Week 4

Data Science Laboratory 1

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

Measuring Central Tendency is my value to the left/right of the middle value?

Measuring the Dispersion how far is the value from the middle value?

Multivariate data mean, median

Summary statistics

- Now that we have seen ways to load and visualise data, we next want to understand how to statistically analyse data. This will be key to being able to explore our data.
- For the most part, almost all of the following can be easily calculated in Python either through `NumPy` or `scipy.stats`.

Summary statistics

- Now that we have seen ways to load and visualise data, we next want to understand how to statistically analyse data. This will be key to being able to explore our data.
- For the most part, almost all of the following can be easily calculated in Python either through `NumPy` or `scipy.stats`.
- In the interests of keeping things brief, we will focus on two main areas of basic statistical descriptions: measures of central tendency, and dispersion of the data.
- Roughly speaking, the former measures the location of the middle/centre of our data. Intuitively, we are asking where most of an attribute's values fall. Related to this will be the mean, median, mode and midrange.
- The latter assesses the central tendency of our data set. In other words, we want to know how spread out our data is. Here, we are interested in variance, standard deviation, range, quartiles and interquartile ranges.

Measuring central tendency (Mean)

- Suppose we have some attribute of our data X (could be salary, height, weight etc.) which has been recorded for a set of objects.
- We let $\{x_1, x_2, \ldots, x_N\}$ be the set of N observations for X (e.g. we have weighed a person N times over a period of weeks).
- If we were to plot the observations for salary/height/weight, where would most of the values fall? This gives us an idea of the central tendency of the data.

Measuring central tendency (Mean)

- Suppose we have some attribute of our data X (could be salary, height, weight etc.) which has been recorded for a set of objects.
- We let $\{x_1, x_2, \ldots, x_N\}$ be the set of N observations for X (e.g. we have weighed a person N times over a period of weeks).
- If we were to plot the observations for salary/height/weight, where would most of the values fall? This gives us an idea of the central tendency of the data.
- The most common numeric measure of the 'centre' of a set of data is the (arithmetic) mean. For the data set above, the mean of this set of values is given by,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Example

- Suppose the following values (in thousands) are the salaries of 11 individuals, shown in increasing order: 17, 20, 23, 31, 40, 55, 60, 62, 64, 84, 99.

Example

- Suppose the following values (in thousands) are the salaries of 11 individuals, shown in increasing order: 17, 20, 23, 31, 40, 55, 60, 62, 64, 84, 99.

  Then, $\bar{x} = \frac{1}{11}(17 + 20 + \cdots + 99) = 50.45k$.

Weighted mean

- Sometimes, each value x_i in a set may be associated with a weight w_i for $i \in \{1, \ldots, N\}$.
- The weight reflects the importance, or occurrence frequency attached to their respective values.

Weighted mean

- Sometimes, each value x_i in a set may be associated with a weight w_i for $i \in \{1, \ldots, N\}$.
- The weight reflects the importance, or occurrence frequency attached to their respective values.
- In this case, we can compute the *weighted arithmetic mean/weighted average* as follows:

$$\bar{x}_W = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}.$$

Benefits and problems of the mean

- The main benefit of the mean (and weighted mean) is that is is easy to compute and gives a useful insight into the data set.

[1] See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2625386/pdf/577.pdf.

Benefits and problems of the mean

- The main benefit of the mean (and weighted mean) is that is is easy to compute and gives a useful insight into the data set.
- However, it does have a major problem since it is sensitive to extremes (e.g. outliers).
- For example, if in the above example we included one more individual, whose salary was £20,800,000 (looking at you, Kevin De Bruyne) then the average will be massively skewed, even though it is just one person.

[1] See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2625386/pdf/577.pdf.

Benefits and problems of the mean

- The main benefit of the mean (and weighted mean) is that is is easy to compute and gives a useful insight into the data set.
- However, it does have a major problem since it is sensitive to extremes (e.g. outliers).
- For example, if in the above example we included one more individual, whose salary was £20,800,000 (looking at you, Kevin De Bruyne) then the average will be massively skewed, even though it is just one person.
- A classic case of this is when discussing life expectancy of earlier periods in history. For example, in Victorian times the average age for a baby boy was just 40. However, this is in large part a consequence of the high infant mortality rate in such times. Once this is stripped out, life expectancy at 5 years was 75 for men, remarkably similar to today. Indeed, life expectancy for mature men has not changed dramatically for over 3000 years (the same is not true, however, for women).[1]

[1] See `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2625386/pdf/577.pdf`.

Dealing with extremes

- To offset the effect caused by a small number of extreme values, we can instead use the *trimmed mean*, which is the mean obtained after removing a few values at the high and low extremes.
- In this case, a percentage p between 0 and 100 is specified, the top and bottom $(\frac{p}{2})\%$ of the data is thrown out, and the mean is then calculated in the normal way.
- In this way, we can think of it as a generalisation of the mean where the standard mean corresponds to $p = 0\%$.

Dealing with extremes

- To offset the effect caused by a small number of extreme values, we can instead use the *trimmed mean*, which is the mean obtained after removing a few values at the high and low extremes.
- In this case, a percentage p between 0 and 100 is specified, the top and bottom $(\frac{p}{2})\%$ of the data is thrown out, and the mean is then calculated in the normal way.
- In this way, we can think of it as a generalisation of the mean where the standard mean corresponds to $p = 0\%$.
- e.g. Consider the values $\{1, 2, 3, 4, 100\}$. Then the mean of these values is $\frac{110}{6} \approx 18.3$.
- However, if we take the trimmed mean with $p = 40\%$ then the trimmed mean is 3.

Dealing with extremes

- To offset the effect caused by a small number of extreme values, we can instead use the *trimmed mean*, which is the mean obtained after removing a few values at the high and low extremes.
- In this case, a percentage p between 0 and 100 is specified, the top and bottom $(\frac{p}{2})\%$ of the data is thrown out, and the mean is then calculated in the normal way.
- In this way, we can think of it as a generalisation of the mean where the standard mean corresponds to $p = 0\%$.
- e.g. Consider the values $\{1, 2, 3, 4, 100\}$. Then the mean of these values is $\frac{110}{6} \approx 18.3$.
- However, if we take the trimmed mean with $p = 40\%$ then the trimmed mean is 3.
- Of course, the key is to trim off the right amount. Too much and we may lose valuable information. Too little and we aren't removing the problem.

Measuring central tendency (Median)

- For skewed (asymmetric) data, a better measure of the centre of data is the *median*, which is the middle value in a set of ordered data values.
- It is the value that separates the higher half of a data set from the lower half.

[2]Popular sorting algorithms: merge sort, quicksort, bubble sort, insertion sort.

Measuring central tendency (Median)

- For skewed (asymmetric) data, a better measure of the centre of data is the *median*, which is the middle value in a set of ordered data values.
- It is the value that separates the higher half of a data set from the lower half.
- To compute, let $X = \{x_1, \ldots, x_N\}$ be sorted in non-decreasing order[2] so that $x_1 = min(X)$ and $x_N = max(X)$.
- Then the median is defined as follows:

$$median(X) = \begin{cases} x_{r+1}, & \text{if } m \text{ is odd, i.e. } m = 2r+1; \\ \frac{1}{2}(x_r + x_{r+1}), & \text{if } m \text{ is even, i.e.} m = 2r. \end{cases}$$

[2] Popular sorting algorithms: merge sort, quicksort, bubble sort, insertion sort.

Example

- Go back to $\{1, 2, 3, 4, 100\}$.
- As it is ordered, the first step is to remove 1 and 100. Then we remove 2 and 4. We are therefore left with 3 and so that is the median.

Example

- Go back to $\{1, 2, 3, 4, 100\}$.
- As it is ordered, the first step is to remove 1 and 100. Then we remove 2 and 4. We are therefore left with 3 and so that is the median.
- Suppose we add an element so our set becomes $\{1, 2, 3, 4, 5, 100\}$.
- We do the same thing but are now left with $\{3, 4\}$. Take the mean of these and so the median in this case is 3.5.

Example

- Go back to $\{1, 2, 3, 4, 100\}$.
- As it is ordered, the first step is to remove 1 and 100. Then we remove 2 and 4. We are therefore left with 3 and so that is the median.
- Suppose we add an element so our set becomes $\{1, 2, 3, 4, 5, 100\}$.
- We do the same thing but are now left with $\{3, 4\}$. Take the mean of these and so the median in this case is 3.5.
- Observe that the median is just the trimmed mean with $p = 100\%$.

Benefits and approximations

- While the median benefits from being easy to compute, it is expensive to compute when we have a large number of observations.
- Nevertheless, for numeric attributes (i.e. attributes like salary or weight which we can represent with numbers, as opposed to attributes like sex which we generally represent with strings (M or F). we can easily *approximate* the value.

Benefits and approximations

- While the median benefits from being easy to compute, it is expensive to compute when we have a large number of observations.
- Nevertheless, for numeric attributes (i.e. attributes like salary or weight which we can represent with numbers, as opposed to attributes like sex which we generally represent with strings (M or F). we can easily *approximate* the value.
- Assume that the data are grouped in intervals according to their x_i data values and that the frequency (i.e. the number of data values) of each interval is known.
- For example, we can group employee's annual salaries into the intervals £10,000-£20,000, £20,000-£30,000, and so on.

Approximation formula

- Let the interval that contains the median frequency be the *median interval*. We can then approximate the median of the entire data set (e.g. the median salary) by interpolation using the formula,

$$median \approx L_1 + \left(\frac{\frac{N}{2} - (\sum freq)_l}{freq_{median}} \right) width,$$

where

- L_1 is the lower boundary of the median interval;
- N is the number of values in the entire data set;
- $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval;
- $freq_{median}$ is the frequency of the median interval;
- $width$ is the width of the median interval.

Measuring central tendency (mode)

- For a set of data, the mode is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative (non-numerical) and quantitative (numerical) attributes.

Measuring central tendency (mode)

- For a set of data, the mode is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative (non-numerical) and quantitative (numerical) attributes.
- It is possible for the greatest frequency to correspond to several different values which results in more than one mode.
- Data sets with one/two/three modes are respectively called *unimodal/bimodal/trimodal*.
- In general, a data set with two or more modes is *multimodal*.
- At the other extreme, if each data value occurs only once, then there is no mode.

Measuring central tendency (mode)

- For a set of data, the mode is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative (non-numerical) and quantitative (numerical) attributes.
- It is possible for the greatest frequency to correspond to several different values which results in more than one mode.
- Data sets with one/two/three modes are respectively called *unimodal/bimodal/trimodal*.
- In general, a data set with two or more modes is *multimodal*.
- At the other extreme, if each data value occurs only once, then there is no mode.
- e.g. The mode of $X = \{1, 1, 2, 3, 4, 4, 4, 5, 6, 7\}$ is 4 (unimodal) and the mode of $Y = \{1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7\}$ is 3 and 4 (bimodal).

Measuring central tendency (mode)

- For a set of data, the mode is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative (non-numerical) and quantitative (numerical) attributes.
- It is possible for the greatest frequency to correspond to several different values which results in more than one mode.
- Data sets with one/two/three modes are respectively called *unimodal/bimodal/trimodal*.
- In general, a data set with two or more modes is *multimodal*.
- At the other extreme, if each data value occurs only once, then there is no mode.
- e.g. The mode of $X = \{1, 1, 2, 3, 4, 4, 4, 5, 6, 7\}$ is 4 (unimodal) and the mode of $Y = \{1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7\}$ is 3 and 4 (bimodal).
- For unimodal numeric data that are moderately skewed (asymmetrical), we have the following approximation:

mean - mode ~~ 3 x (mean - median)

$$mean - mode \simeq 3 \times (mean - median).$$

Measuring central tendency (Midrange)

- The *midrange* is the average of the largest and smallest values in the set.[3]

[3] If using SQL, use the aggregate functions `max()` and `min()`.

Measuring central tendency (Midrange)

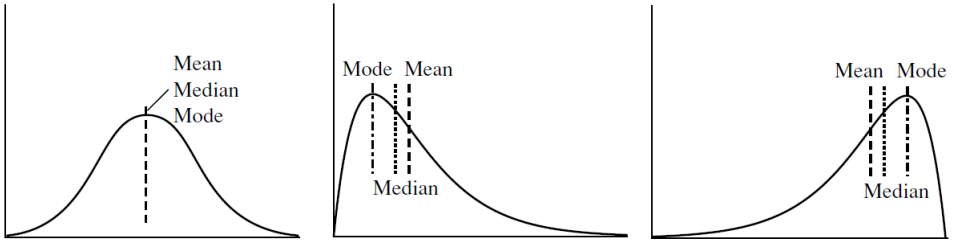
- The *midrange* is the average of the largest and smallest values in the set.[3]
- e.g. Return to the salary example from earlier, with values (in thousands) given by 17, 20, 23, 31, 40, 55, 60, 62, 64, 84, 99. Then the midrange is $\frac{17000+99000}{2} = 58000$.

[3] If using SQL, use the aggregate functions `max()` and `min()`.

Comparing measurements of central tendency

- We can conclude by noting that in a unimodal frequency curve with perfect symmetric data distribution, the mean, median and mode are all the same center value, as shown on the next slide. (a on the next slide)
- Of course, data in most real applications are not symmetric.
- Instead, they may either be *positively skewed*, where the mode occurs at a value that is smaller than the median (b on the next slide),
- or *negatively skewed*, where the mode occurs at a value greater than the median (c on next slide).

Comparing measurements of central tendency



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Multivariate data

- Importantly, the above can be generalised for measures of location for data that consists of several attributes (multivariable/multivariate data), $\mathbf{x} = (x_1, \ldots, x_N)$.
- For computing the mean and median, the we just do the obvious; namely, we compute the mean and median for each attribute separately.
- So, if $\bar{x}_i$ is the mean of the i^{th} attribute, then

$$\bar{\mathbf{x}} = (\bar{x_1}, \ldots, \bar{x_N}).$$

Measuring the dispersion (Interquartile range)

- Again, let $x_1, \ldots, x_N$ be a set of observations for some numeric attribute X.
- The *range* of the set is the difference between the largest and the smallest values. If these are ordered so that $x_1 = min$ and $x_N = max$, then this is simply $x_N - x_1$.

Measuring the dispersion (Interquartile range)

- Again, let $x_1, \ldots, x_N$ be a set of observations for some numeric attribute X.
- The *range* of the set is the difference between the largest and the smallest values. If these are ordered so that $x_1 = min$ and $x_N = max$, then this is simply $x_N - x_1$.
- Keep this assumption that our data is ordered. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets.
- For example, if our set is $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ then we might split this set into $\{x_1, x_2\} \cup \{x_3, x_4\} \cup \{x_5, x_6\}$.
- These data points are called *quantiles* and they are points taken at regular intervals of a data distribution, dividing it into essentially equal size sets.

Measuring the dispersion (Interquartile range)

- Again, let $x_1, \ldots, x_N$ be a set of observations for some numeric attribute X.
- The *range* of the set is the difference between the largest and the smallest values. If these are ordered so that $x_1 = min$ and $x_N = max$, then this is simply $x_N - x_1$.
- Keep this assumption that our data is ordered. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets.
- For example, if our set is $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ then we might split this set into $\{x_1, x_2\} \cup \{x_3, x_4\} \cup \{x_5, x_6\}$.
- These data points are called *quantiles* and they are points taken at regular intervals of a data distribution, dividing it into essentially equal size sets.
- The k^{th} *q-quantile* for a given data distribution is said to be the value x such that, at most, $\frac{k}{q}$ of the data values are less than x, and at most $\frac{q-k}{q}$ of the data values are more than x, where k is an integer such that $0 < k < q$. There are $q - 1$ q-quantiles.

Median, quartiles and percentiles

- The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median.
- The 4-quantiles are the three data points that split the data distribution into four equal parts (each part represents one-fourth of the data distribution). They are more commonly referred to as quartiles.
- The 100-quantiles are more commonly referred to as percentiles. They divide the data distribution into 100 equal-sized consecutive sets.
- The median, quartiles and percentiles are the mostly widely used forms of quantiles.

IQR

- The quartiles give an indication of a distribution's centre, spread and shape.
- The first quartile (denoted by Q_1) is the 25^{th} percentile. It cuts off the lowest 25% of the data.
- The second quartile is the 50^{th} percentile. As the median, it gives the centre of the data distribution.
- The third quartile, denoted by Q_3 is the 75^{th} percentile. It cuts off the lowest 75% of the data.

IQR

- The quartiles give an indication of a distribution's centre, spread and shape.
- The first quartile (denoted by Q_1) is the 25^{th} percentile. It cuts off the lowest 25% of the data.
- The second quartile is the 50^{th} percentile. As the median, it gives the centre of the data distribution.
- The third quartile, denoted by Q_3 is the 75^{th} percentile. It cuts off the lowest 75% of the data.
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the *interquartile range (IQR)* and is defined as

$$IQR = Q_3 - Q_1.$$

Example

- Suppose we have a set $\{1, 3, 4, 4, 5, 6, 8, 14, 14, 20, 22, 22\}$ which has 12 data points.
- The quartiles are the three values that split the sorted data into four equal parts. So, in this case, the quartiles are the third, sixth and ninth values.
- Thus, $Q_1 = 4$, $Q_2 = 6$, $Q_3 = 14$ and $IQR = 14 - 4 = 10$.

Measuring the dispersion (Variance and standard deviation)

- Variance and standard deviation indicate how spread out a data distribution is.
- A low standard deviation means the data tends to be close to the mean, while a high standard deviation indicates that the data are spread out over a large range.

Measuring the dispersion (Variance and standard deviation)

- Variance and standard deviation indicate how spread out a data distribution is.
- A low standard deviation means the data tends to be close to the mean, while a high standard deviation indicates that the data are spread out over a large range.
- As usual, we let $x_1, \ldots, x_N$ be our N observations of a numeric attribute X. Then the variance σ_X^2 of X is given by,

$$\sigma_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

the sum of all the x values from 1 to n

Measuring the dispersion (Variance and standard deviation)

- Variance and standard deviation indicate how spread out a data distribution is.
- A low standard deviation means the data tends to be close to the mean, while a high standard deviation indicates that the data are spread out over a large range.
- As usual, we let $x_1, \ldots, x_N$ be our N observations of a numeric attribute X. Then the variance σ_X^2 of X is given by,

$$\sigma_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

- The thinking here is that we want to know how far away each data point is from the mean. We then want to sum up these differences to get a picture of the total variance from the mean. The only problem is that some values are less than the mean (so will give a negative value), while some are greater than the mean (so will give a positive value). Thus, we will get cancellations.

Fixing this problem

- In fact, we will get 0, as the following shows:

$$\begin{aligned}\sum_{i=1}^{N}(x_i - \bar{x}) = & \quad \sum_{i=1}^{N}(x_i) - N\bar{x} \\ = & \quad \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i \\ = & \quad 0.\end{aligned}$$

Fixing this problem

- In fact, we will get 0, as the following shows:

$$\begin{aligned}\sum_{i=1}^{N}(x_i - \bar{x}) = & \quad \sum_{i=1}^{N}(x_i) - N\bar{x} \\ = & \quad \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i \\ = & \quad 0.\end{aligned}$$

- One way around this is to take the absolute value (i.e. just take the number and drop the sign) but this isn't that nice (i.e. it isn't smooth).
- However, there is another operation which *is* smooth and also kills minus signs - squaring. This is where the variance comes into play and it turns out to have a number of nice properties.

Central Limit Theorem (Again)

- One nice property is the presence of the Central Limit Theorem, which is at work whenever we measure the mean and standard deviation of a distribution we assume to be normal (which is a lot, particularly in nature).
- In a nutshell, the Central Limit Theorem (developed by de Moivre in 1733) states that as the size of a sample expands, the distribution of the mean among multiple samples will be approximately Gaussian. See `https://towardsdatascience.com/why-is-central-limit-theorem-important-to-data-scientist-49a40f4f0b4f`.

** Gaussian = normal distribution

- We can use this theorem to make all sorts of predictions about the entire distribution since a normal distribution is completely specified by its mean and standard deviation.

Standard deviation

- With the above in mind, we have introduced a square which in some ways is not what we want.
- So, to undo this in some sense, we can take the square root.
- This is what we call the standard deviation. In other words, the standard deviation is given by

$$\sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

A subtlety

- In the literature, there is another definition for variance which has a denominator N, *not* $N - 1$.
- The reason for this stems from the idea of bias. Without going into too much detail, if our statistic is *not* an under- or overestimate of a population parameter, then that statistic is said to be unbiased.

A subtlety

- In the literature, there is another definition for variance which has a denominator N, *not* $N - 1$.
- The reason for this stems from the idea of bias. Without going into too much detail, if our statistic is *not* an under- or overestimate of a population parameter, then that statistic is said to be unbiased.
- So, in other words, an unbiased estimator is an accurate statistic that is used to approximate a population parameter.
- If an over- or underestimate does happen, the mean of the difference is called a bias.

An example

- Suppose we want to find the average amount people spend on food per week in the UK.
- Obviously we cannot survey the whole population, so instead we take a sample of, say, 1,000.

An example

- Suppose we want to find the average amount people spend on food per week in the UK.
- Obviously we cannot survey the whole population, so instead we take a sample of, say, 1,000.
- From this, we find that the average amount spent on food per week is £60 per person. If £60 is indeed the average for the whole population, then this is unbiased. Otherwise, it is biased.

An example

- Suppose we want to find the average amount people spend on food per week in the UK.
- Obviously we cannot survey the whole population, so instead we take a sample of, say, 1,000.
- From this, we find that the average amount spent on food per week is £60 per person. If £60 is indeed the average for the whole population, then this is unbiased. Otherwise, it is biased.
- The causes of bias are all to do with how we took our sample. We need to know if the questions were biased (e.g. do we mean how much a person spent in that *given* week, or on an *average* week?) or whether the way we chose the sample was biased or if we have excluded parts of the population, and so on.

Bias and our formula

- Our formula (the one with $N - 1$ in the denominator) is called an *unbiased* estimate, or the *sample variance*.
- The use of $N - 1$ is called Bessel's correction as it corrects the bias in the estimate of the population variance.
- It also partially corrects the bias in the estimation of the population standard deviation. On average, with this formula, the sample variance is equal to the unknown population variance (the variance we would get if we did ask every person in the UK how much they spend on food).
- If there was an N in the denominator, the corresponding formula is called a biased estimate, or the population variance if we did know the actual mean of the whole population (i.e. we did ask everyone in the UK).

Example

- This example was taken from
  `https://towardsdatascience.com/variance-sample-vs-population-3ddbd29e498a`
  and also includes a brief discussion of the above, along with the Bessel factor.

Example

- This example was taken from
  `https://towardsdatascience.com/variance-sample-vs-population-3ddbd29e498a`
  and also includes a brief discussion of the above, along with the Bessel factor.
- Imagine a forest of 10,000 oak trees. This is the entire population and we want to estimate the distribution of heights. If we counted all 10,000, then we would find out that the heights are normally distributed with an average of 10m and a standard deviation of 2m. These are the statistical parameters of the *entire* population.
- However, we do not want to count 10,000 oak trees. We want to try and estimate these values. So instead, we take a sample of 20 random oak trees, measure their heights and then repeat this experiment 100 times.

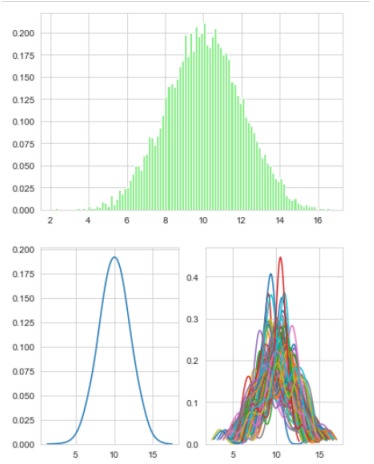
Example (Code)

```
import numpy as np
import pandas as pd
from random import sample,choice
import matplotlib.pyplot as plt
import seaborn as sns
import scipy


np.random.seed(1234)
mu = 10
sigma = 2
pop = np.random.normal(mu, sigma, 10000)
count, bins, ignored = plt.hist(pop, 100, density=True, color =
'lightgreen')
sns.set_style('whitegrid')
tests = 100
sam = []
mean =[]
std_b = []
std_u = []
fig, axs = plt.subplots(ncols=2)
sns.kdeplot(pop, bw=0.3, ax=axs[0])
for i in range(tests):
    sam_20 = np.random.choice(pop, 20)
    sns.kdeplot(sam_20, bw=0.3, ax=axs[1])
    sam.append(sam_20)
    mean.append(np.mean(sam_20))
    std_b.append(np.std(sam_20))
    std_u.append(np.std(sam_20, ddof=1))


frame = {'mean':mean, 'std_b': std_b, 'std_u': std_u}
table = pd.DataFrame(frame)
```

Example

- In this way, for each of the 100 experiments, we get an unbiased ($ddof = 1$ is the same as $N - 1$ instead of N) standard deviation '*std_u*', a biased standard deviation '*std_b* and a sample mean '*mean*'.
- Graphically, this can be seen in the following figure.

Example (Graphics)



Example

- Of course, 20 samples are a tiny subset of the 10,000 in the whole population.
- As such, every time we run the test, we get different distributions.
- The point is that, *on average*, we get a reasonable estimate of the real mean and real standard deviation.
- We can see this in Python using the command `table.describe` which shows that the unbiased sample standard deviation is on average nearer to the value of the population parameter than the biased one.

Example (Table)

	mean	std_b	std_u
count	100.000000	100.000000	100.000000
mean	10.095473	1.916363	1.966147
std	0.473076	0.279373	0.286630
min	9.104423	1.321537	1.355868
25%	9.781574	1.752017	1.797532
50%	10.046942	1.910105	1.959726
75%	10.479769	2.091813	2.146155
max	11.436420	2.819455	2.892699

AAD and MAD

- Recall, we discussed the sensitivity of the mean to outliers.
- Since the variance is computed using the mean, it is also sensitive to outliers. Indeed, as we have used the square difference, the variance is particularly sensitive to outliers.
- As a result, we sometimes want to use more robust estimates of the spread of a set of values.

AAD and MAD

- Recall, we discussed the sensitivity of the mean to outliers.
- Since the variance is computed using the mean, it is also sensitive to outliers. Indeed, as we have used the square difference, the variance is particularly sensitive to outliers.
- As a result, we sometimes want to use more robust estimates of the spread of a set of values.
- Three such measures are the *absolute average deviation (AAD)*, the *median absolute deviation (MAD)* and the *interquartile range (IQR)* which we have already discussed. The first two are defined as follows:

$$AAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

$$MAD = median(\{|x_1 - \bar{x}|, \ldots, |x_N - \bar{x}|\}).$$

Multivariate data

- As with the mean and median, we can generalise measures of dispersion to multivariable data. Once again, we can do this by looking at the spread of each attribute separately.
- For data with continuous variables, the spread of the data is most commonly captured by the *covariance matrix*, S, whose ij^{th} entry s_{ij} is the covariance of the i^{th} and j^{th} attributes of the data.

Covariance matrix is a type of matrix that is used to represent the covariance values between pairs of elements given in a random vector. This is because the variance of each element is represented along the main diagonal of the matrix.

Multivariate data

- As with the mean and median, we can generalise measures of dispersion to multivariable data. Once again, we can do this by looking at the spread of each attribute separately.
- For data with continuous variables, the spread of the data is most commonly captured by the *covariance matrix*, S, whose ij^{th} entry s_{ij} is the covariance of the i^{th} and j^{th} attributes of the data.
- For ease, suppose we have just two variables with the following samples: $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_N\}$. Then the covariance is given as follows:

$$cov(x, y) = \frac{1}{N-1} \sum_{k=1}^{N} (x_i - \bar{x})(y_i - \bar{y}).$$

Multivariate data

- As with the mean and median, we can generalise measures of dispersion to multivariable data. Once again, we can do this by looking at the spread of each attribute separately.
- For data with continuous variables, the spread of the data is most commonly captured by the *covariance matrix*, S, whose ij^{th} entry s_{ij} is the covariance of the i^{th} and j^{th} attributes of the data.
- For ease, suppose we have just two variables with the following samples: $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_N\}$. Then the covariance is given as follows:

$$cov(x, y) = \frac{1}{N-1} \sum_{k=1}^{N} (x_i - \bar{x})(y_i - \bar{y}).$$

- $covariance(x, x) = variance(x)$.

Covariance Matrix

- There are four possible combinations of x, y, leading to four different covariance calculations. We arrange this in the aforementioned covariance matrix, as follows:

$$\begin{pmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{pmatrix}.$$

- Of course, we can do this for any number of variables. If we have three variables x, y, z, each with N samples then we simply perform 9 covariance computations with the above formula and arrange them in a 3×3 covariance matrix.

Example

Suppose we have the following two samples: $\{4, 6\}$ and $\{2, 3\}$:

$$S = \begin{pmatrix} var(\mathbf{x}) & cov(\mathbf{x}, \mathbf{y}) \\ cov(\mathbf{y}, \mathbf{x}) & var(\mathbf{y}) \end{pmatrix}.$$

Example

Suppose we have the following two samples: $\{4, 6\}$ and $\{2, 3\}$:

$$S = \begin{pmatrix} var(\mathbf{x}) & cov(\mathbf{x}, \mathbf{y}) \\ cov(\mathbf{y}, \mathbf{x}) & var(\mathbf{y}) \end{pmatrix}.$$

First, we need to compute the means:

- $\bar{x} = \frac{4+6}{2} = 5$;
- $\bar{y} = \frac{2+3}{2} = \frac{5}{2}$.

Next, we need to compute the variance of each:

- $var(\mathbf{x}) = \frac{1}{1}((4-5)^2 + (6-5)^2) = 2$;
- $var(\mathbf{y}) = \frac{1}{1}((2-\frac{5}{2})^2 + (3-\frac{5}{2})^2) = \frac{1}{2}$.

Example

Suppose we have the following two samples: $\{4, 6\}$ and $\{2, 3\}$:

$$S = \begin{pmatrix} var(\mathbf{x}) & cov(\mathbf{x}, \mathbf{y}) \\ cov(\mathbf{y}, \mathbf{x}) & var(\mathbf{y}) \end{pmatrix}.$$

First, we need to compute the means:

- $\bar{x} = \frac{4+6}{2} = 5$;
- $\bar{y} = \frac{2+3}{2} = \frac{5}{2}$.

Next, we need to compute the variance of each:

- $var(\mathbf{x}) = \frac{1}{1}((4-5)^2 + (6-5)^2) = 2$;
- $var(\mathbf{y}) = \frac{1}{1}((2-\frac{5}{2})^2 + (3-\frac{5}{2})^2) = \frac{1}{2}$.

Finally, we compute the off-diagonal elements of S:

- $cov(\mathbf{x}, \mathbf{y}) = \frac{1}{1}[(4-5)(2-\frac{5}{2}) + (6-5)(3-\frac{5}{2})] = 1$;
- $cov(\mathbf{y}, \mathbf{x}) = \frac{1}{1}[(2-\frac{5}{2})(4-5) + (3-\frac{5}{2})(6-5)] = 1$.

Example

Putting this all together, the covariance matrix is given by,

$$S = \begin{pmatrix} 2 & 1 \\ 1 & \frac{1}{2} \end{pmatrix}.$$

x = y =

Exercise: Consider the following samples: $\{92, 60, 100\}$ and $\{80, 30, 70\}$. Show that the covariance matrix is as follows:

x mean = 84 y mean = 60
xi - xmean = {8, -24, 16}
(xi - xmean)^2 = {64, 576, 256}
cov(x,x) = 448

$$S = \begin{pmatrix} 448 & 520 \\ 520 & 700 \end{pmatrix}.$$

Code in Spyder

Correlation

- The covariance of two attributes is a measure of the degree to which two attributes vary together and depends on the magnitudes of the variables.
- A value near 0 indicates that two attributes do not have a (linear) relationship, but it is not possible to judge the degree of relationship between two variables by looking only at the value of the covariance.
- Nevertheless, we can roughly say that if x and y are completely unrelated to each other, we would expect $\text{cov}(x, y) = 0$.
- Likewise, if high values of x generally correspond to low values of y then we would expect $\text{cov}(x, y) < 0$, and vice versa $\text{cov}(x, y) > 0$.
- In this way, covariance allows us to assess whether two variables are correlated.

Using NumPy

In Python we can calculate the covariance matrix using NumPy, as follows:

```
import numpy as np

# Generate some anti-correlated data

# X1 and X2 are like the columns of some data
X1 = np.array([-1,0,1,2,3])
X2 = np.array([3,2,1,0,-1])

# Stack these into a 2d array
X = np.vstack((X1,X2))

# Calculate the covariance matrix
C = np.cov(X)
#[[2.5,-2.5,]
# [-2.5,2.5]]
```

The highly negative off-diagonal terms in this example show that `X1` and `X2` are strongly anti-correlated, such that as one increases the other decreases in lockstep.

Pearson's correlation coefficient

- Because the correlation of two attributes immediately gives an indication of how strongly two attributes are (linearly) related, correlation is preferred to covariance for data exploration.
- We define the *correlation matrix* R as follows: the ij^{th} entry of R is the correlation between the i^{th} and j^{th} attributes of the data.
- If $x_i,\ x_j$ are the i^{th} and j^{th} attributes, respectively, then

$$r_{ij} = correlation(x_i,\ x_j) = \frac{cov(x_i,\ x_j)}{\sigma_i \sigma_j},$$

  where $\sigma_i,\ \sigma_j$ are the standard deviations of $x_i,\ x_j$, respectively.
- The diagonal entries of R are $correlation(x_i,\ x_i) = 1$, while the other entries are between -1 and 1.
- This correlation measure is called *Pearson's correlation coefficient*.

Interpreting Pearson's correlation

- If $r_{ij} = 0$ then there is no correlation between x_i, x_j.
- On the other hand, if $r_{ij} = 1$ then x_i, x_j are perfectly correlated.
- Finally, if $r_{ij} = -1$, then x_i, x_j are perfectly anti-correlated.

In Python

```
from scipy import stats

# Assuming x and y are numpy arrays containing the samples
r,p = stats.pearsonr(x,y)
```

Note that the function returns both the correlation coefficient and a *p*-value. The *p*-value is the probability of *x* and *y* producing a value $\geq R_{xy}$ if they are drawn from an uncorrelated distribution. In other words, a measure of the significance of the result.

Summary

- We have seen measures of central tendency. These include:
  - mean
  - median
  - mode
  - midrange
- We have seen measures of dispersion. These include:
  - variance
  - standard deviation
  - IQR
- Multivariate statistics
  - covariance
  - Pearson's correlation