# Introduction to Working with Data Collected Over Time (a Time Series).

It is well known that one of the four V’s of Big Data are Velocity, Volume, Veracity and Variety. To focus on the last ‘V’, Variety, data can come in all sorts of shapes, sizes and even formats. Given the increase in computing power, many unstructured data sources such as text, audio and image data have become important new sources of information for both science and industry. However, one structured source of data that has stood the test of time, no pun intended, is time series data.

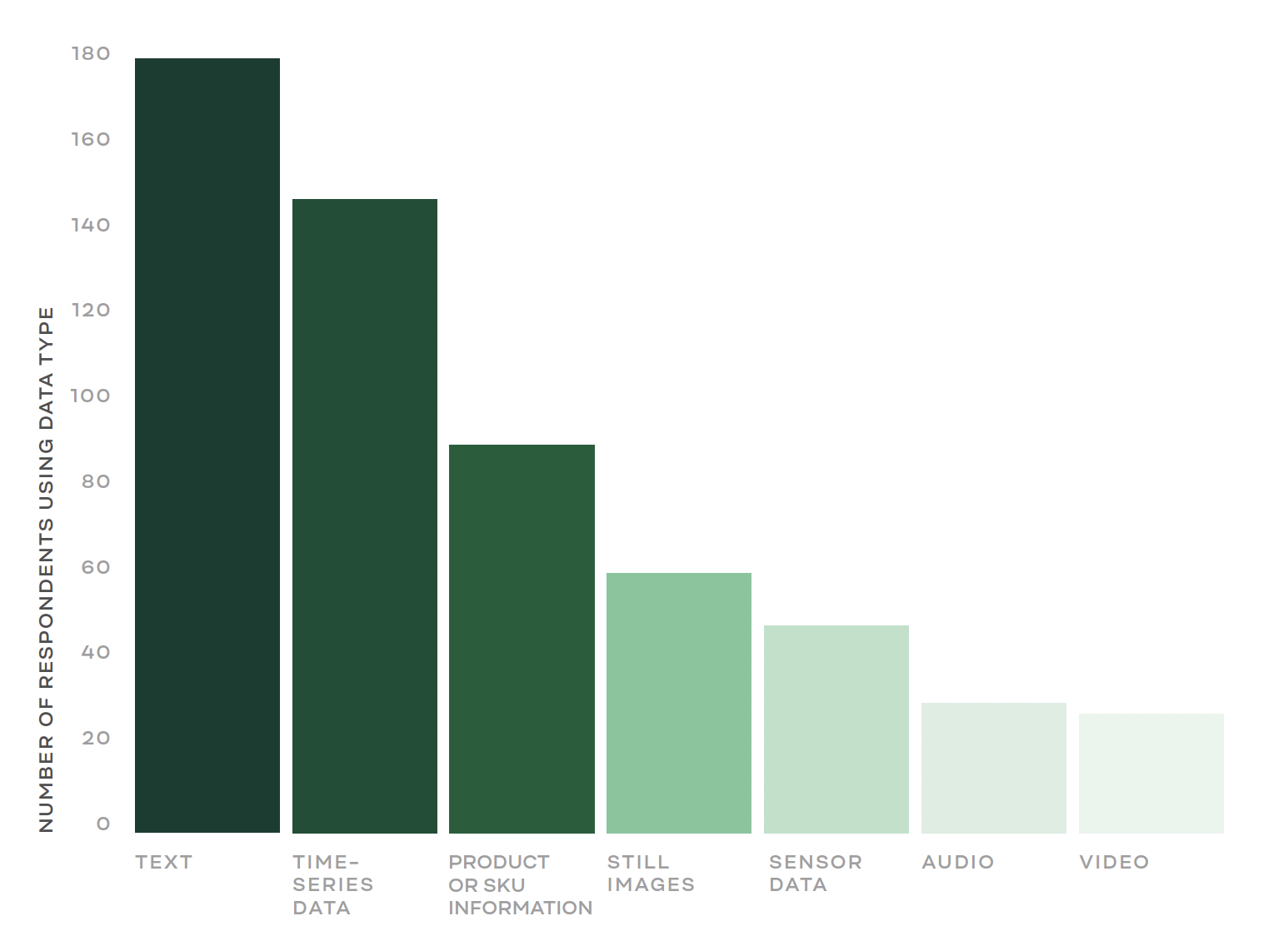


Fig 1.1. Data Types Encountered by a Survey of Data Scientists (Figure 8 Data Scientist Report 2018)

## What is a Time Series?

### 1.1.1Notation

Let’s begin by looking at an example. Below is a plot of the sales of a new shampoo at a particular local Walmart. The *x*-axis is time (day) and the *y*-axis is the response (sales). As a matter of convention, is known as the response or dependent variable, the independent variable is time.

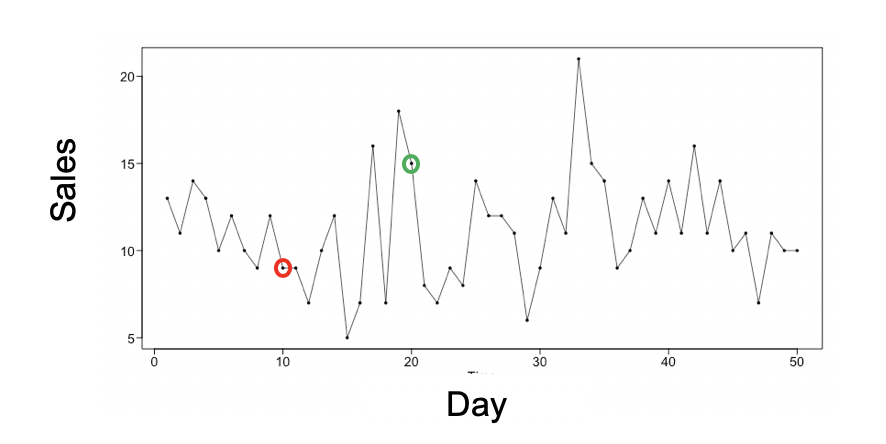


Fig 1.2. Shampoo sales recorded at a particular Walmart

Formally, we will define to be the random variable that represents sales at day *t* and will define the collection of all (one for each *t*) as a *stochastic process* and is also referred to as a *time series*. Furthermore, we will let *xt* represent a particular value of from the sample. For example, in the plot above, = 9 and similarly = 15.

### 1.1.2 Realizations

Once we have observed and recorded *x1, x2, x3,* … *xn* (where n is the size of our sample) we have what is called a *realization* of the time series which can be thought of as a particular instance of the time series.

Consider again the shampoo sales data in Figure 1.2. These data represent a realization of the Walmart shampoo time series. Next assume that the data in Figure 1.3 represent the sales of the shampoo at 4 additional Walmarts. If we can assume that these additional 4 Walmarts are very similar to each other, we can consider these data to be 4 additional realizations of the same time series.

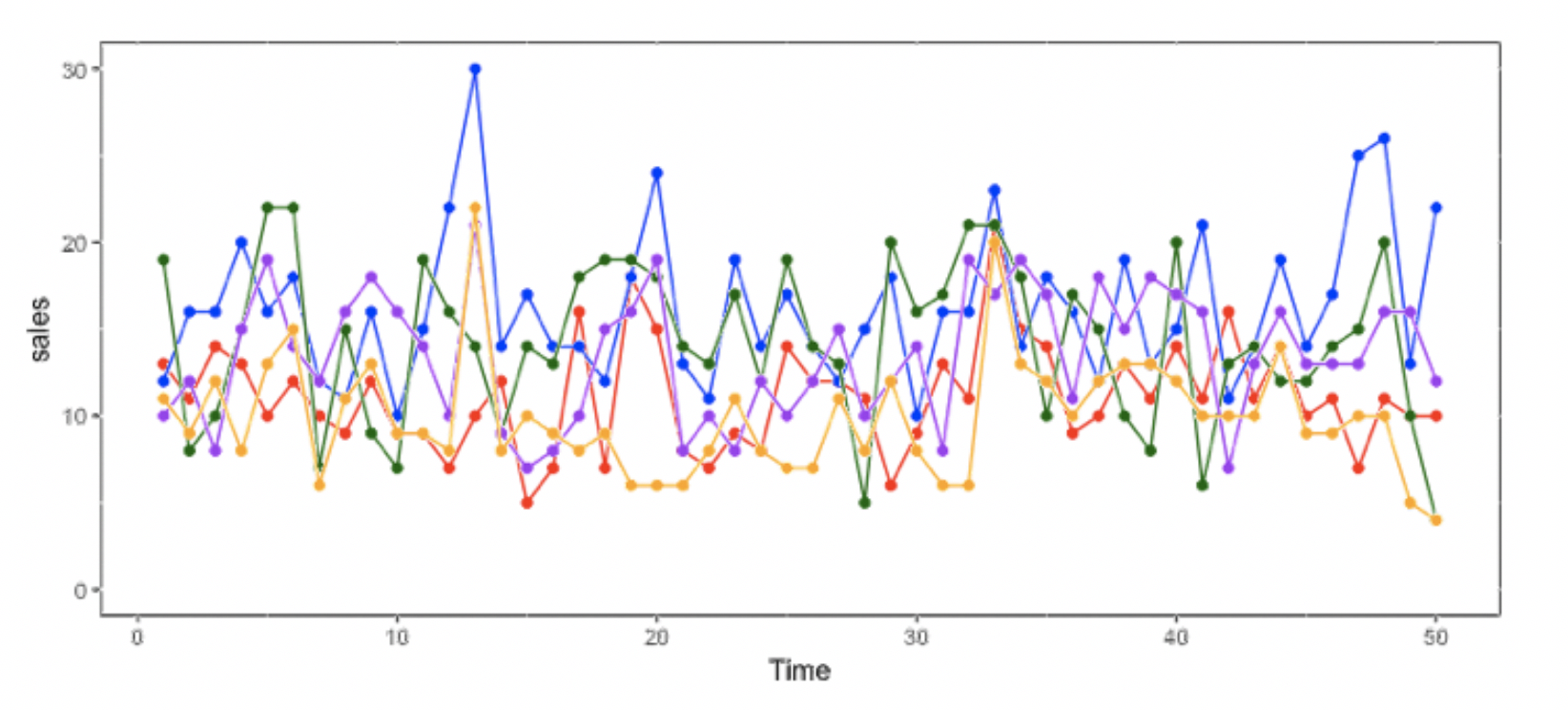


Fig 1.3. Shampoo sales recorded over the same period at 4 very similar Walmarts

In most cases it is impossible to observe more than one realization. Take for instance the data in Figure 1.4 which represent the West Texas Crude Oil Price from December 2009 to December 2019. Note that in this time series it is only possible to observe one realization since it is impossible to role back time and see how the prices would play out in an alternate reality.

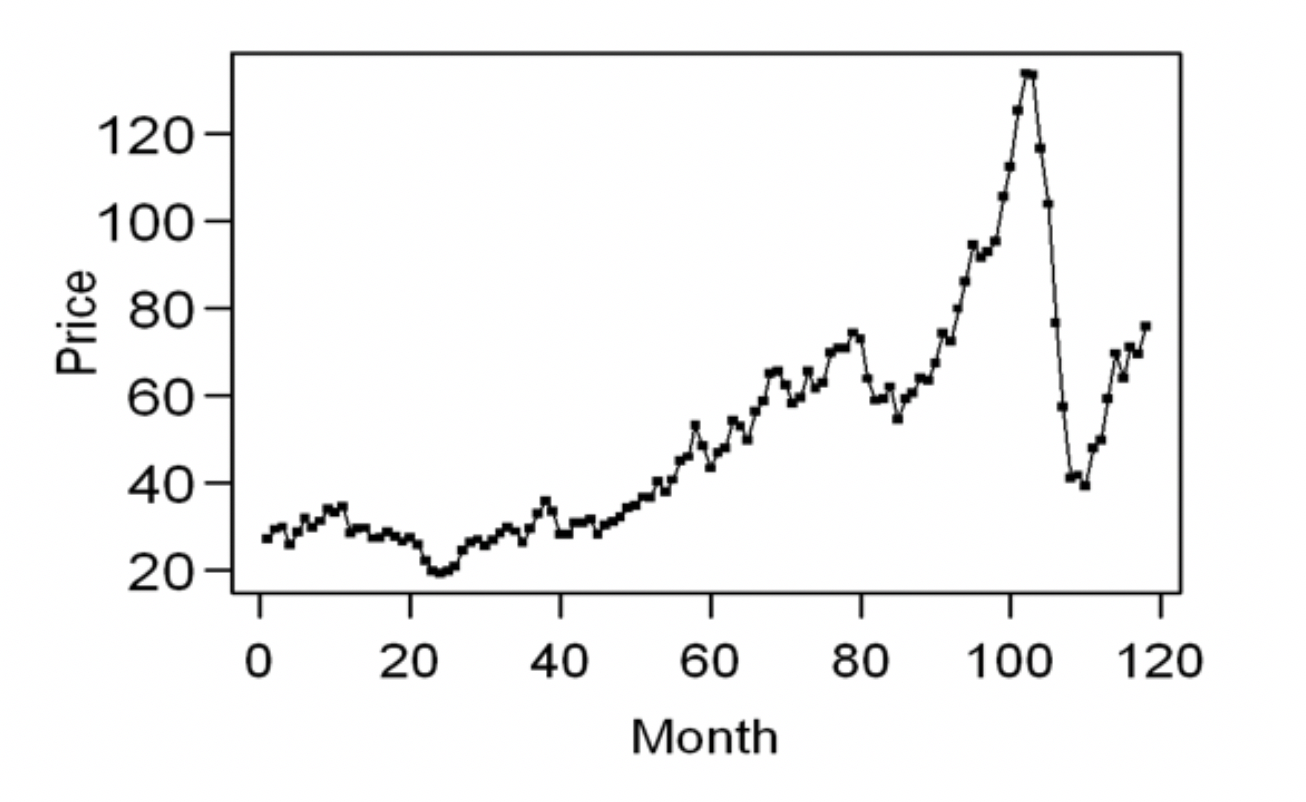


Fig 1.4: West Texas Crude Oil Prices from December 2009 to December 2019.

On the other hand, if we were able to observe all possible realizations of a time series we would have the *ensemble* of the time series. In most practical situations, the ensemble must be imagined which is often where domain knowledge, the understanding of the given practical situation, plays a crucial role.

### 1.1.3 Examples

In Fig 1.5 we see several realizations of various time series that provide evidence of the wide variety of characteristics that they may possess. The data can seem to wander (as in (a) and (b)) and may have increasing or decreasing trend as is evident in the global temperature data and stock data (a and c below). A time series may also display a strong seasonal component as seen in the sunspot data (c) or may even have a combination of these behavior as can be the case in the beer sales data (d). Time series such as the Doppler signal (e) and seismic data (g) display the possibility that these characteristics themselves may change over time.

|  |  |  |
| --- | --- | --- |
| a | b | c |
| d/var/folders/9j/d95918x55sbf4r8nq2_lj6s40000gn/T/com.microsoft.Word/Content.MSO/D1668F38.tmp | e | f |

Fig. 1.5. Example of 6 different time series: a) global temperature data, b) Google Stock Price, c) Sunspot Data, d) Beer Sales Data, e) Doppler Signal, f) Seismic Data.

# Chapter 2

***In our discussion of time series analysis, it will be very important to have a solid understanding of the concepts of expected value, variance, covariance and correlation. While it is assumed that the reader has been exposed to these concepts, we will provide a brief review of each fundamental concept before addressing it in the time series setting****.*

## Expected Value

### In General

1. **Discrete**

The expected value of a random variable *X* **(*E[X]*) is the mean** or intuitively, the long-run average of the event that variable represents.

For example:

Consider a variable *X* that represents the amount of money won in a particular casino game. *X* represents the amount won or lost (negative amount) and *P*(*X*) represents the probability of that event. We could organize this idea in a table:

|  |  |
| --- | --- |
| ***X*** | ***P(X)*** |
| $2 | .1 |
| $0 | .3 |
| –$1 | .6 |

Table 2.1: Discrete Probability Distribution of X

This table indicates that there is a 10% chance that the player will win $2, a 30% chance that they will break even (win / lose $0), and a 60% chance that they will lose $1.

Should casino goers play this game? If they played this game many times, what would be their average winnings or losings?

These questions can be addressed by looking at the expected win or loss (the expected value) each time they play.

**–$0.4**

In this game, each time the player plays this game, he or she is expected to lose, on average, $0.4, or 40 cents.

Spoiler alert, all games in Las Vegas have a negative expected value!

In general, we have calculated the overall mean of an infinite number of outcomes from an infinite number of plays of this game, That is, the expected value is the population mean:

1. **Continuous**

Recall, in the case when X is a continuous random variable, we will assign probabilities to ranges of X through a probability density function (pdf), f(x), and will then calculate the expected value of X by:

For example, when X can take on any value between 2 and 6 with equal probability (a uniform distribution) we have the situation as seen below:

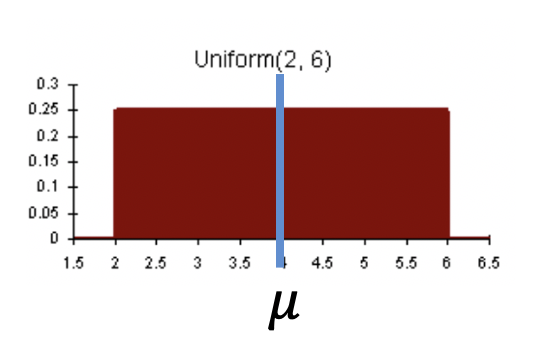


Figure 2.6 Continuous Probability Distribution of X: (Uniform(2,6))

=

All pdfs must be nonnegative and must integrate to one; however, they may take on many different shapes. An example involving the well-known normal distribution is seen to the below. Note that the probability distributions have very different shapes but have the same expected values.

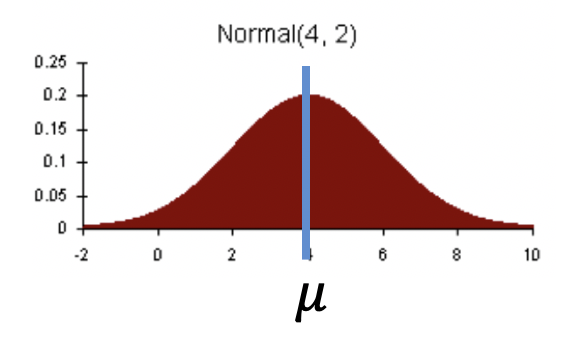
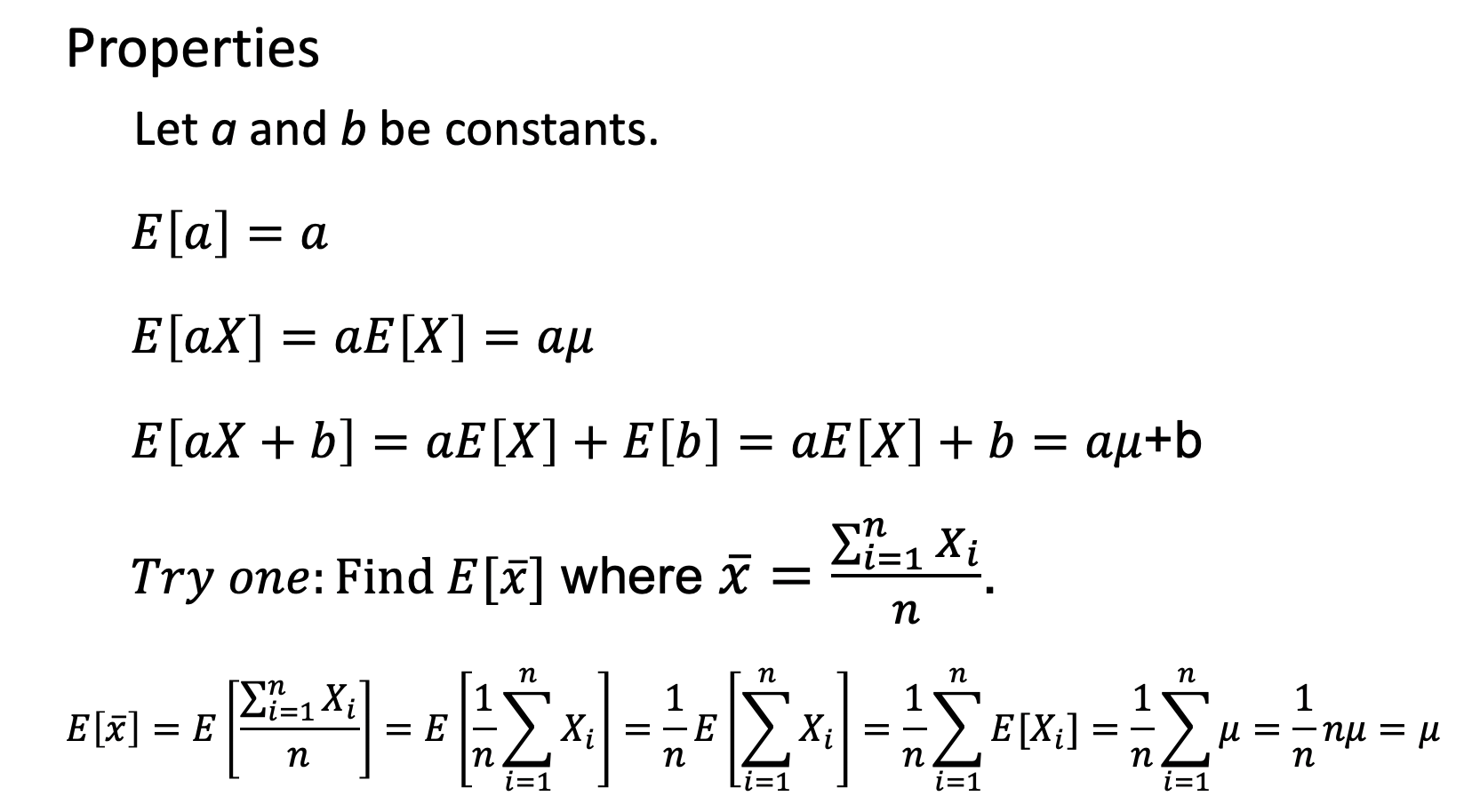


Fig 2.7: Continuous Probability Distribution of X: Normal(4,2)

Throughout this book we will study the properties and characteristics of several models and the expected value will be one of these characteristics. Below are properties of the expected value that we will use in our exploration of these models:



### In Time Series

Recall from earlier that we will be working with time series data recorded as We will be discussing (and estimating) the expected value of the random variable .

Mean of :

Note here that is the mean of all possible realizations of ***for a fixed t.*** To be clear, by fixing t, we are only addressing those values recorded at that time.

For example, we do not have the ensemble of all realizations from the Walmart example, but from the five we are given, let’s estimate for *t* = 10.

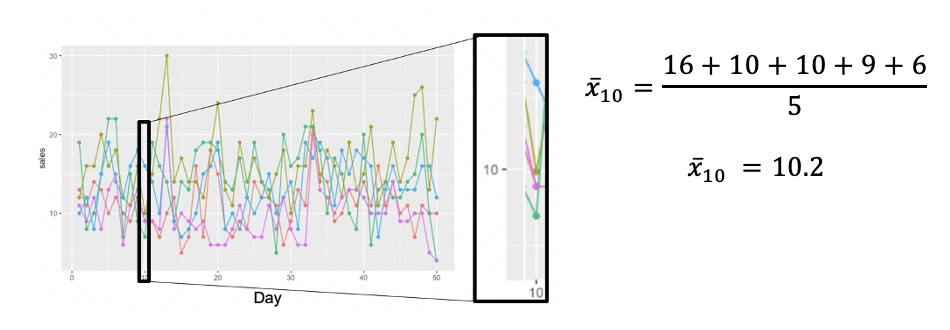


Fig 2. 8: Plot of 5 realizations of Walmart Data with focus on the 5 observations at t = 10.

Note that we only looked that the observed values of and 6. The other values of the series were not used in the estimation of the expected value.

Looking at another example below for t = 33, we see that the five observed values are all greater than the five we observed for t = 10: 20,20,21,24 and 17. This may be evidence that the process that generated the data has a higher population mean at t = 33 than it does at t =10 or it may be that they share the same population mean and the 5 observations at t = 33 just all tended to be somewhat higher. Given that we only have a sample of 5 at each value of t, we will never know or for certain. However, we do have evidence (the data).

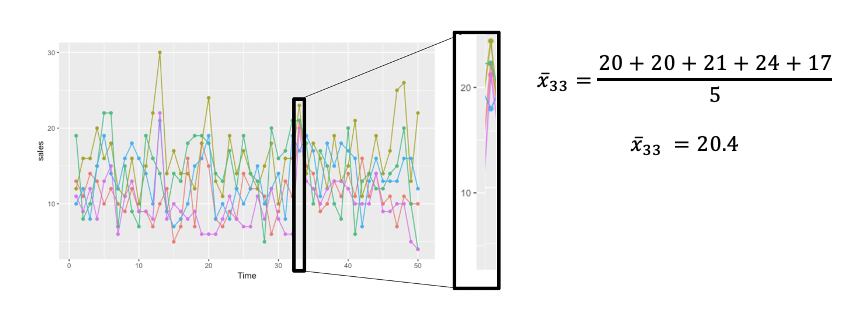


Fig 2.9: Plot of 5 realizations of Walmart Data with focus on the 5 observations at t = 33.

## Variance

### In General

#### Discrete

While the expected value effectively measures the center of the distribution, it is also useful to measure how spread out the values are. There are many measures of this spread and one of the most widely used measure is the *variance:****Var(X)***. Specifically, the variance measures the extent to which the values are dispersed around the expected value **(*E[X]*).** In fact, this measure can be expressed as the expected value off this difference:

For example, for the discrete variable X defined in the casino example from before, the variance is defined as

From above, we see that the variance is a function of the distance each value is from the mean. As a mathematical note, if we were to simply add these distances together (, the result would be exactly zero. Therefore, we will sum the squared differences () which will provide us with the sum of only positive values so that the variance is directly proportional to the magnitude of the sum of the squared differences. Finally, each distance is weighted by its probability of occurring, () yielding the equation seen in (reference of the Var(X) equation above.)

In the casino example we have:

|  |  |
| --- | --- |
| ***X*** | ***P(X)*** |
| $2 | .1 |
| $0 | .3 |
| –$1 | .6 |

= 0.408 dollars2

Since it is reasonable to work with the original scale of the variable (dollars instead of squared dollars in this case), we will often work with the standard deviation of X (SD(X)) which is simply the square root of the variance:

= dollars = $0.6387

1. **Continuous**

The analog for the variance of a continuous random variable is similar.

In contrast to a discrete random variable, when our random variable is continuous there are an uncountably infinite number of values that X can take on. The above integral represents the average squared distance from the population mean of every possible value of X. This is assumed here to be an infinite number of values thus highlighting one of the great powers of calculus (no one has the time to average an infinite number of values!). Furthermore, it may be that some values of X are more probable to occur than others and should thus receive more weight in the average. This “weight” is supplied by the in the equation and represents the probability density function or pdf. You have likely heard of a very widely used pdf: the normal distribution.

In practice, we will need to estimate the variance. We will naturally estimate with the sample mean ( and we will essentially take the average of the squared differences of and the n values in our sample: . Note that in the equation below we are dividing by n-1 rather than n. This is known as the degrees of freedom and provide the statistic desired mathematical properties (see reference for more information.) For our purposes it is safe to think of the variance as the average squared difference an observation is from the sample mean.

As before, since the variance is measured in the square of the units that the data was originally recorded in. It will be preferred in a practical sense to work with the standard deviation which is measured in the units the data are recorded in. Again, as before, the estimated standard deviation is simply the square root of the estimated variance:

### In Time Series

Again, working with the Walmart data, we will assume that the sales ( are continuous and will now investigate the variance of . We recall that the mean of for a given time is noted as so that:

Mean of :

Therefore,

Where is the total number of observations of the ensemble (the complete set of possible realizations) at time t.

Since the ensemble is rarely obtainable (most of the time we may only have one possible realization of the ensemble) we will need to estimate with the sample variance:

With respect to the Walmart data from above, we could estimate a separate variance estimate for each t. For example, when t = 10, we witnessed earlier that the values of the realization were =16,=10,=10,=9 and =6 and the sample mean was thus 10.2. Therefore, the estimate of the variance is

and thus,

Similarly, for t = 33 we have:

and thus,

|  |
| --- |
|  |
|  |
| Fig 2.10 Side by side plots comparing the sample mean and variance of t = 10 versus t = 33. |

A natural question at this point is if there is enough evidence to suggest the process that is generating these realizations has a different mean and/or variance at t = 10 and t = 33. If these are independent realizations, we could conduct a simple hypothesis test for the differences of means and variances. However, what happens if we only have one realization? We will return to that question in the next chapter! For now, we are on track if the concept variance and how it applies to a fixed time over multiple realizations is understood.

## Covariance and Correlation

### In General

1. Covariance

Next, consider two random continuous variables *X* and *Y*. Another tool we would like is one that will measure the extent to which X and Y move together. That is, as *X* gets bigger, is there evidence that *Y* tends to get smaller or larger? This can be measured with the **covariance** of *X* and *Y*: ***Cov(X,Y).***

Before we define Cov(X,Y), let’s observe a plot in which as X gets larger, Y also tends to get larger (Fig 2.11A). In Fig 2.11 B we see the same plot although this one indicates where the average of the y coordinates of the points is (the horizontal line) and the average of where the x coordinates is (the vertical line). Note in the case where as X increases Y tends to increase we have that when x values are below their mean, the y values tend to also be below their mean (lower left of Fig 2.11B) and when the x values are above their mean the y values tend to be above their mean (upper right of Fig 2.11B).

|  |  |
| --- | --- |
| Fig 2.11 A | Fig 2.11 B |
| Fig 2.11: Cov(X,Y) > 0 | |

Now let’s define the Cov(X,Y) as:

And we will estimate this metric with

Note that is a function of the products of the differences of the x and y coordinates and their means. In the case described above, when is negative will tend to be negative and their product will be positive; furthermore when is positive will tend to be positive and their product will again be positive. Therefore, since most of the products will be positive (and the ones that are negative will have relatively small magnitude) the overall sum and thus will tend to be positive.

Alternatively, when we have the case as in Fig 2.12A where, as X increases, Y tends to decrease we have the opposite behavior. When the x coordinates are below their mean the y values tend to be above their mean and will tend to be negative and when the x coordinates are above their mean will tend to be negative as well (and again the ones that are positive will have small magnitude). Therefore, in this case, we see that the sum of all the products () and thus will tend to be negative.

|  |  |
| --- | --- |
| Fig 2.12 A | Fig 2.12 B |
| Fig 2.12: Cov(X,Y) < 0 | |

Finally note the case when X and Y are not related as is evident in Figure 2.13. When the x coordinates are below their mean, the corresponding y coordinates are scattered rather randomly above and below the mean so that will be both positive and negative and with similar magnitudes. This behavior will be similar when the x coordinates are above their mean as well. Therefore, for all points, this will lead to be as likely to be positive as it is to be negative with similar magnitudes so that will be close to zero.

|  |  |
| --- | --- |
| Fig 2.13A | Fig 2.13 B |
| Fig 2.13: Cov(X,Y) 0 | |

Note that in this last case, for low values of X we observed both high and low values of Y and for high values of X we also observed high and low values of Y. It appears here that knowing the value of X does not tell us any information about the corresponding value of Y. This quality is consistent with the concept of X and Y being uncorrelated and leads us to the following fundamental result:

If X and Y are uncorrelated then their covariance is 0: Cov(X,Y) = 0.

However, carefully note that the converse is not true. That is, if Cov(X,Y) = 0, this does not imply that X and Y are uncorrelated. As a quick example, note Fig 2.14 in which we see that knowing X clearly gives some information as to where to expect its corresponding Y. However, the covariance of X and Y is a measure of the ***linear*** relationship between X and Y and thus, in this case, is 0. This is precisely the reason why any assessment of being uncorrelated with respect to the covariance should be accompanied with a visual inspection of the scatterplot.

|  |
| --- |
|  |
| Fig 2.14: In this case, X and Y are clearly associated although the covariance is near 0. |

1. Correlation

While the covariance is certainly a very useful statistic, one of its shortcomings is that the magnitude of Cov(X,Y) does not easily reflect the strength of the linear relationship between X and Y. Take for instance Fig. 2.15a and b in which there is clearly a much stronger linear relationship between X and Y in (b) than there is in (a). However, we note that the Cov(X,Y) = 557.143 in (a) and is only Cov(X,Y) = 4.771 in (b). This is simply because the scale in (a) is much larger than the scale in (b) (note the y axis of each plot.) Since the covariance is a function of the distance between x and and y and and these distances are much larger in (a).

|  |  |
| --- | --- |
| a)  Cov(X,Y) = 557  Corr(X,Y) = .082 | b)  Cov(X,Y) = 4.771  Corr(X,Y) = .990 |
| Fig 2.15 a) Data in which X and Y do not appear to have a very strong linear relationship. b) Data in which X and Y appear to have a very strong linear relationship. | |

A simple adjustment to the covariance is to divide (or scale) it by the standard deviation of X and Y. The resulting statistic is called the correlation of X and Y and is estimated by:

The effect of this scaling of the Cov(X,Y) is that the Cor(X,Y) is bounded by -1 and 1 inclusively. No matter what the scale of X and Y, the stronger the linear relationship, positive or negative, the closer the Cor(X,Y) will be to 1 or -1 respectively. In addition, the weaker the linear relationship between X and Y, the closer the correlation will be to 0. Note again Figure 2.15 and notice that although (a) has the larger covariance (Cov(X,Y) = 557), it has a correlation much closer to 0 (Corr(X,Y) = .082) and while (b) has the smaller covariance (Cov(X,Y) = 4.771) its correlation is much closer to 1 (Corr(X,Y) = .990) which reflects the clear visual difference in the strength of the linear relationship in (a) and (b).

### In Time Series (Autocovariance and Autocorrelation)

Recall above that we discussed the possibility of the mean (E[Xt]) and variance of Xt may depend on t. That is E[Xt] may be different than E[Xk] when t ≠ k and that Var(Xt) may be different than Var(Xk) when t ≠ k. *In this section it will be critical to assume that the mean and variance are constant across the entire realization; for any t.* That is, to discuss autocovariance and autocorrelation, we will assume that E[Xt] = E[Xk] and Var(Xt) = Var(Xk) for any t and k.

2.3.2.1 Uncorrelated versus Serially Dependent Data

One of the assumptions of ordinary least squares (OLS) regression is that the observations (and thus the residuals) are uncorrelated of one another. As you might imagine, and have seen in some of the examples thus far, this assumption is frequently violated when we record data over time (when the explanatory variable is time).

Take for instance the residual plots in Figure 2.16 a and b. This figure displays the residuals from a model with time as the explanatory variable where the solid horizontal line at zero represents the constant mean for all t (E[Xt] = for all t). Note also here that while we haven’t assigned it a value, the variance is assumed to be constant for all t: (Var(Xt) = ).

In fig 2.16a, the observations appear to be uncorrelated; that is, knowing a particular data point does not tell us any information about any other data point. Fig 2.16a shows observations x59 and x60 circled and note that they are both above zero (the mean). However, Fig 2.16b has x60 and x61 circled and illustrates that it is not unusual for observations one unit apart in time to be on opposite sides of the mean as well. In fact, close inspection of these plots will show that for a given point in time (xt), the next point (xt+1) is above or below the mean of 0 with approximately equal frequency. In other words, knowing the value of the observation “today” does not seem to tell us any information about where “tomorrow’s” observation will be; regardless of the value today (above or below the mean) there is equal probability of the next observation being above or below the mean tomorrow.

|  |  |
| --- | --- |
| a | b |
| c | d |
| Figure 2.16 a) and b) represent uncorrelated residuals. c) and d) represent residuals that are not uncorrelated. | |

As a heuristic exercise, consider Figure 2.16b and assume you have just observed the last observation: x100 = -0.6. Where do you think observation (x101) will be? What do you think of the chances of it being above or below the mean of 0?

It is reasonable here to believe, given the past values of the series, that there would be a 50%-50% chance of the next observation being above or below the mean of 0. This is consistent with uncorrelated data.

Alternatively, compare Fig 2.16 a and b with Fig 2.16 c and d. In the latter we see that values above the mean tend to be followed by values that are also above the mean (examples boxed in Figure 2.16c) and similarly with values below the mean (examples boxed in Figure 2.16 d).

This time consider Figure 2.16d and assume you have observed x100 = -13.5. Given what you have seen in observations to , where do you think will be?

It is reasonable here to believe, given the past values of the series, is likely to be less than 0. This is an example of strong (intuitive) evidence these data are correlated. In the discussion to follow, we will quantify this dependency / relationship using covariance and correlation similar to our discussion above, except in the time series setting, we will call these *autocovariance* and *autocorrelation*.

2.3.2.1 Autocovariance and Autocorrelation

Whereas the covariance quantified the linear relationship between two separate variables X and Y, we will begin by quantifying the linear relationship between observations that are one time unit apart: Xt and Xt-1. We will call this parameter the *autocovariance* at *lag 1,* where *lag* describes the distance between the observations*,* and will denote it as (pronounced “gamma sub 1”):

A popular estimate of this parameter is (pronounced “gamma hat sub 1”):

Note that the autocovariance at lag 1 is simply the covariance between observations that are one unit apart. As we will discuss in detail in the next chapter, we will assume that Xt and X1-1 have the same mean (mu) which is reflected in the above definition of , and estimated by in the statistic .

As we will see in the forthcoming example, we run into similar issues with interpreting the autocovariance as we did with the covariance and will thus introduce a standardized autocovariance which will call the *autocorrelation*. The autocorrelation at lag 1 is denoted (pronounced “rho sub 1”) and is defined as:

This parameter is commonly estimated by (pronounced “rho hat sub one”):

Note here that it is assumed that Xt and Xt-1 not only share the same mean , but the same standard deviation as well which is precisely why = in the above estimate. We will talk much more in depth about this assumption in the next chapter.

Consider again the uncorrelated data represented in Fig 2.16a and b. Figure 2.17a displays the first 10 observations of this series and Figure 2.17b displays a chart that displays these values listed as nine ordered pairs of the form: (xt-1, xt). Note that when k = 1 that we cannot pair the first observations since we do not have xt-1 when t = 1. Take a moment and inspect Figure 2.17b and note the fact that each pair consists of an observation (xt) and the observation one time unit before it (xt-1). Figure 2.17c shows the scatterplot of all 99 ordered pairs using all 100 data points. Note the scatterplot of all 99 ordered pairs (Figure 2.17c) shows that for low and high values of Xt-1 we observe low and high values of Xt. This is reflected in the autocovariance and autocorrelation that are to the right the plot in Figure 2.17c. Note that while it may be difficult to interpret the autocovariance, the autocorrelation is close to 0 as we would expect with uncorrelated data.

|  |  |  |
| --- | --- | --- |
| a. | b. | c. |
| Fig 2.17a -c | | |

Next, let’s look at the data we believe to be serially correlated which are displayed in Fig 2.16c and d. Fig 2.18a displays the first 10 observations while Figure 2.18b shows the corresponding first 9 ordered pairs of lag 1. (Note again, that when k = 1 that we cannot pair the first observations since we do not have xt-1 when t = 1.) In contrast to the uncorrelated case, this time, high values of Xt-1 tend to be followed by relatively high values of the next observation (Xt) and low values of Xt-1 tend to be followed by relatively low values of the next observation (Xt). This behavior will result in a positive autocovariance and autocorrelation which is illustrated in the scatterplot of the full series and its corresponding autocovariance and autocorrelation which can be viewed in Fig 2.18c. Again, the autocovariance ( ) is not as interpretable as the autocovariance (which is now much closer to its upper limit of 1.

|  |  |  |
| --- | --- | --- |
| a. | b. | c. |
| Fig 2.18a -c | | |

So far we have described the autocovariance and autocorrelation at lag 1 and are now in a position to generalize these parameters for any lag k:

whose estimators can naturally be generalized for any lag k as:

As an example, let’s now consider data in which Xt is correlated with Xt-2. Given our discussion above, think for a moment what this might imply about the autocorrelation and scatterplot of Xt and Xt-2. Fig. 2.19a shows the data from the uncorrelated data once again and 2.19b shows the first 8 pairs that are 2 units apart in time: (xt-2,xt). Note, analogous tot eh case when k = 1, when k = 2 we cannot pair the first two observations since we do not have xt-2 when t = 1 and 2. Figure 2.19c shows the scatter plot of the complete set of 98 coordinate pairs that can be formed from the original 100 data points. Note that, as expected, the sample autocorrelation at lag 2 for the uncorrelated data is close to zero since the data are uncorrelated at any lag. Contrast this with the corresponding data and plots in figure 2.19 d,e and f for the correlated data. Here we see that there is strong evidence of autocorrelation even at lag 2

|  |  |  |
| --- | --- | --- |
| a. | b. | c. |
| d. | e. | f. |
| 2.19 a - f | | |

We are now in a good position to generalize our discussion of autocovariance and autocorrelation to any lag k. For simplicity, we will restrict our discussion to autocorrelation although the results for autocovariance is analogous and left as an exercise. In review, we have noted that the sample autocorrelation for the white noise realization at lag 1 and 2 are and respectively; and that, for white noise, we expect the autocorrelation to be close to zero for any lag. We can check this by looking at the first 20 sample autocorrelations with the autocorrelation function or the “acf” (base R function acf()). Figure 2.20 shows the acf for lags 1 – 20 of the white noise realization while Figure 2.20b shows a table of the numerical values that correspond to the bars on the plot. (Take a moment and verify the first two autocorrelation on the plot and in the table!). Note that for lags 3 – 20, the autocorrelations are also relatively close to zero which is consistent with white noise.

|  |
| --- |
| a. |
| b. |
| Figure 2.20 a - b |

Let’s also verify the autocorrelations of the correlated realization by inspecting its acf in Figure 2.21 a. Recall that the first two sample autocorrelations were and which can be verified graphically in Fig. 2.21 a and numerically in Fig. 2.21 b. Note that for this series, the autocorrelations tend to damp slowly which is a characteristic we will study in depth in future chapters!

|  |
| --- |
| a. |
| b. |
| Figure 2.21 a - b |

***2.4 Conclusion / Summary***

In this chapter, we reviewed the fundamental statistical concepts of expected value, variance, covariance and correlation and their extensions with respect to data collected over time (time series). We began with a discussion and definition of expected value and variance and how, in the time series domain, we will be most concerned with calculating and evaluating these statistics for fixed values of time (conditional on time). We then extended the concepts of covariance and correlation by defining the *autocovariance* and *autocorrelation* *at lag k* to be the covariance and correlation between xt and xt-k. In the next chapter, we will use these constructs to define the concept of *stationarity* which is fundamental to a large class of widely used and extremely useful class of time series models.

**THE END as of August 19, 2020**