INTRODUCTION TO STATISTICS

Institute of Trading & Portfolio Management

Contents

What is Statistics?

Statistics is a mathematical science concerned with the collection, analysis, interpretation or explanation, and presentation of data. It is often described as the science of decision-making under conditions of uncertainty and as such, a basic knowledge of statistics can contribute significantly to trader's and portfolio manager's toolsets.

Statistics can be split into two broad categories:

- **Descriptive Statistics** are used for organizing, presenting and summarizing data.
- **Inferential Methods** are often the next stage of analysis, where conclusions are drawn about a population based on data observed in a sample.

Populations *refer to the entire group of data that you want to draw conclusions about. Smaller* ***Samples*** *are used to infer conclusions about the larger population. Most of the time, statistical methods are conducted on samples of data, not populations.*

Quantitative vs Qualitative Data

At the highest level, two types of data exist – quantitative and qualitative. Quantitative data is in numeric form and is a result of either counting or measuring attributes of an observation or dataset. Qualitative data usually categorizes or describes attributes of observations or datasets. Statistical methods used in this video series will only be utilizing quantitative data.

Time-Series vs Cross-Sectional Data

Time-series data refers to a set of observations taken over time at specific and equal intervals. For example, the Price/Earnings ratio of Microsoft on a weekly basis from 1st January 2000 to the current date.

Cross-sectional data refers to a set of observations taken at a specific point in time, but across a range of observational units. For example, the Price/Earnings ratio of all the companies in the S&P500 index on the 27th October 2011.

Univariate vs Multivariate Analysis

Univariate analysis is the most basic form of analyzing data. "Uni" means one and so it refers to analyzing data with only one variable.

- Example: Analyzing the daily returns of Apple stock over time

Multivariate analysis examines two or more variables. Often, it involves testing the relationship between a dependent variable and a number of other independent variables.

- Example: Analyzing and testing for relationships between the returns of both Apple and Amazon stock over time.

The vast majority of statistical analysis in ITPM educational courses is in the form of univariate analysis. The three major characteristics typically analyzed in a single variable are:

- The distribution
- The central tendency
- The dispersion

The following pages discuss these attributes in more detail.

The Distribution

A dataset's (empirical) distribution is a summary of the frequency of individual data-points that lie within certain ranges of values for a given variable. Typically, the distribution encompasses all the data within a sample and the ranges of values are ordered from smallest to largest. Charts can be used to allow for easy visualization of the values/observations and the frequency with which they appear.

Example – US Adult Male Height

Dataset: Heights of 1000 randomly selected adult males within the United States

*Note that this is a **sample** dataset, which we can analyze to draw conclusions about the **population** dataset (all adult US males)*

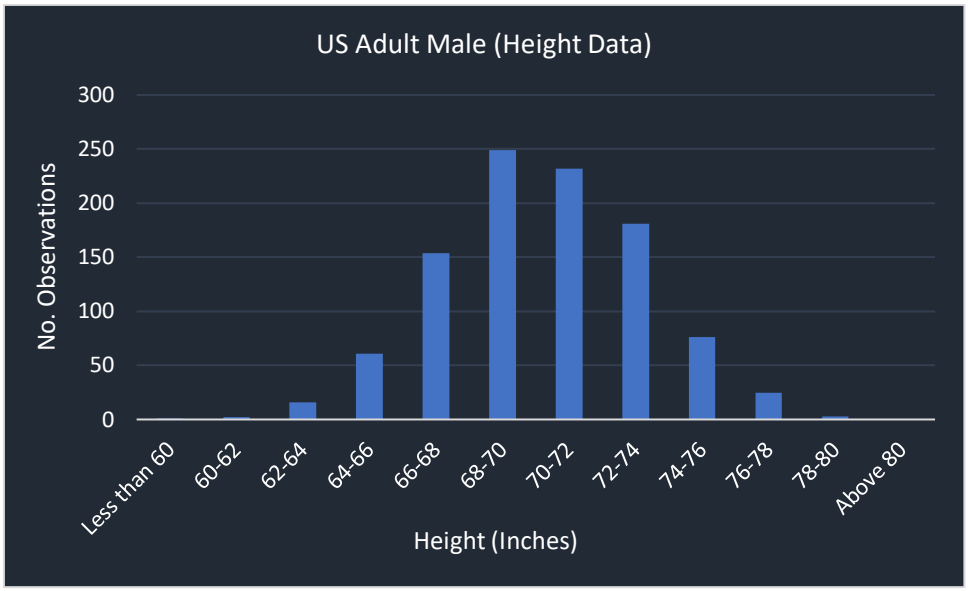
The following frequency table was generated from the sample data:

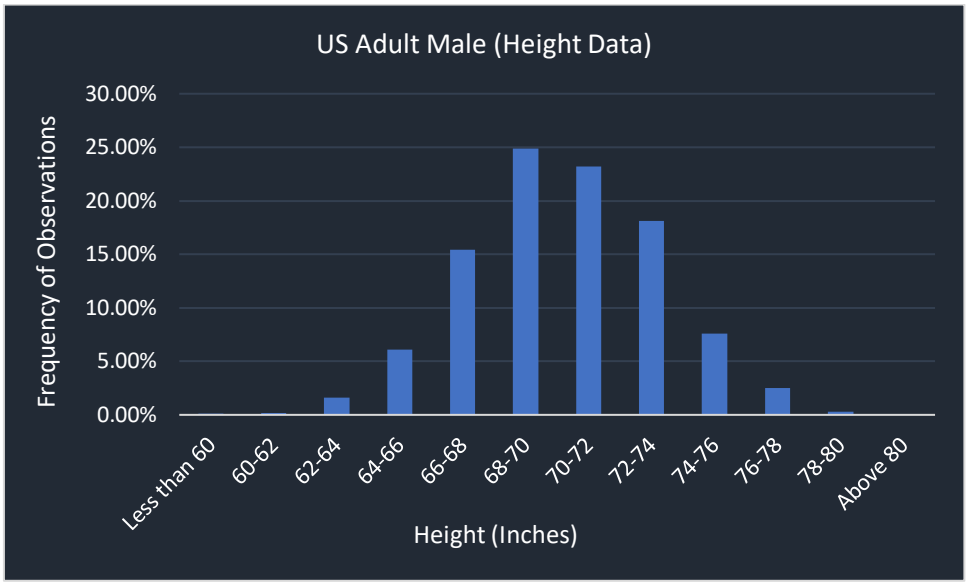
Frequency Table			
Height (Inches)	**Frequency**	**Probability**	**Cumulative Probability**
Less than 60	1	0.10%	0.10%
60-62	2	0.20%	0.30%
62-64	16	1.60%	1.90%
64-66	61	6.10%	8.00%
66-68	154	15.40%	23.40%
68-70	249	24.90%	48.30%
70-72	232	23.20%	71.50%
72-74	181	18.10%	89.60%
74-76	76	7.60%	97.20%
76-78	25	2.50%	99.70%
78-80	3	0.30%	100.00%
Above 80	0	0.00%	100.00%

The frequency table counts the number of observations (US adult males' height measurements) that lie within certain ranges. The data tells us that out of the 1000 observations, 232 US adult males are between 70-72 inches (177.8-182.9 centimetres) tall. 232 observations equates to 23.2% of the entire sample (of 1000 observations) and so we might infer that 23.2% of all US adult males are within that height range. We also might infer that for any randomly selected adult male in the United States there is a 23.2% probability that their height will be within the range of 70-72 inches. In a robust analysis, we would then test those hypotheses for statistical significance, but that is beyond the scope of this Introduction to Statistics PDF.

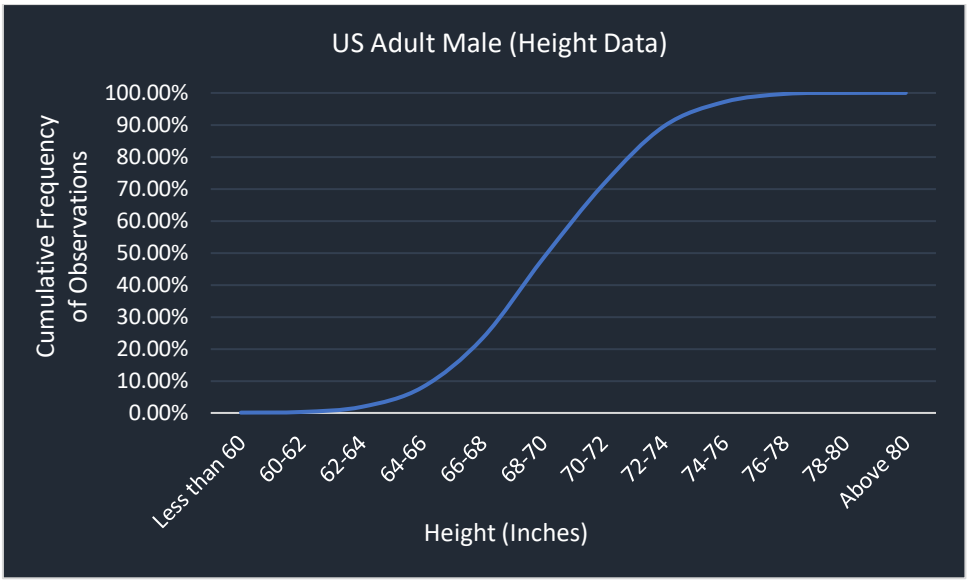
The final column in the Frequency Table above shows the cumulative percentage of observations that lie within all the height ranges up to and including the current one. For instance, we can say that 89.60% of observations have a value of 74 inches or less, and from that data we can infer conclusions about the population in a similar manner as we did previously.

To make all this data more visually accessible we can view it as a column chart or histogram, counting either the number of observations or percentage number of observations for each range of heights:





We can also view the cumulative probability distribution as a line chart:



Normal (Gaussian) Distributions

Normal or "Gaussian" Distributions are most commonly found in nature and human behaviour. To give some practical examples, a Normal Distribution can explain the probability of being a certain height, weight or IQ level. They exhibit a bell-shaped curve, which is symmetrical around the mean (arithmetic average) of the sample being analysed. Larger datasets will result in more accurate probability distributions and so in cases where a normal distribution applies, more data will show empirical analysis tending towards normality. We will take a look at the "normality" of our US Adult Male Height distribution as we continue into the next sections of the PDF.

It is important to note that the Normal Distribution is the most commonly used probability distribution within the finance industry when it comes to creating quantitative models, largely because of the simplicity in its calculation. For example, it is not uncommon for an asset's or portfolio's risk to be calculated under the assumption of "normality" when considering their return distributions, and the Black-Scholes Options Pricing Model also has this assumption embedded within it. However, as we shall see later in the document, and in video 3 and 4, traders/investors should be aware that it is often not a perfect representation of asset returns.

The Central Tendency

When analyzing a dataset, one of the first calculations usually done is to find the "middle" of the data. However, defining the middle (or average) isn't as obvious as you might think. There are several things you will need to think about when deciding what measure to use for the "middle", but first let me list the various measures that you might use:

Mean

The arithmetic mean if the measure most people think about when trying to find the middle, or the average of a dataset. The arithmetic mean is simple the sum of observational values divided by the number of observations. Consider the following dataset of 10 observations:

1, 2, 3, 4, 4, 4, 5, 7, 9, 10

$$Arithmetic\ Mean = \frac{1+2+3+4+4+4+5+7+9+10}{10} = 4.9$$

There are many other different types of means, and each of them have their own special use cases and characteristics. For instance, there is the *geometric mean* (often used for calculating compounded annual growth rates), the *weighted mean* (which could be used to find the portfolio returns given you know the asset weightings within the portfolio and their individual returns) and the *harmonic mean* (which might be used to calculate an average P/E ratio, for example). Understanding when to use each other type of mean is unnecessary for the analysis we will be doing throughout this PDF and most of the ITPM video series', and so there is no need to go into any more detail about these.

Median

The median is the middle observation in a dataset, of the observation that splits the lower and upper halves of the data in a dataset. For example, in a dataset of 9 observations arranged in ascending order of values, the median would be the 5th number. In a dataset with an even number of observations, the arithmetic mean of the middle two observations is taken as the median. Consider the previous dataset of 10 observations, the middle two numbers are observations 5 and 6 (with values of 4 and 4 respectively) when organised into ascending order. Note that we take the values of both observations 5 and 6 because that splits the data in half (there are 4 observations below these values and 4 above). So, in this case:

$$Median = \frac{4+4}{2} = 8$$

Why might the median be preferable to the mean? Whilst the median can be used for more types of data than the mean, probably the most important reason to use the median instead of the mean is in a circumstance where the mean might be affected significantly by outliers in the data. For example, imagine if the dataset for Height of US Adult Males included an observation with a value of 1 billion. Clearly this is an outlier (and in this case an error in the data). It would skew our calculation for the arithmetic mean much higher than it probably should be, and so the median then becomes a more reliable measure of centrality within the dataset.

Mode

The mode is the value that appears most frequently in the dataset, or in other words the value that is most likely to occur when sampling the dataset. Much of the time, the mode is not that useful within financial analysis since it can easily take multiple values, or as in many continuous datasets, the observations likelihood of the same value occurring more than once is very small.

Central Tendency in US Adult Male Height Data

Below is a table of descriptive statistics for the US Adult Male Height dataset:

Descriptive Statistics	
Mean	70.19
Median	70.10
Mode	68.30
Variance	9.12
Standard Dev	3.02
Minimum	58.80
Maximum	79.60
Range	20.80
25th Percentile	68.20
50th Percentile	70.10
75th Percentile	72.30
Skewness	-0.05
Kurtosis	-0.11
Count	1000

The top three statistics are the measures of central tendency we have discussed. In a Normal Gaussian distribution, the Mean, Median and Mode are all equal to each other. In this case, the sample data shows the mean and the median are very similar, but the mode is slightly different with a value of 68.30. When the dataset is analyzed more closely, it becomes clear that this is just one of many values for the mode. The value of 68.30 occurs 17 times in the data set, but so do the values of 68.9, 69.5, 69.6, 71.8 and so on. So, in this case it's probably best to assess the "Modal Range" rather than the mode itself and we can see from our histogram data that the most frequently occurring range is 68-70, closely followed by 70-72 which leads us to the conclusion that the true Mode is likely to be very close to our Mean and Median values. We know that in a Normal Distribution the Mean = Median = Mode and therefore at this stage we can already be fairly certain that the sample distribution of US Adult Male heights is close to being normally distributed.

The Dispersion

So far, we have analyzed both the empirical distribution and the central tendency of our data. Now we move on to assessing the variability of values around the mean, and there are several ways to do this:

Range

The range of the dataset is a very simple measurement which calculates the difference between the maximum and minimum values observed. For example in the US Male Height dataset the minimum value is 58.80 inches and the maximum is 79.60. Therefore:

$$Range = 79.60\ inches - 58.80\ inches = 20.80\ inches$$

Inter-quartile ranges

We can take our analysis of the range a step further by arranging the datasets values low to high and splitting the dataset into 4 equally sized groups, meaning the same number of observations fall into each group. So, the bottom 25% of values would fall into the first quartile, and next 25% into the second quartile, and so on. This type of analysis is like a simplified version of our

frequency table, and we can see work out the ranges of these quartiles by finding the observation values at the 25th, 50th and 75th percentiles of the distribution, and then calculating the difference between these values:

$$1st\ Quartile\ Range = 25th\ Percentile - Minimum = 68.20 - 58.80 = 9.4\ inches$$

$$2nd\ Quartile\ Range = 50th\ Percentile\ (Median) - 25th\ Percentile = 1.9\ inches$$

$$3rd\ Quartile\ Range = 75th\ Percentile - 50th\ Percentile\ (Mediand) = 2.2\ inches$$

$$4th\ Quartile\ Range = Maximum - 75th\ Percentile = 79.60 - 72.30 = 7.3\ inches$$

Mean Deviation

The mean deviation calculates how far, on average, all the individual observations are from the arithmetic mean of the dataset. To do this, the absolute difference of each observation to the mean is summed and then divided by the total number of observations. Mathematically:

$$Mean\ Deviation = \frac{\sum |X - \mu_x|}{N}$$

Where: X = observation value

μ_x = arithmetic mean of the dataset

N = number of observations in the dataset

$\sum$ is shorthand for "sum of"

Note that we take the absolute difference (denoted by the modulus symbols | |) because we do not care whether the values lie above or below the mean, but only their difference to the mean in absolute terms (absolute terms means taking any value and making it positive).

Variance

Variance and standard deviation (to follow) are typically preferred measurements of average dispersion from the mean to the mean deviation. Variance is calculated by taking the sum of the squared differences from the mean of each observation, and dividing that by the total number of observations. The reason we square these numbers is because we still only care about the differences of each observation to the mean and not whether they are positive or negative. Squaring these differences is a more algebraically sound method of doing this that taking the modulus as in the mean deviation equation.

$$Variance = \frac{\sum (X - \mu_x)^2}{N}$$

Where: X = observation value

μ_x = arithmetic mean of the dataset

N = number of observations in the dataset

$\sum$ is shorthand for "sum of"

There is a slight difference between the equations for calculating variance for a population and variance for a sample but this isn't something we need to go into. When using calculations for variance and standard deviations in Excel, typically we will show you the sample functions for

these calculations as most of the time we will be taking a sample of data to draw conclusions about the population.

The Variance of our US Adult Male Height dataset is equal to 9.12. Unfortunately, this figure isn't easily understandable because its units are inches squared, not inches, which leads us onto the Standard Deviation.

Standard Deviation

Standard deviation is the most commonly used value for volatility in the finance industry, and one of the statistics we will be using most in this course. It is calculated by simply taking the square root of the variance, and by doing so changes the units of the output back to the same as those of the dataset.

$$Standard\ Deviation = \sqrt{Variance}$$

Within Normal probability distributions:

1 Standard Deviation either side of the mean accounts for 68.27% of the data

2 Standard Deviations either side of the mean accounts for 95.45% of the data

3 Standard Deviations either side of the mean accounts for 99.73% of the data

An empirical distribution can be analyzed to assess the percentage of data that lies within 1, 2 and 3 standard deviations from the mean in order to compare the results with that which we expect in a Normal distribution. For example, for our Adult US Male Height dataset:

Standard Deviations	Lower Bound	Upper Bound	No. Observations	Frequency %	Normal Freq %
1	67.17	73.21	682	68.20%	68.27%
2	64.14	76.23	957	95.70%	95.45%
3	61.12	79.25	998	99.80%	99.73%

We can see that 682 observations lie within 1 standard deviation from the mean (67.16 inches to 73.21 inches), which as a percentage equates to 68.20% of the data. When the same analysis is done for 1, 2 and 3 standard deviations from the mean we can see that the percentage frequency of observations we see in the empirical data is remarkably close to what we see in a Normal distribution. This is another way to conduct a quick test for normality in the dataset. You could then go a step further and formally conduct a hypothesis test to see if the difference between the frequency % and normal frequency % values are statistically significant. However, as outlined earlier, this is beyond the scope of this document.

The Central Moments

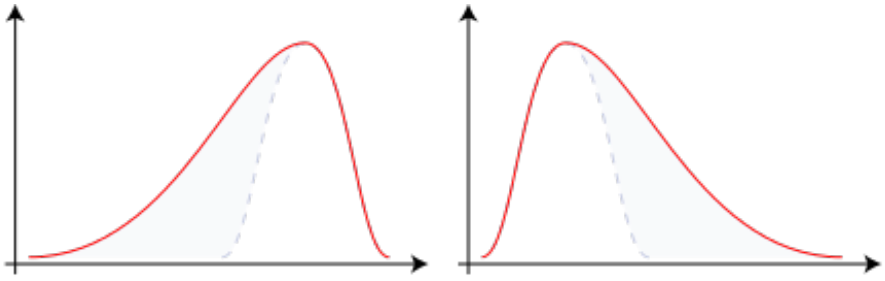
Finally, we will consider central moments. In statistics, the central moments describe the shape of a probability distribution around it's mean - essentially providing us with values that help us usefully characterize and understand the data. Without going into too much detail, there are three values that you need to know something about, and they are the 2nd, 3rd and 4th central moments. We've already

covered variance, which is the 2nd central moment and describes a dataset's dispersion around its mean. So that leaves us with the 3rd and 4th central moments - Skewness and Kurtosis respectively.

Skewness

Skewness is the third central moment and measures the tendency for the data to be distributed either above of below the mean. Normally distributed data has a skew of 0, whilst negative or left-skewed data has value below 0 and positive of right-skewed data has a value above 0.

It is easiest to understand skewness visually through charts. As you can see below, negatively skewed data has a long tail to the left, whilst positively skewed data has a long tail to the right.



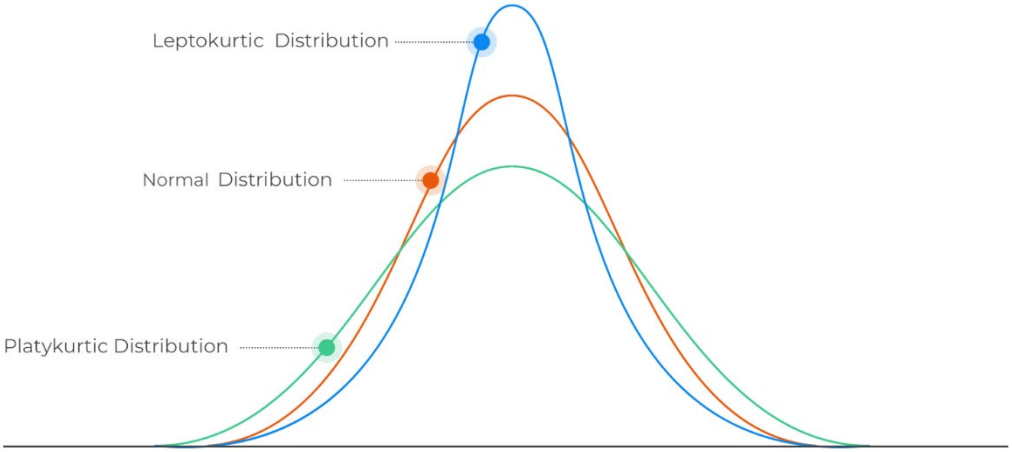
Negative Skew Positive Skew

If you imagine this in terms of asset returns, a negative skew would mean that extreme values are more likely to be negative than positive, which is often the case in reality (although to a lesser extent than the charts above depict).

Looking back at our Height dataset, we can see that it has a skewness value of -0.05 which is very close to 0 and so further indicates a normally distributed profile.

Kurtosis

The fourth central moment, kurtosis, is a measure of the "peakiness" or "tailedness" of a distribution. Normal distributions have a kurtosis value of 3 and an excess kurtosis value of 0. Distributions with excess kurtosis values above 0 are "peakier" around the mean, meaning that a larger portion of the probability distribution is localised around the mean than a Normal distribution would predict. Positive excess kurtosis also exhibits "fatter tails" which indicates that both extreme positive and negative values occur more frequently than a Normal distribution would predict. Distributions with positive excess kurtosis are called Leptokurtic distributions, as opposed to Platykurtic distributions which have negative excess kurtosis.



The descriptive statistics table for our height dataset is generated in Excel and the Kurtosis value listed is actually the excess kurtosis. So, since that value is -0.11 it is also in agreement with all of previous analysis in that it indicates that the dataset of US Adult Male Heights is very close to a Normal distribution.

Correlation & Covariance

Covariance

Covariance provides insight into how two variables are related to one another. More precisely, covariance refers to the measure of how two random variables in a data set will change together. A positive covariance means that the two variables at hand are positively related, and they move in the same direction. A negative covariance means that the variables are inversely related, or that they move in opposite directions.

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Where:

$$x = variable\ 1$$

$$y = variable\ 2$$

$$\bar{x} = mean\ of\ x\ variable$$

$$\bar{y} = mean\ of\ y\ variable$$

$$n = number\ of\ observations\ in\ each\ data\ set$$

$$\sum_{i=1}^{n}(\dots) = sum\ of\ \dots from\ observation\ 1\ to\ n$$

Correlation

Correlation is very similar to covariance and can is sometimes referred to as the "normalised" covariance. It's a statistical measure that indicates the extent at which two variables fluctuate with each other, so essentially the degree to which two variables are related. There are actually many different measures of correlation but the most common one, and the one we use throughout our courses at ITPM is the Pearson Coefficient of Correlation.

The Pearson Coefficient Correlation is defined below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Where:

$$x = variable\ 1$$

$$y = variable\ 2$$

$$\bar{x} = mean\ of\ x\ variable$$

$$\bar{y} = mean\ of\ y\ variable$$

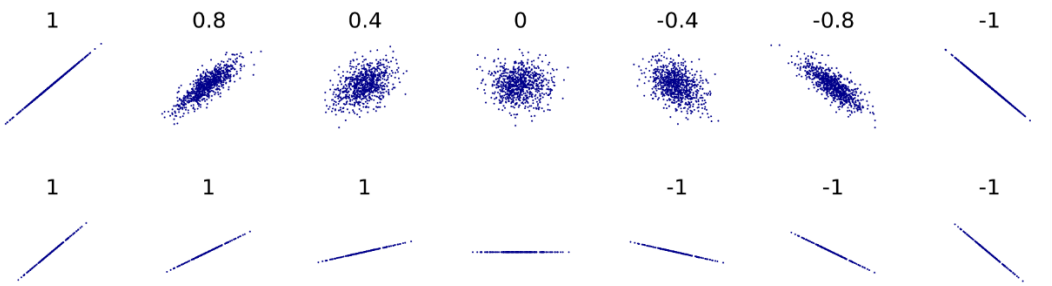
$$n = number\ of\ observations\ in\ each\ data\ set$$

$$\sum_{i=1}^{n}(\dots) = sum\ of\ \dots from\ observation\ 1\ to\ n$$

The formula calculates the "pairwise" correlation between two variables, in linear terms. Although it isn't necessary to use this formula directly (there are shortcuts in Excel to get to the same answer!) it is useful to understand how it works. For example, you might think that if both variables you are analysing have positive data, then the relationship or correlation between them must also be positive. Unfortunately, you would be wrong in that assumption! What actually happens is that every observation of both the X and Y variables are converted or normalised to values that represent their distances from the mean. Those values are then compared against each other to compute an average relationship over the dataset. The output shows whether the two variables have a positive of negative linear relationship and how strong that relationship is.

Output values of the Pearson Correlation Coefficient range between values of +1 and -1, or 100% and -100%, where +1 represents perfect positive correlation and -1 perfect negative correlation. A measure of 0 would suggest the two variables are perfectly uncorrelated, and there is no linear relationship between them. However, that doesn't necessarily mean the variables are independent – as they might have a relationship that is not linear.

Scatterplot charts are a good way of visualizing various values for correlation. Scatterplots are charts in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present:

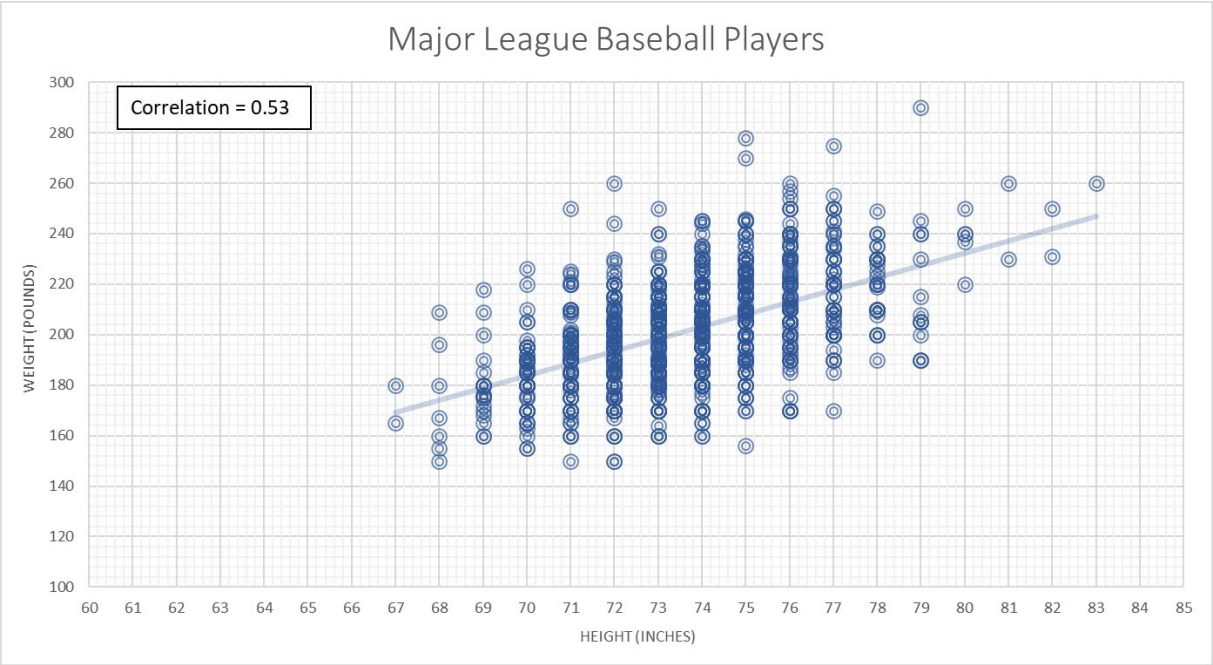


Essentially, you can see that an upward sloping relationship shows a positive correlation, while a downward sloping relationship shows a negative one. The dispersion around that linear relationship shows the strength of weakness of correlation, which you can see in the top row of charts.

An important point here is that the slope or steepness of the line is somewhat irrelevant to the strength of correlation. The slope shows positive, negative or 0 correlation, whilst dispersion shows the strength of the correlation.

Major League Baseball Players

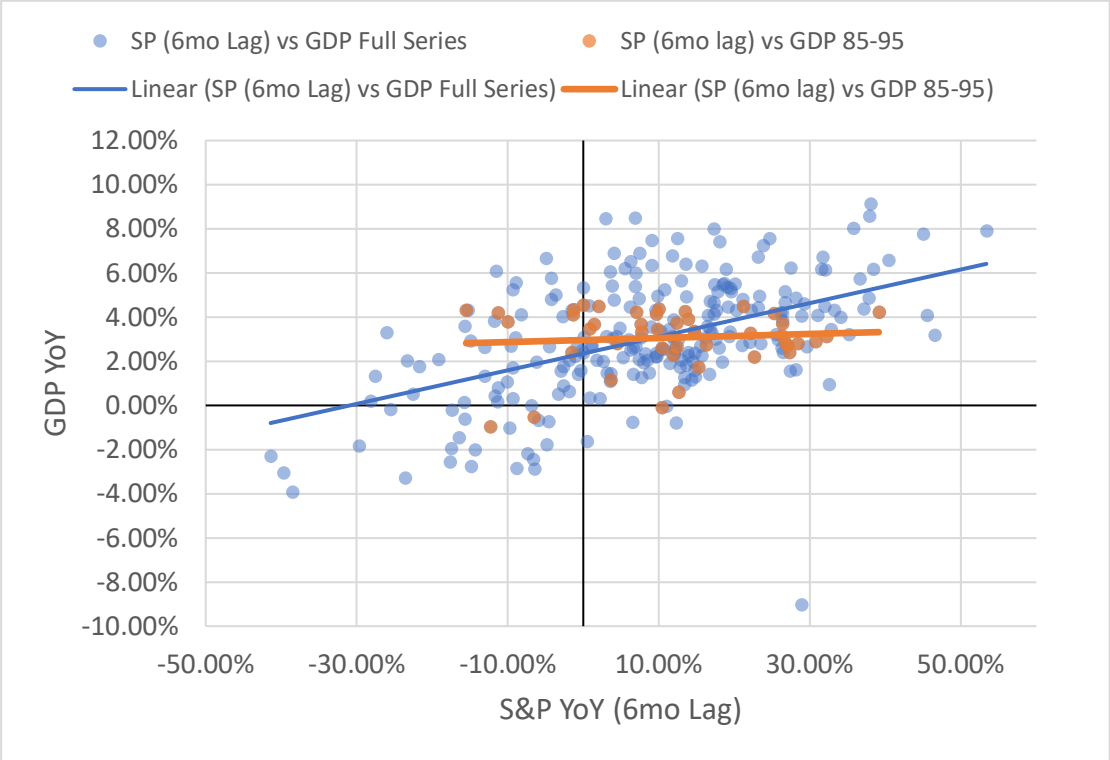
Let's look at a real work example now. Below is a scatterplot showing height vs weight for Major League Baseball players:



As you can see, there is a fairly strong positive correlation (0.53) between player's height and weight, which is what we would expect. The line of best fit through the middle of the data is drawn by minimizing the sum of the squared distances of each observation to the mean of the data set. It basically represents a kind of average linear relationship. An upwards sloping line (bottom left to top right) shows that the correlation between the two variables (height and weight) is positive. The variance with which the observations for each player are dispersed around the line dictates the strength of the correlation value.

S&P 500 vs GDP

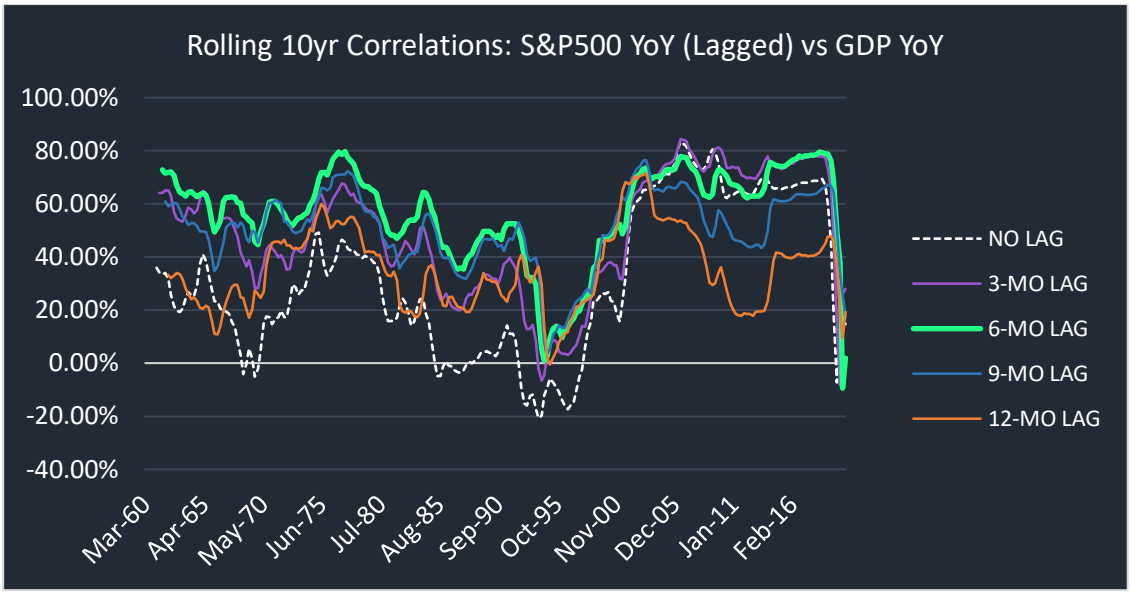
Now let's look at a more relevant example for us as traders and portfolio managers.



The chart above shows US Real GDP Year on Year percentage values on the y-axis (vertical), plotted against S&P500 Year on Year percentage returns on the x-axis (horizontal). S&P500 returns have been lagged by 6 months, since the aim of this data analysis is to understand if there is a predictive linear relationship between the two variables.

The blue data represents the entire data set available, and the orange data represents data only between the years of 1985-1995. We can see that in the data set as a whole the upward sloping blue line shows there is a positive correlation between the variables (Pearson Coefficient = 0.25) but yet if you strip out the years from 1985-1995 the relationship weakens, and the orange line is (close to) flat which means that there is no identifiable linear relationship over that period.

You can see this quite clearly in the following rolling correlation chart which calculates the rolling 10 year correlation between S&P500 YoY and GDP YoY observations, with the S&P500 data being lagged by various amounts (to see which time lag provides the strongest linear explanation for GDP YoY values).



As you can see, the rolling 10-year correlation between S&P500 YoY (6mo lagged) returns and GDP YoY values is around 0% in the year 1995 which makes sense given the data we looked at in the scatterplot previously.

Hopefully these examples give you a better understanding of what correlation is, how it can be used and why it is useful to us as traders and portfolio managers.

Common Pitfalls of Correlation Analysis

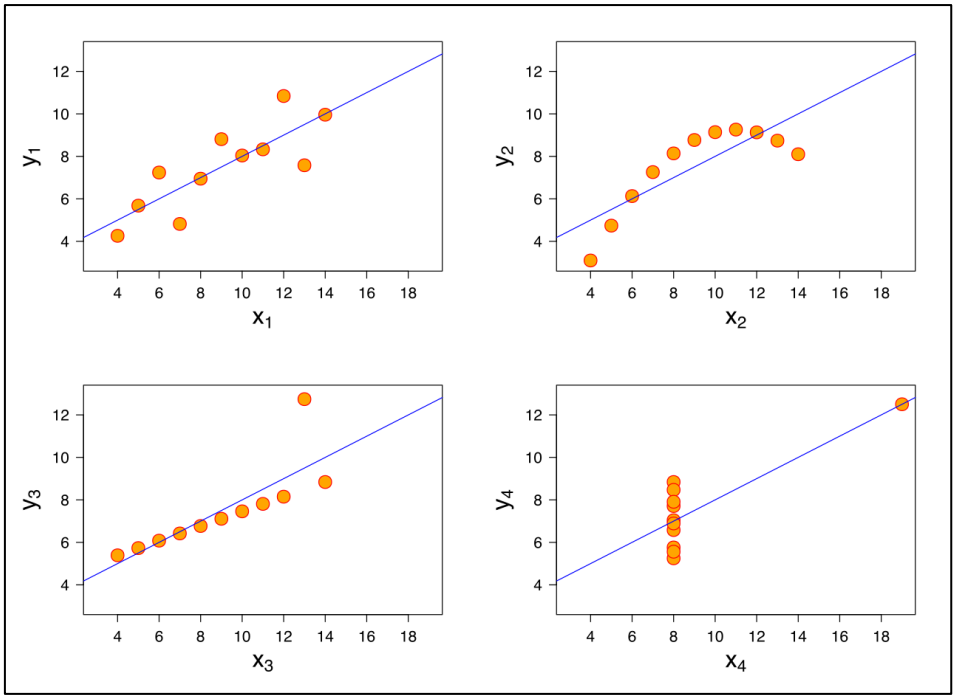
To finish off this section on correlation, it is important to understand some of the common pitfalls and drawbacks of correlation analysis so you are best prepared when you do your own research.

1. Correlation is not Causation!

   Just because two variables have showed a historic correlation doesn't mean that one of the variables *causes* the other to move. The causation of the two variables moving with a positive or negative correlation could be a third completely unconsidered variable, OR a combination of many factors. In theory, we want to try and understand the causes for relationships between variables so we can have a more accurate idea about when those relationships might change and if they will. The reality is that this is very hard to achieve and so practically speaking correlation is analysis is often used to surmise relationships and use them as forward-looking predicator under the caveat that we understand it is likely that there are many factors at play that are responsible for the causation of the relationship.

2. Correlation is just a number!

   Correlation can be easily misinterpreted if you don't understand it properly. If you want to truly understand the relationship between two variables you must do much more than just look at some descriptive statistics and correlations. **Anscombe's Quartet** is a great example of this:



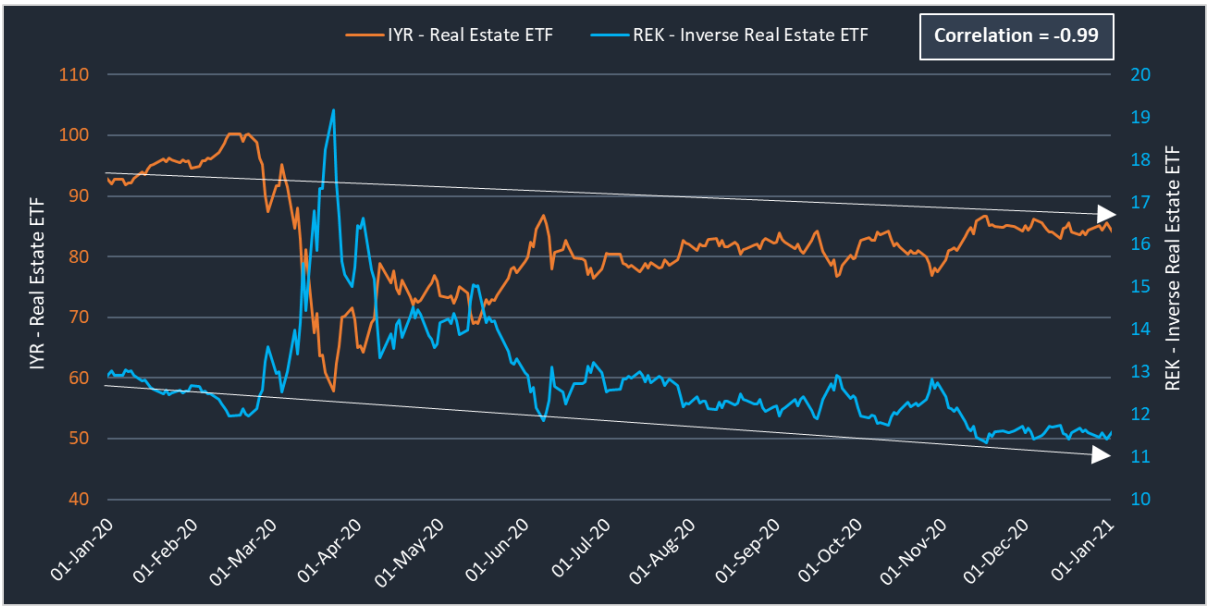
The above four datasets are called Anscombe's Quartet. They all have the following characteristics:

Characteristic	Value
Mean (x)	9
Variance (x)	11
Mean (y)	7.5
Variance (y)	4.125
Correlation	0.816
Linear Regression	y = 3 + 0.5x
Coefficient of Determination	0.67

As you can see, the four datasets are all very different, yet they have exactly the same descriptive statistics. The point here is that correlation, like any other statistic, is just a number. You need to understand when it is useful and when it isn't, and often to do that you will have to use other investigative methods to really understand the data.

3. Correlations can be strongly positive (negative) whilst the data trends in the opposite (same) direction!

The Pearson Coefficient of Correlation relies on the relative difference of each observation from the mean of the dataset, but the means of the x and y variables can be very different. A good example of this is seen when looking at the Inverse Real Estate ETF (REK) and the Real Estate ETF (IYR).



These two assets have an almost perfect negative correlation, as you would expect, however over the year 2020 they both trended in the same direction. Positive (negative) correlations do NOT mean that data has to trend in the same (opposite) direction.

4. Correlation is dependent on the dataset used, even if the two variables are the same!

This sounds obvious, but if we think about it in terms of financial assets, it means the correlation between the returns of two assets will change dependent on the time period and frequency of data used to calculate correlation.

A very important example of correlations changing over time is looking at what happens during crises, when liquidity in the market dries up. In this table below, correlations have been calculated between the S&P 500 and various other assets by first filtering for days on which the VIX is at various levels, to act as a proxy for liquidity levels. You can see that as the VIX increases, correlations across assets, particularly in equities, tend to 1. This have many implications for traders and portfolio managers, not least of which is a double-whammy effect on portfolio volatility – not only do individual asset volatilities increase, but correlations magnify this effect at the portfolio level!