# Chapter 2
# Math Review

## 1 Notation and Math Review

In this class, we will most commonly be forming quantitative forecasts using time-series data on one variable collected at regular intervals. In order to make these forecasts, we will need to use mathematical formulas - these require some notation, which I introduce in the next subsection. In subsection after that, I will do some basic math/statistics review, including measure of central tendancy and spread, as well as summation notation.

### 1.1 Notation

I will do my best to remain consistent with the notation in this course. When we use time series data, the entire past history of the "object" or variable that we are interested in forecasting will be a vector, $Y$, which has the following properties:

- $Y$ holds observations over time.
- $t$ is any arbitrary time period.
- $t = 1$ represents the first time period in which data was observed.
- $t = T$ represents the most recent (i.e. last) period in which data was observed.
- Therefore, the vector $Y$ holds observations at times $t = \{1, 2, \cdots, T\}$
- $y_t$ refers to a specific observation in $Y$ at any arbitrary point in time, $t$.
- The first observation is $y_1$, and the last is $y_T$.

With all of this outlined, we can see that the vector, $Y$, containing all of our observed data can be written as:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}$$

### 1.1.1 Example

Suppose we have data on the US unemployment rate, collected monthly, for January, 1948 through March, 1948. The following table illustrates the notation above:

Table 1: Example of Time Series Data

| Period | $t$ | $Y$ | Data |
|---|---|---|---|
| January, 1948 | 1 | $y_1$ | 3.4% |
| February, 1948 | 2 | $y_2$ | 3.8% |
| March, 1948 | 3 | $y_3$ | 4.0% |

Another way to write this data would be:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 3.4\% \\ 3.8\% \\ 4.0\% \end{bmatrix}$$

In either case, we see that since there are three observations, $T = 3$. We also see that the first observation, $y_1 = 3.4\%$, the second observation, $y_2 = 3.8\%$, and the third (last) observation, $y_3 = 4.0\%$.

### 1.1.2 Forecasting Notation

The notation above covers how we will denote in-sample data. However, when we forecast we will need a little more notation. Typically, we will be interested in a specific **forecast horizon** — the number of periods in the future that we are interested in forecasting. We will use $h$ to denote the forecasting horizon. For example, if $h = 1$ and we are using monthly data, this means that we want to forecast one month ahead. If $h = 1$ and we have weekly data, this means that we are interested in forecasting one week ahead.

We will also draw a distinction between the forecast horizon, which is just one period, and a series of forecasts, which are forecasts for each period up to the forecast horizon. If we want a series of forecasts, we will instead use a capital $H$.

For example, if we wanted to form the two-period ahead forecast, $h = 2$, we would only come up with a forecast for two periods ahead, and we would NOT form or pay attention to the one-period ahead forecast. However, if we were interested in forecasting *through* two-periods ahead, we would write $H = 2$ and we would form the forecasts for both $h = 1$ (one-period ahead) and $h = 2$ (two-periods ahead).

Point Forecasts Finally, when we forecast in this class we will be most interested in the **point forecast** — a single number that summarizes our best guess about the variable $h$-periods ahead. We will denote the $h$-period ahead forecast as $\hat{y}_{T+h|T}$ where the "hat" on top of $y$ denotes that this is not real, actual, collected data, but is rather the forecast from a quantitative model. We write $T + h|T$ to denote that this forecast is for period $T + h$, and was formed using data up to and including time period $T$.

Interval Forecasts While we will not discuss them as extensively as point forecasts, the computer software we use will produce *interval forecasts* (sometimes called "forecast intervals" or "prediction intervals"). These intervals help calrify how much uncertainty we have about our forecast. For example, if we use a 50% forecast interval, the actual data should have a 50/50 chance of lying within that interval. If we used a 95% interval, the actual data should lie in that interval with 95% probability. Mathematically, we will denote an interval forecast for period $T + h$ by:

$$[\hat{y}^L_{T+h|T}, \hat{y}^U_{T+h|T}]$$

where the $L$ in $\hat{y}^L_{T+h|T}$ stands for the "lower" bound on the prediction interval and the $U$ in $\hat{y}^U_{T+h|T}$ stands for the "upper" bound on the prediction interval.

## 1.2 Math Review

Now that we have our basic notation down, let's turn to our math review.

### 1.2.1 Summation Notation

In many of our formulas, we will often encounter **summation notation**. The summation operator is denoted by $\Sigma$ (pronounced "sigma"). This is used as a compact way to denote that we want to add all the elements of our vector together. If our time series is relatively short, like it was in the last example, it's not too big of a deal to just write, for example, $y_1 + y_2 + y_3$. However, if our time series is longer (say, 250 observations), it would certainly waste a lot of our time if we were to write out the addition with 250 terms.

When we use the summation operator, it typically looks something like this:

$$\sum_{t=1}^{T} y_t$$

The first element used in the addition is denoted by what's beneath $\Sigma$. In the example above, the term beneath the $\Sigma$ is $t = 1$. This means that we are going to vary over the variable $t$, and we will start with $t = 1$. The number above the $\Sigma$ tell us what the last term in the addition will be. In the example above, the term above $\Sigma$ is $T$. This means that we will stop adding once we get to $t = T$.

Let's consider an example with numbers. Let's continue with the unemployment example from the previous section, in which we have US unemployment data for three months in 1948. We know that $T = 3$ since we have three observations. Therefore we would have:

$$\sum_{t=1}^{T} y_t = \sum_{t=1}^{3} y_t$$

Again, this means that we will first plug in "1" for $t$, then "2", and we will continue until we hit $t = 3$, at which point we will stop. Therefore,

$$\sum_{t=1}^{3} y_t = y_1 + y_2 + y_3$$

Plugging in the actual unemployment rates, we would have:

$$\sum_{t=1}^{3} y_t = y_1 + y_2 + y_3$$
$$\sum_{t=1}^{3} y_t = 3.4\% + 3.8\% + 4.0\%$$
$$\sum_{t=1}^{3} y_t = 11.2\%$$

### 1.2.2 Measures of Cenral Tendency

With summation notation in hand, let's turn to a review of some basic statistics. After we collect data and we have a data set, it is often useful to calculate some simple measures. For example, we might have unemployment rate data back to 1948, and it would be interesting to know what the most "typical" unemplyment rate has been over the past 70 years. We will consider three measures of central tendency that you have almost certainly seen before: mean, median, and mode.

The mean (i.e. the arithmetic mean) is simply the average of the past data. This is simply the sum of all observations divided by the number of observations, $T$:

$$\text{mean} \ = \frac{y_1 + y_2 + \cdots + y_T}{T}$$

We can rewrite this using summation notation as:

$$\text{mean} \ = \frac{\sum_{t=1}^{T} y_t}{T}$$

or:

$$\text{mean} \ = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The median is the middle-most observation (ie. the 50th percentile). In other words, if you sorted the data from the smallest number to the largest number (instead of sorting it by time observed), the median would be the middle-most number. If the total number of observations, $T$, is odd, then the median is simply the middle-most number. If the total number of observations,

$T$, is even, then the median is the average of the two observations closest to the middle.

Finally, the mode is the observation that occurs most frequenly in the data set. If all numbers in the data set only occur one time (so that there is no clear number that occurs with the greatest frequency), the mode is not well defined. We could either say every observation is the mode, or the mode does not exist.

### 1.2.3 Measures of Spread

While measures of central tendency tell us what a "typical" value of our variable is, it doesn't tell us anything about how spread out the values tend to be. Even if two separate variables have the same mean as each other (say, mean $= 0$ for both vairables), if one varies between $-1,000,000$ and $1,000,000$, while the other varies between $-0.01$ and $0.01$, the two variables are much different. Therefore it is common to consider measures of spread. We will consider three measures, all of which you have probably heard of before: Range, Inter-Quartile Range (IQR), and Standard Deviation.

The range is computed as the largest observation minus the smallest observation. Note that it is not represented as an interval (-1,1), but instead as a single number: $1 - (-1) = 2$. This gives us a sense of the largest observed spread of the distribution of a variable.

The inter-quartile range (IQR) is computed as the value at the $75^{\text{th}}$ percentile of the distribution minus the value at the $25^{\text{th}}$ percentile of the distribution. Again, like the range this is not an interval, but a single number. Since it only considers the spread of the middle 50% of the distribution, it is less sensitive to outliers then the range. For example, if 999 of 1,000 observations are between -1 and 1, but there is one outlier where the variable is equal to 10, the IQR might provide a better sense of the typical spread of the data.

Finally, the standard deviation is another measure of spread it is computed as:

$$s = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

where $\bar{y}$ is the mean of the series. Like the range and IQR, the standard deviation will always be a positive number, with smaller values indicating less spread. One way to think about standard deviation is that if you were to repeatedly pick observations completely at random, on average they will tend to be $s$ units from the mean.

## 1.3 Hypothesis Testing

We will come across hypothesis testing in this course. You should have covered the basics of hypothesis testing in STAT 220, so I won't cover them in great detail here. The basics are as follows.

Suppose you have some hypothesis, for instance, you believe that the mean of a random variable, $Y$, might be exactly equal to zero. However, due to the limitations of finite samples, it is almost surely the case that the mean of $Y$ will NOT be exactly equal to zero. So you conduct a hypothesis test.

In any hypothesis test, our hypothesis is called the "null hypothesis" ($H_0$). The alternative is called the "alternative hypothesis" ($H_1$). Then, we assume that the data was generated under the null hypothesis, and generate the probability (*p-value*) that the observed data came from a process under which the null hypothesis was true. Finally, in order to determine whether or not we can reject the null hypothesis, we use what's called a *critical value*, $\alpha$. A critical value is a cut-off signifying the power of your test. Typically, we set $\alpha = 0.05 = 5\%$, so we only reject the null hypothesis if there is less than a 5% chance that the data we are seeing could have been generated according to a process in which the null hypothesis was true.

Based on the evidence you have gathered, there will be two possibilities:

1. You do not have enough evidence to reject the null hypothesis.

2. You have enough evidence to reject the null hypothesis.

Note that you can never confirm that your hypothesis is true, only that you can either reject it or fail to reject it.

To help illustrate this point, let's continue with the example laid out above, where we believe that the mean of random variable, $Y$, is exactly equal to zero. In this example, the null hypothesis is that the mean of $Y$ is zero, and the alternative is that it is different than zero:

$$H_0: \ \text{mean}(Y) = 0$$
$$H_1: \ \text{mean}(Y) \neq 0$$

Then, we assume that the data was generated from a process that actually had a mean of zero. Based on that assumption, we compute the probability of actually getting a sample of data that looks like $Y$. Maybe that probaility is 0.01=1%, so it is unlikely that the true mean of the data is zeror. Since 1% is less than our critival value of 5%, we **reject** the null hypothesis in favor of the alternative.

## 1.4 Expected Value

In this course, we will use the **expected value** operator fairly extensively. Recall from STAT 220 that the expected value of a random variable is an average over all possible outcomes, with each possible outcome weighted by its probability. In the past, you have probably used the expected value mathematically. We will typically use it more theoretically.

Let's start with a mathematical example. Suppose we were going to play a game together. You will flip a (fair) coin. If it comes up Heads, I will give you a dollar. If it comes up tails, you will give me a dollar. From your point of view, we could then define the random variable, $X$ as: where $P(X)$ is the probability of each event - in this case, with a fair coin, both outcomes have a 50% probability of occuring.

| Outcome | $X$ | $P(X)$ |
|---|---|---|
| Heads | 1 | 0.5 |
| Tails | -1 | 0.5 |

Now we want to find the expected value of this bet. To do so, we can multiply the value of the random variable under each outcome by the probability of that outcome occurring. In general, if there are $n$ total possible outcomes, then:

$$E(X) = p_1 X_1 + p_2 X_2 + \cdots + p_n X_n$$
$$E(X) = \sum_{i=1}^{n} p_n X_n$$

In our specific example:

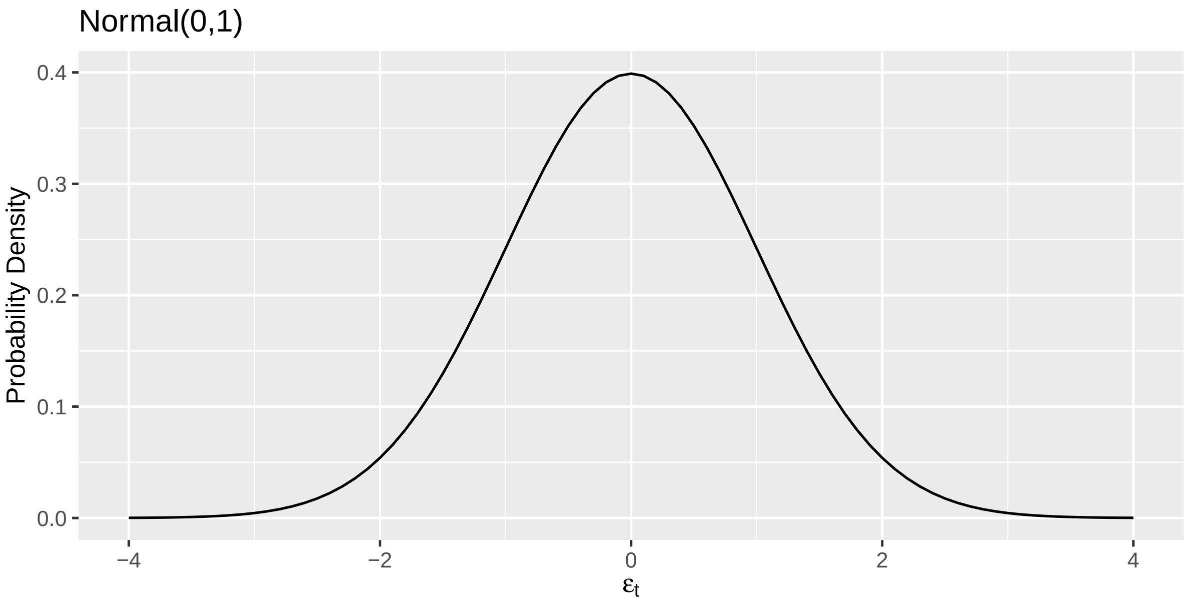
$$
\begin{aligned}
E(X) &= \sum_{i=1}^{n} p_n X_n \\
E(X) &= \sum_{i=1}^{2} p_n X_n \\
E(X) &= p_1 X_1 + p_2 X_2 \\
E(X) &= 0.5(1) + 0.5(-1) \\
E(X) &= 0.5 - 0.5 \\
E(X) &= 0
\end{aligned}
$$

So the expected value of our bet is \$0. In the long run, if you played this game a large number of times, you should expect to come away with exactly as much money as you started with.

In this course, we will often use the expected value operator a little differently. We will usually be dealing with continuous random variables, so $n \to \infty$. Clearly we can't just multiply the outcomes by their associated probabilities! Instead, we will work in more abstract terms. For example, suppose that we know a random variable, $Y$, is generated according to the following process:

$$
\begin{aligned}
y_t &= 5.0 + \varepsilon_t \\
\varepsilon_t &\sim N(0, \sigma)
\end{aligned}
$$

In words, this is saying that the variable $Y$ at time $t$ is equal to 5 plus some random shock (i.e. error). This error term comes from a Normal distribution with mean zero and standard deviation $\sigma$. For example, if $\sigma = 1$, then the distribution of all possible error terms would look like this:



But suppose we wanted to find the expected value of $Y$ at time $t$, $y_t$. How could we do that? Let's use the expected value operator to take the expected value of both sides of the equation:

$$y_t = 5.0 + \varepsilon_t$$
$$E(y_t) = E(5.0 + \varepsilon_t)$$

A nice property about the expected value operator is that it is additively seperatble, so that in general:

$$E(a + b) = E(a) + E(b)$$
$$\text{and}$$
$$E(a - b) = E(a) - E(b)$$

Therefore, we can separate the constant (5.0) from the error term ($\varepsilon_t$):

$$E(y_t) = E(5.0 + \varepsilon_t)$$
$$E(y_t) = E(5.0) + E(\varepsilon_t)$$

We know that $\varepsilon_t$ has mean 0 - when we set the problem up we said that $\varepsilon_t$ was normally distributed with **mean 0**. Therefore, we know the expected value (i.e. the mean) of $\varepsilon_t$ is zero:

$$
\begin{aligned}
E(y_t) &= E(5.0) + E(\varepsilon_t) \\
E(y_t) &= E(5.0) + 0 \\
E(y_t) &= E(5.0)
\end{aligned}
$$

Finally, the expected value of a constant is just that constant. For example, if you know that you have exactly one dollar in your pocket, then the expected value of the amount of money in your pocket is \$1. It's like taking the average of just one number. Therefore:

$$
\begin{aligned}
E(y_t) &= E(5.0) \\
E(y_t) &= 5.0
\end{aligned}
$$

### 1.4.1 Properties of Expected Value Operator

There will be a few properties of the expected value operator that we will use again and again:

1. Additively Seperable: If we have two random variables, $X$ and $Y$, then:

$$
\begin{gathered}
E(X+Y) = E(X) + E(Y) \\
\text{and} \\
E(X-Y) = E(X) - E(Y)
\end{gathered}
$$

2. NOT Multiplicatively Seperable: If we have two random variables, $X$ and $Y$, then in general:

$$
\begin{gathered}
E(XY) \neq E(X)E(Y) \\
\text{and} \\
E\left(\frac{X}{Y}\right) \neq \frac{E(X)}{E(Y)}
\end{gathered}
$$

3. Additively and Multiplicatively Seperable with Constants: If we have a constant, $a$, whose

value we know with certainty, and a random variable $Y$, then:

$$E(a+Y) = a + E(Y)$$
$$E(a-Y) = a - E(Y)$$

and

$$E(aY) = aE(Y)$$
$$E\left(\frac{Y}{a}\right) = \frac{E(Y)}{a}$$

### 1.4.2 Conditional Expected Value and Information Sets

In the previous subsection, we covered *unconditional expected value*. However, there are many situtations in which we have relevant information that will change the expected value. For example, consider the average high termperatures in Minneapolis: Let the random variable $Y$ be the average high temperature in Minneapolis. Then, the expected value of $Y$ would be roughly:

| Month | Avg. High Temp. |
|---|---|
| January | 23.7 |
| February | 28.9 |
| March | 41.3 |
| April | 57.8 |
| May | 69.4 |
| June | 78.8 |
| July | 83.4 |
| August | 80.5 |
| September | 71.7 |
| October | 58.0 |
| November | 41.2 |
| December | 27.1 |

$$E(Y) = \frac{1}{T}\sum_{t=1}^{12} y_t$$
$$E(Y) = \frac{1}{12}(23.7 + 28.9 + 41.3 + 57.8 + 69.4 + 78.8 + 83.4 + 80.5 + 71.7 + 58.0 + 41.2 + 27.1)$$
$$E(Y) = \frac{1}{12}(661.8)$$
$$E(Y) = 55.15$$

In other words, if we were to select a day completely at random, our best guess of the high temperature would be roughly 55 degrees. However, we will typically have more information that we can use to get a more precise expected value. For example, suppose you had a family member who was thinking of visiting Minneapolis next July. You could tell your family member that on the average day in Minneapolis, the temperature is 55 degrees. But in this case, we have more useful information, since there is predictable seasonal variation in the temperature. Instead, you should use the conditional expected value:

$$E(Y|\text{month} = \text{July}) = 83.4$$

We read $E(Y|\text{month} = \text{July})$ as "the expected value of $Y$ given that the month is July". Even if the trip is planned months in advance, this will most likely be a better estimate than the unconditional expected value. If we have more then one piece of information (perhaps we know that the trip will be in July and that next July is projected to be warmer than average), we collect all the infomation in an **information set**, $\Omega$ ("omega"). Then, instead of writing:

$$E(Y|\text{month} = \text{July AND warm})$$

we could simply write:

$$E(Y|\Omega)$$

where $\Omega$ contains all the information that we have.

Finally, note that all the properties we discussed in the previous section are applicable to conditional expected values as well. For example, conditional expected values are additive. If we have random variables $X$ and $Y$, then:

$$E(X + Y|\Omega) \equiv E(X|\Omega) + E(Y|\Omega)$$

For our purposes in this course, the information set will typically be all of the past historical data. However, when we discuss linear regression, our information set can expand to include past historical data on the current series, $Y$, as well as past historical data of other related series: $X_1, X_2, \cdots, X_k$.

## 2 Exercises

1. Use the following data set to answer the questions below.

Table 2: Average Nightly Hours of Sleep

| Period | Data |
|---|---|
| May, 2017 | 7.5 |
| June, 2017 | 8.5 |
| July, 2017 | 8.5 |
| August, 2017 | 8.0 |
| September, 2017 | 6.5 |

(a) What type of data is this?

(b) What is the total number of observations, $T$?

(c) Which period corresponds to $t = 2$?

(d) What is $y_4$?

(e) What would be the notation for the forecast for October, 2017? (Hint: I want you to plug in values for $T$ and $h$ into the generic forecast notation: $\hat{y}_{T+h|T}$)

(f) What is the mean of the data?

(g) What is the median of the data?

(h) What is the mode of the data?

(i) What is the standard deviation of the data?

(j) What is the range of the data?

(k) What is the inter-quartile range of the data (Hint: You can find this using RStudio...if we haven't used RStudio yet in class, hold off on this.)

2. Suppose you hold the following beliefs about the number of wins the Vikings will have this season:

Table 3: Number of Wins of MN Vikings

| Wins | Probability |
|---|---|
| 8 | 0.10 |
| 9 | 0.30 |
| 10 | 0.35 |
| 11 | 0.15 |
| 12 | 0.10 |

   (a) What is the expected value of wins by the Vikings?

3. Explain the following notation in words. Be as specific as possible.

$$E(Y|\Omega)$$

4. Why is the expected value operator so important?

5. Suppose $X$ and $Y$ are random variables and $a$ and $b$ are known constants. Simplify the following as much as you can:

   (a) $E(a)$

   (b) $E(a+b)$

   (c) $E(Y+X)$

   (d) $E(YX)$

   (e) $E(aY+bX)$

   (f) $E(ab+YX)$

   (g) $E(Y|X)$

   (h) $E(a+Y|X)$

   (i) $E(a+bY|X)$