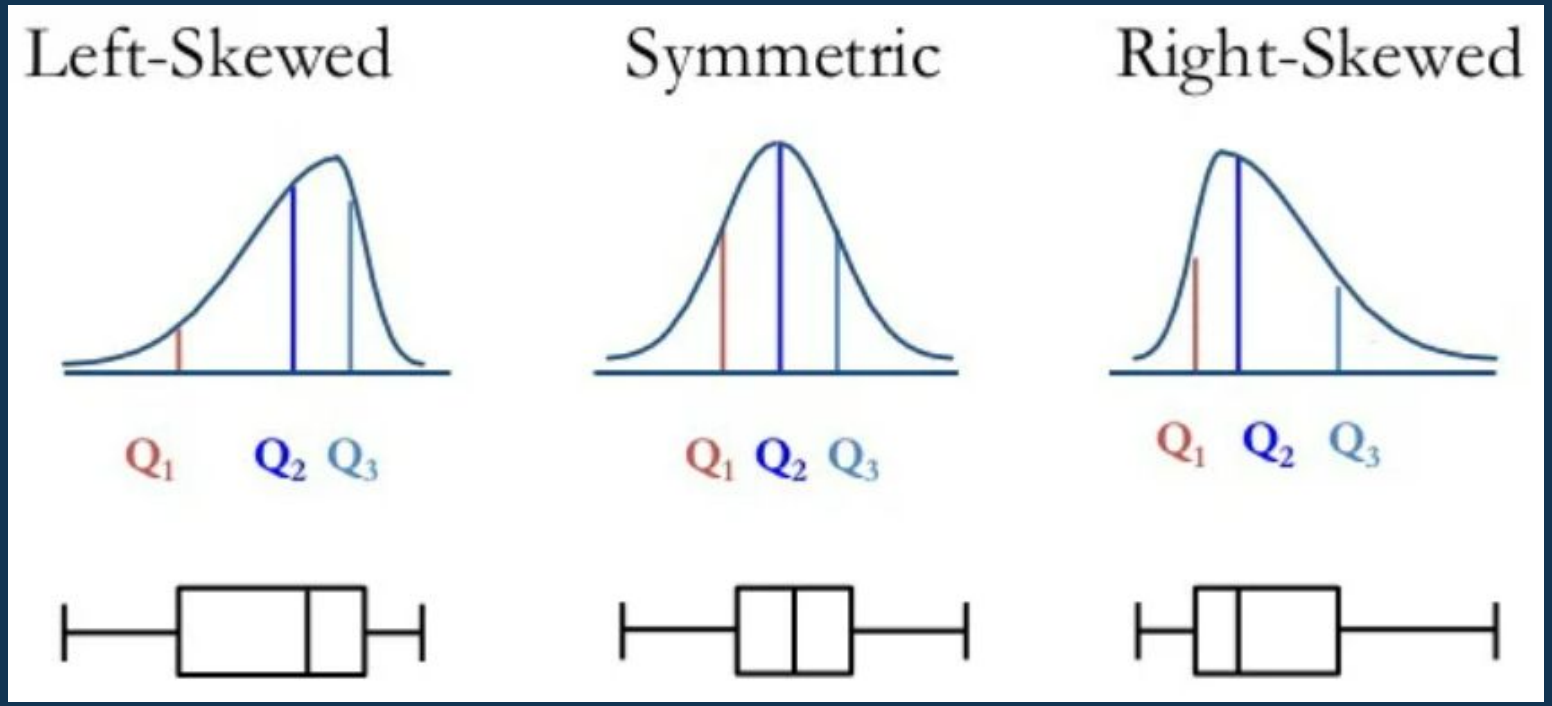
The weights of male and female individuals participating in a medical study are summarized in the box plots below:

We can make a lot of powerful inferences about our data using the above plots.

Inferences on data

1. The median weight of males is about 162 lbs
2. About 50% of male participants have weights between 150 and 185 lbs
3. About 25% of female students have weights more than 130 lbs
4. Male participants have more variability in weights compared to female participants
5. The mean weight of female participants is probably around 120 lbs because of symmetry… symmetry? Let's talk about that.

Inferences on data



Inferences on data

- Left-skewed is also referred to as negatively skewed.
- Right-skewed is also referred to as positively skewed.

Before even getting into our problem or investigation, the descriptive statistics can tell us so much about our data. That is the power of statistics and the reason why it is important to carry out some exploratory data analysis before going head-first into problem solving.

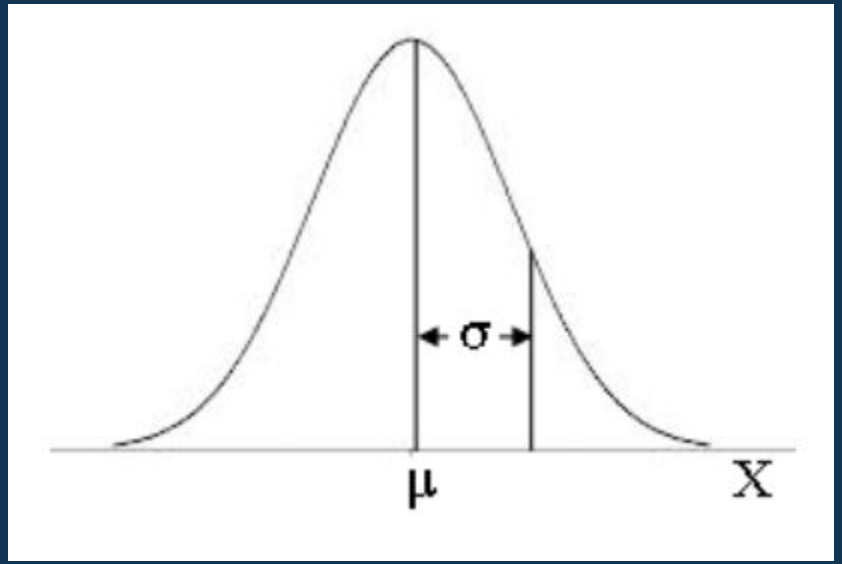
Probability Distributions

Random variables follow probability distributions. You can think of probability distributions as functions that allow us to find the probability that our random variable takes on some value or range of values.

We will use the weights of females to illustrate the use of probability distributions. Given that the weights of females appear to be distributed symmetrically, we can use a Normal distribution to model the weights of females.

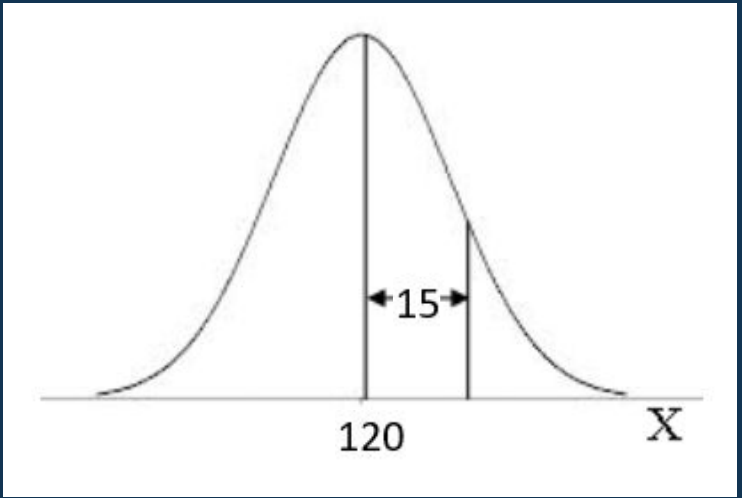
Probability Distributions

If X follows a Normal distribution, we would say $X \sim N(\mu, \sigma^2)$



Probability Distributions

If X represents the weights of females, then $X \sim N(\mu = 120, \sigma^2 = 15^2)$



Important note: the variance σ^2 is a (good) approximation

Probability Distributions

How is this useful?

With this distribution, we can evaluate many probabilities. For example, what is the probability that a female will weigh less than 110 lbs? Written differently, what is Pr(X < 110) ?

We can use Python to evaluate this.

Probability Distributions

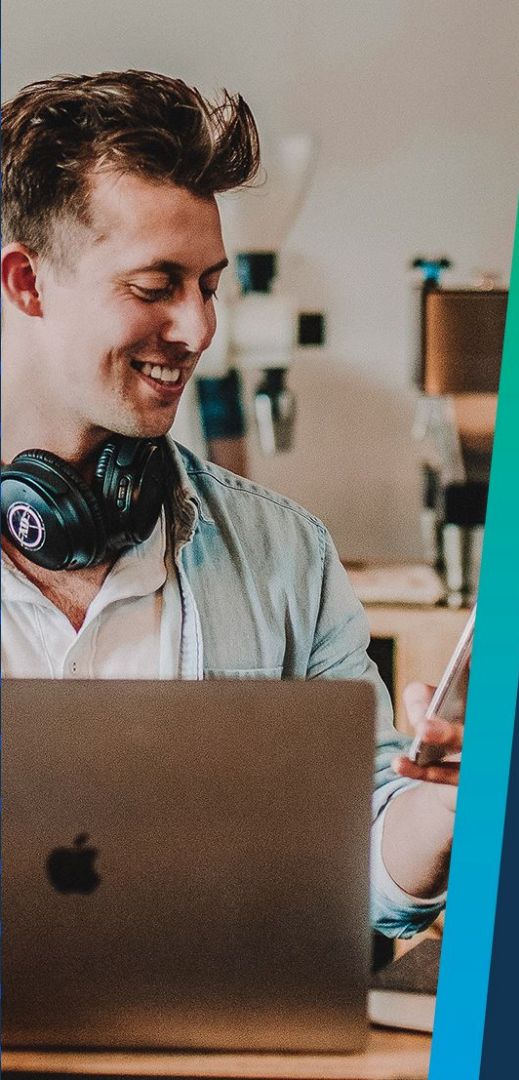
```
1   from scipy.stats import norm
2
3   # give the mean and standard deviation for X
4   mu = 120
5   sigma = 15
6
7   # calculate the probability of X being less than a value
8   value = 110
9   probability = norm.cdf(value, loc = mu, scale = sigma)
```

Which yields Pr(X < 110) = 0.2524925 ≈ 0.25

Hyperiondev

Q & A Section

Please use this time to ask any questions relating to the topic explained, should you have any



Hyperiondev

Thank you for joining us