Contents

Probability and Statistics

Brynnydd Hamilton

August 17, 2022

1 Introduction/Philosophy

Many people have taken classes in probability and statistics, but I always found those terms very vague. What are probability and statistics, and what's the difference between them?

Probability theory is what we use to understand the **likelihood of random events**. It tells us how we can manipulate probabilities of different events. As applied scientists (i.e., not mathematicians), we don't *really* think about probability theory that much, we're more concerned about describing the characteristics of our dataset through statistics.

However, having a good understanding of probability theory is essential to understanding statistics, since all statistics are just applications of probability theory. If statistics is painting, probability is perspective and color theory. It's the underlying mechanics of statistics. You can kind of think of probability as more cut-and-dry (there is usually a "right" answer, as opposed to in statistics, where there are better and worse ways to do things, but no "right" answers).

I really like that metaphor, because, much like many people can look at the same scene and paint different pictures of it, there are infinite ways to use statistics in your research to analyze your dataset. Every choice that you make in experiment setup, data processing, and analytical techniques will have an impact on your resultant statistics. That can be overwhelming, but it's a key thing to keep in your head at all times. And it's also important to keep in mind that statistics is always a simplification, so you will always be "losing" information. However, we have to lose some information to come up with an understandable system.

"Whenever you're dealing with a model or with a situation, there are zillions of details in that situation. And when you come up with a model, you choose some of those details that you keep in your model, and some that you say, well, these are irrelevant" - John Tsitsiklis

2 Probability

2.1 How to Talk about Probability

When people want to talk about probability theory, they will often work with random variables denoted by a capital letter, commonly an X or Y. Technically speaking, a random variable is something called a *measurable function*, which is a function between two sets in two measurable spaces, one of which is a probability space, that preserves the structure of the space. It's important to note that these can be discrete (our output is a finite set) or continuous (our output is a number anywhere within some range, an infinite set).

Oftentimes there can be many ways to describe a sample space, for the same process (e.g. flipping a coin: one sample space is heads, tails, another sample space is heads, tails and it's raining, tails and it's not raining). Note that the way we define a sample space will guide our interpretation of the problem. Therefore, it's important to think about having the right "granularity" in your definition of your sample space for a specific problem.

We refer to a **event** as some subset of the sample space.

We assign probability to our events, with a probability between 0 (not gonna happen!) and 1 (absolutely gonna happen!). We will use the notation that $\mathbf{P}(A)$ is the probability that event A occurs.

Some axioms of probability are...

1. nonnegativity
2. normalization (probability of entire sample space = 1)
3. additivity

   If two events do not intersect

$$A \cap B = 0 \tag{1}$$

   Then the union of the two events is equal to the sum of their independent probabilites

$$P(A \cup B) = P(A) + P(B) \tag{2}$$

When defining a sample space, we want the events of the sample space to have the following characteristics

1. mutually exclusive
2. collectively exhaustive

For a sample space, we will define a "probability law", which tells us how probability is distributed in the sample space. We can use probability laws to analyze a random variable. These can come from physics or data.

If you take a probability class, you'll spend hours playing around with these statistics. I'm going to focus on some applications of conditional probability, because I think that's one of the more relevant avenues of probability theory. Conditional probability is how we incorporate our prior knowledge of a situation into our estimate of probability.

2.2 Conditional Probability

We define the notation $\mathbf{P}(A|B)$ as the probability of A, given that B occurred.

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \tag{3}$$

A consequence of this definition is what we call the **Joint Probability**

$$\text{Joint Probability: } \mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A|B) \tag{4}$$

and because an intersection is order-invariant, we can also say

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B|A) \tag{5}$$

2.2.1 Independence

One concept that impacts a lot of calculations is **independence**, which tells us if two probability distributions are related to one another.

If an event A is **independent** from event B, then $\mathbf{P}(A|B) = \mathbf{P}(A)$. For our earlier example where we defined our sample space as follows

$$S = \begin{cases} \text{heads} \\ \text{tails and its raining} \\ \text{tails and its not raining} \end{cases} \tag{6}$$

we would find that $\mathbf{P}(\text{tails}|\text{raining}) = \mathbf{P}(\text{tails})$, indicating that the probability of rain is independent from the probability of our coin landing on tails.

2.2.2 Example: False Alarms

See the attached .ipynb with a Julia example of this situation

Imagine we have a piece of equipment that detects whales. We know that for our region of study, there are whales there 5% of the time.

$$\mathbf{P} = \begin{cases} 0.95 & \text{no whale} \\ 0.05 & \text{whale} \end{cases} \tag{7}$$

Furthermore, our equipment will detect the whale 99% of the time **if there is a whale there** (if we miss a whale: Type II error, false negative). It will detect a whale 10% of the time **if there is no whale there** (if we detect a whale: Type I error, false positive).

If we get a detection signal, what's the probability that it's actually a whale? Hint: calculate

$$\mathbf{P}(\text{whale}|\text{detection}) = \frac{\mathbf{P}(\text{whale} \cap \text{detection})}{\mathbf{P}(\text{detection})} \tag{8}$$

2.2.3 Bayes Theorem/Law/Rule

This example helps show the power of what's called Bayes Theorem/Law/Rule (depending on who you ask - and there's a lot of strong opinions about it!). We can use it to incorporate prior information into our analysis. Specifically, Bayes' theorem is used to incorporate prior knowledge about the situation into the estimate of the probability of an event, A_i. A formal definition of it is given in Equation 9.

$$\text{Bayes' Theorem: } \mathbf{P}(A_i|B) == \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(B)} \tag{9}$$

An intuitive way to think about this theorem is that it takes the probability of both events occurring, and divides it by the normalizing factor. There is a whole field of analysis surrounding this fundamental idea called **Bayesian Analysis**. If you want to see a good example of using Bayesian analysis, check out this Github repo from Chris Piecuch.

3 Statistics

3.1 Probability Density Functions

Statistics is what we we use whenever we want to think about analyzing and making decisions about a dataset. In statistics, we'll be working with **probability density functions** (PDFs) and **cumulative distribution functions.**. Although we think about PDFS more often, it can be easier to think about the CDF first.

We will define a CDF with a capital P, and it will be defined for the random variable x $P_x()$. CDFs are defined such that $P_x(X)$ is the probability that $x \leq X$. At $-\infty$, the CDF is 0, and at $+\infty$, the CDF is 1.

We will define a PDF with a lowercase p and will be defined similarly to the CDF, $p_x()$. The PDF is the derivative of the CDF and is normalized so that its integral over all X is 1.

An example of the PDF and CDF are shown in Figure 1 for the Normal distribution.

3.2 Statistics for Understanding Distributions

3.2.1 Ways to Describe One Distribution

Two terms we'll use all the time to talk about distributions are the **mean** and **variance**.

The mean is also referred to as the first moment, average, or expectation value. One way to think about it is that it is the "center of mass" of the distribution.

The variance helps us define the "spread" of the distribution, or how far we vary away from the mean, on average.

In Table 1, I show the mathematical definitions for mean and variance, for continuous and discrete cases. The continuous case is what we use when we talk about functions defined for all real numbers. However,

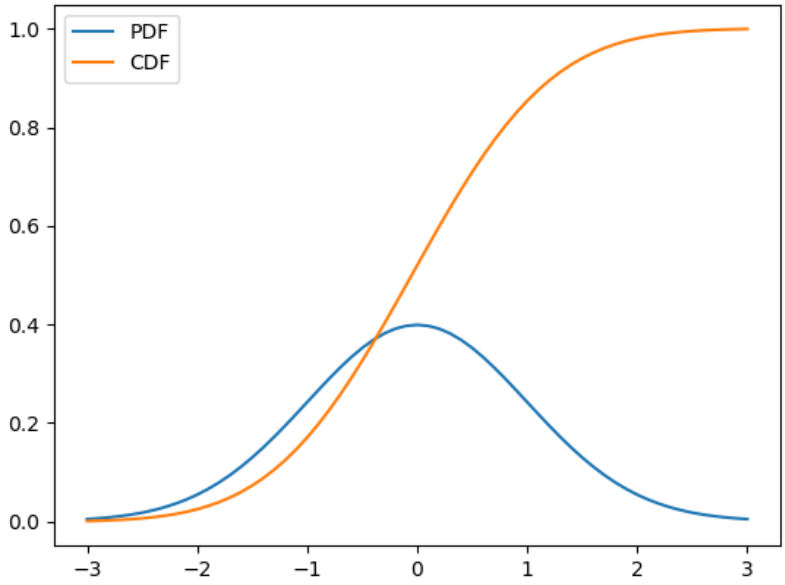

Figure 1: The normal distribution as a PDF and CDF.

	Continuous	Discrete
Mean $< x >, \mu$	$\int_{\infty}^{\infty} X p_x(X) dX$	$\frac{\sum_i^N x_i}{N}$
Variance $< (< x - < x >)^2 >$	$\int_{-\infty}^{\infty} (X - < x >)^2 p_x(X) dX$	$\frac{\sum_i^N (x_i - \bar{x})^2}{N}$

Table 1: The continuous and discrete definitions of the mean and variance.

when we are dealing with real data, we will usually use the discrete definition. Note that the definition of variance will change if we are talking about "sample" or "population" variances.

We also will frequently talk about **standard deviation, which is the square root of variance**, and another measure of the spread of a distribution.

One further question we can ask is "how does the variance vary?" This is equivalent to asking about the asymmetry of the the distribution or the **skewness**. The skewness comes from the idea of standardized moments (it's the 3rd standardized moment), and is defined as such

$$\text{Skewness: } < \left(\frac{X - \mu_X X}{\sigma_X X} \right)^3 > \tag{10}$$

You can take this one step further, and look at the 4th standardized moment (i.e., raise the quantity in Equation 10 to 4), we will have a measure of the "tailedness" of the distribution, called **kurtosis**.

Finally, you may recall talking about quartiles and quantiles in your previous statistic classes. A quantile is the more general definition, and it can be considered the cut point of a probability distribution into continuous intervals with the same probability. We can have a quantile for any percentage. The 25% quantile is the first quartile, and it is the point where 25% of the data lies "below". The 50% quantile is the median, and 75% quartile is the third quantiles. The space between the first and third quartiles is referred to as the interquartile range.

It is important to be cognizant of how extreme values (sometimes outliers) will affect these statistics. For a statistic like "range" (the difference between the maximum and minimum values of the dataset), it will be heavily impacted by extreme values. Mean and variance will be more impacted by extreme values than the

median and interquartile range.

For discrete sets, we can also look at the following statistics

- median: for the ordered dataset, the middle value (similar to mean, less impacted by extreme values)
- mode: the most common value (not really used a lot, tells us something about frequency)

3.2.2 Ways to Talk About Two Distributions

If we are dealing with two distributions, we can ask if they are related to each other by looking at the correlation (how they are related) or the covariance (how they differ). I show these definitions in Equations 11 and 12, where μ_Z denotes the mean of random variable Z and σ_Z denotes the standard deviation, and $< Z >$ represents taking the mean of whatever is between the brackets.

$$\text{Covariance: } < (X - \mu_X)(Y - \mu_Y) > \tag{11}$$

$$\text{Correlation: } \frac{< (X - \mu_X)(Y - \mu_Y) >}{\sigma_X \sigma_Y} \tag{12}$$

As we see in the above equations, correlation is a scaled covariance. Sometimes you will hear it referred to as a Pearson correlation coefficient, of which there are a few mathematical definitions. It is also sometimes referred to as an r-value. Pearson correlation is most sensitive to linear relationships, and there are other measures of correlation that are more sensitive to nonlinear relationships. As correlations are not scaled, you should not compare correlation coefficients between datasets.

Covariance, which is correlation scaled by standard deviation, ranges from -1 to +1, with -1 indicating perfect anticorrelation, 0 indicating independence, and +1 indicating perfect correlation. Covariance can be compared between datasets.

3.3 Distribution Statistic Example: Average Daily Air Temperatures in Florida

Now, we will do an example applying these statistical measures we've defined. We will look at daily average air temperatures from three different locations in Florida: St Petersburg, Tampa Bay, and Fort Myers. In Figure 2, I plot the three locations of the datasets. A common, and frequently overlooked, part of data analysis is to think about the physics (or what you already know) of the problem before you begin. What do you expect to see?

Some things I would expect to see is the coastal effect (cooler temperatures near the coast) as well as the urban effect (warmer temperatures in urban areas). What would you expect to see in terms of variability? Extreme values?

The next, and still sometimes overlooked, step in data analysis is to make some exploratory graphs of your data to get a feel for its form. Since this is time-series data, I plot it with respect to day. You should be able to get an idea for the mean, variance, extreme values, and correlation from this plot. I also plot a histogram to get a better sense of the distribution (mean value, variance, extreme values) with all temporal aspects removed.

I think two great tenets of data analysis are

1. Have an idea of what you'd expect every time you **calculate** or **visualize** anything
2. Plot everything, in multiple ways.

In the accompanying Jupyter Notebook, I walk through several ways to visualize this dataset. I think it's really important, whenever you're dealing with multiple datasets, to visualize everything together, as well as on its own axes. This really helps to highlights comparative features (i.e., "this mean is higher than that mean") as well as stand-alone features (i.e. "this dataset has some really large extreme values in comparison to the rest of it").

I also walk through calculating all the statistics we previously discuss, and come up with an interpretation of this dataset.

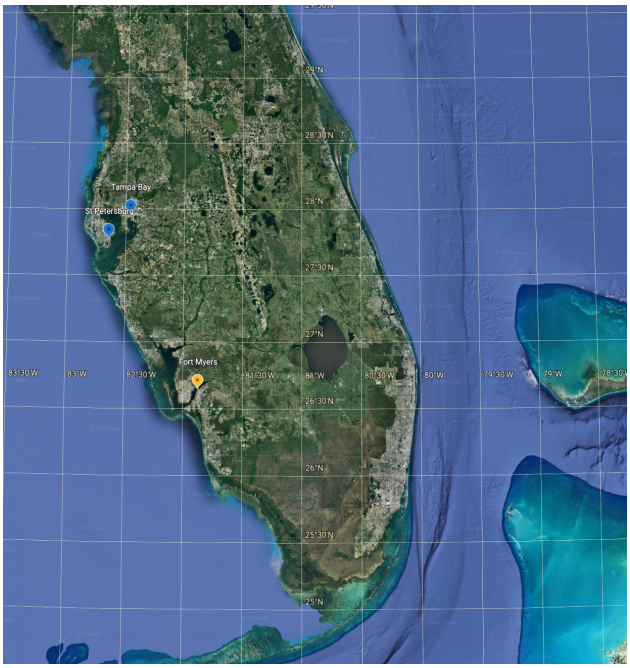

Figure 2: The locations of our three datasets.

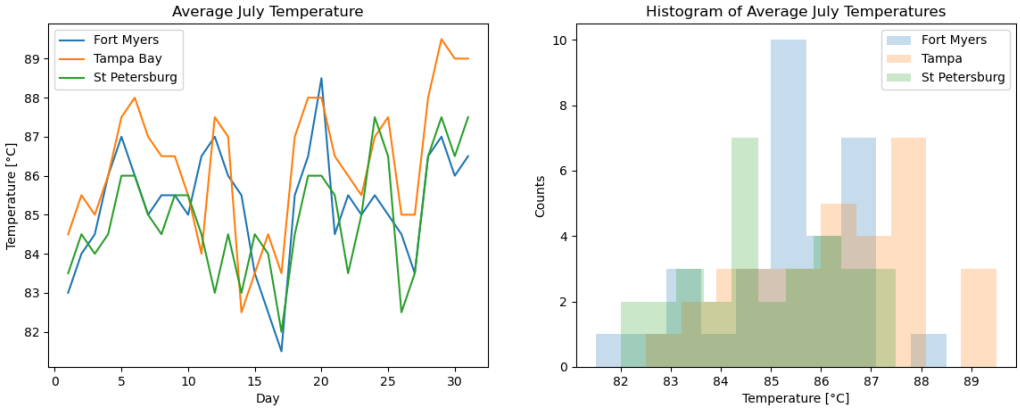

Figure 3: On left, data plotted as timeseries. On right, data plotted as a histogram.

3.4 Normal Distribution and the Central Limit Theorem

"Normally distributed" is a phrase that you'll hear a lot in statistics. The Normal Distribution, also referred to as the Gaussian (pronounced "gau-see-an") Distribution (or the bell-curve) refers to a distribution that takes the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{13}$$

Normal distributions are important because of two interrelated reasons

1. Central Limit Theorem
2. Many statistical techniques apply only to normally distributed variables

What's the Central Limit Theorem? The Central Limit Theorem states that, for any random variable, with any distribution, the fluctuations between it's sample mean and population mean will become normally distributed, when scaled by the number of samples. It can be a little difficult to grasp the magnitude of this theorem, so I present it, step by step here.

First, let's assume we have some weird distribution, like a gamma distribution. This distribution is always positive, has its highest value at 0, and trails off from there. Definitely not a normal distribution! I show this in Figure 4

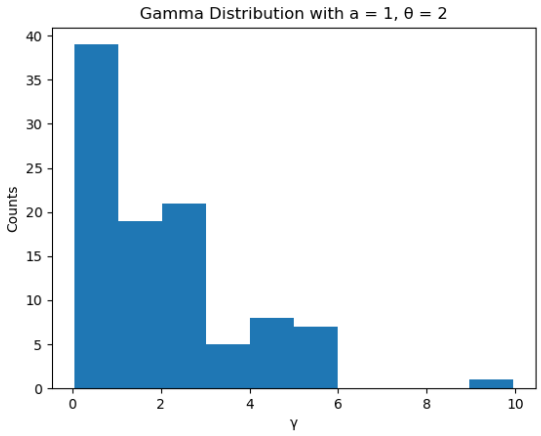

Figure 4: Histogram of 100 points taken from Gamma distribution

The law of large numbers says that, as we take progressively larger samples, the mean of any distribution will probably converge. For posterity, I demonstrate this in Figure 5, where I take progressively larger samples from our gamma distribution and calculate the mean. In Figure 5, we see that there is a lot of

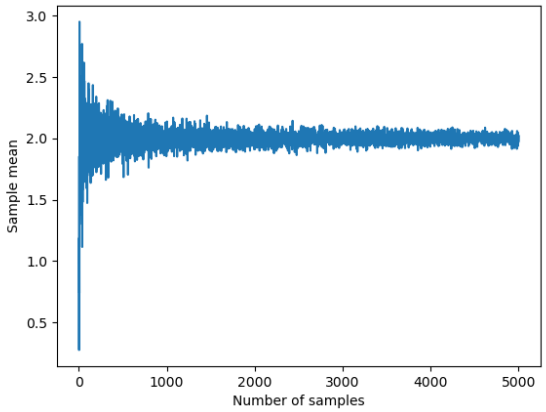

Figure 5: Demonstration of law of large numbers

oscillation for smaller sample sizes (< 1000), and then it seems like our sample mean settles around 2 (but there is still some noise!).

We'll take the mean for some very large sample size and call this our population mean. Next, for progressively larger number of sample sizes, we'll calculate the following scaled fluctuation metric

$$f(n) = \sqrt(n)(\mu_{x_n} - \mu_x) \tag{14}$$

where n is the number of samples, μ_{x_n} is the mean of a n samples taken from our random distribution, and μ_x is our population mean (the horizontal asymptote of Figure 5.

In Figure 6, I show the result of taking histograms of the distribution of values calculated by taking progressively larger samples. We see that it approaches the normal distribution!

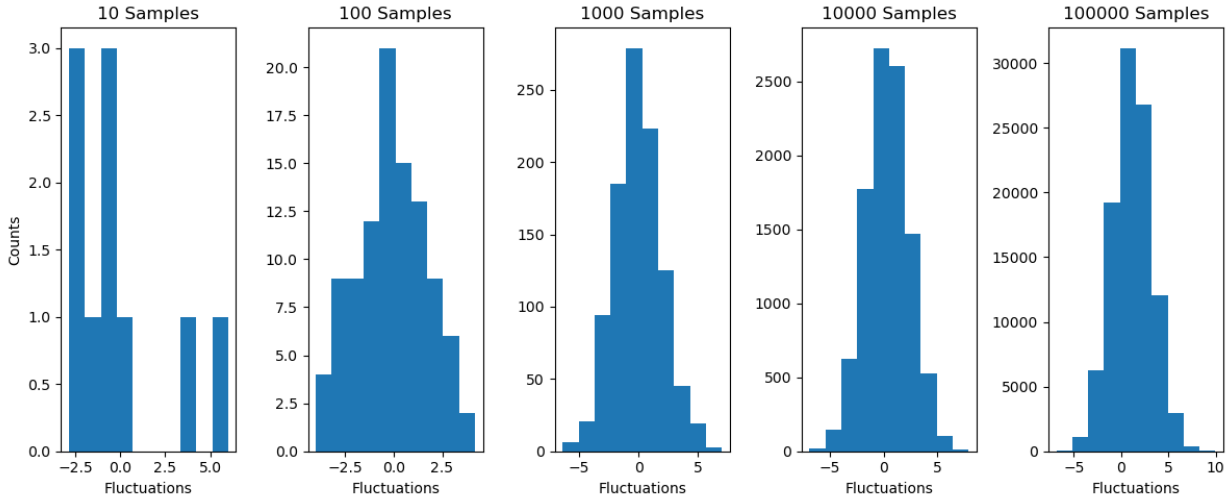

Figure 6

3.5 Hypothesis Testing

An important question in statistics is whether your results are significant. The boundary for significance is very field-dependent, but oftentimes people will use the idea of the null hypothesis and a t-test to prove significance. Here is the protocol for this

1. Formulate a **null hypothesis** and a **alternative hypothesis**. Usually your null hypothesis is that your data has no significance, and your alternative hypothesis is that it does. For my temperature example, I might make the following hypotheses.

   Null Hypothesis: Average temperature in Tampa Bay is less than **or equal to** 85°C in June.

   Alternative Hypothesis: Average temperature in Tampa Bay is greater than 85°C in June.

   Note: your null hypothesis should always have the "equals" because it suggests that this is the most conservative estimate that your null hypothesis could satisfy.

2. Determine the appropriate test statistic. This depends on your alternative hypothesis H_a

   - if $H_a : \mu > \mu_0$: upper-tailed test
   - if $H_a : \mu < \mu_0$: lower-tailed test
   - $H_a : \mu \neq \mu_0$: two-tailed test

   Examples of test statistics include z-tests (commonly used in teaching because its a little easier to understand, but not as applicable for research because we don't always know the population σ) and t-tests. Below, I show the equation for a t-test, one of the more commonly used tests.

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{15}$$

3. Calculate t, and then map it on to a Student's t-distribution. Compute the probability, or p-value, by integrating to the extrema. Compare this to your level of significance (commonly, you want $p < 0.05$ to reject the null hypothesis).

You will very rarely see this process laid out in papers. Oftentimes, people will just report the p-value.

3.6 Linear Regression

Another basic statistical technique is finding the trend in data. There are several ways to do this, but linear regression is one of the most basic and common. It can be a great starting point to inform future analysis. A common way to do this is linear least squares, which minimizes the distance between each datapoint and the regression line. There are several ways to represent this calculation, but I'm going to show it with linear algebra. As you progress to more advanced statistics, most things are represented with linear algebra, so this is a good introduction to thinking about your statistics in terms of vectors and matrices.

Linear regression assumes that we have an independent variable x and a dependent variable y and that they are linearly related, so we can assume a relationship of the following form for every data point

$$y = mx + b \tag{16}$$

In Eq. 16, m is a constant that is the slope, or trend (how much does change in y does a change in x elicit). b is a constant that is the y-intercept, or offset. Here, x and y are our knowns, and we want to solve for m and b. Is there a way we can rephrase this well-known equation so it's purely linear algebra (i.e., we want to put those coefficients into a vector)?

Well, adding a coefficient in an equation like we do with b in Eq. 16 is the same thing as adding a vector of ones multiplied by the coefficent in linear algebra, like so

$$\vec{y} = m\vec{x} + b\vec{1} \tag{17}$$

And remember, when we multiply a matrix by a vector, we're essentially multiplying the first value in the vector by the first column, multiplying the second value in the vector by the second column, and so forth, and then adding all resultant columns together. So we can rewrite Eq. 17 like

$$\vec{y} = \mathbf{E}\vec{c} \tag{18}$$

$\mathbf{E}$ in Eq. 17 is the horizontal concatenation of the one-vector and the $\vec{x}$ vector, like so

$$\mathbf{E} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \dots & \dots \\ x_n & 1 \end{bmatrix} \tag{19}$$

and $\vec{c}$ in Eq. 17 is the coefficient vector, like so

$$\vec{c} = \begin{bmatrix} m \\ b \end{bmatrix} \tag{20}$$

We'll now define something called the noise vector, or the misfit vector, which is the error incurred by assuming values for m and b. This noise vector n is defined as such

$$\vec{n} = \vec{y} - \mathbf{E}\vec{c} \tag{21}$$

Now, the vector $\vec{n}$ will tell us the mismatch between every data point and its value as calculated according to our coefficent vector $\vec{c}$. We want to minimize the magnitude of this vector. We can calculate a vector's magnitude by multiplying its transpose by itself, like so

$$J = \vec{n}^T\vec{n} \tag{22}$$

This is equivalent to squaring every element of the noise vector and adding all the values together to create a scalar. Now, we want to minimize this scalar. The best way to do this is to take the derivative of J with respect to the coefficent vector $\vec{c}$ and set it equal to zero. Recall how our scalar J depends on $\mathbf{E}, \vec{y}$ and $\vec{c}$

$$J = (\vec{y} - \mathbf{E}\vec{c})^T(\vec{y} - \mathbf{E}\vec{c}) \tag{23}$$

By the chain rule for linear algebra,

$$\frac{\partial J}{\partial \vec{c}} = \frac{\partial \vec{n}}{\partial \vec{c}} \frac{\partial J}{\partial \vec{n}} \tag{24}$$

Now we calculate the derivative of the noise vector with repsect to the coefficient vector, finding

$$\frac{\partial \vec{n}}{\partial \vec{c}} = \frac{\partial(\vec{y} - \mathbf{E}\vec{c})}{\partial \vec{c}} = -\mathbf{E}^T \tag{25}$$

Now, we calculate the derivative of the scalar J with respect to the noise vector, finding

$$\frac{\partial J}{\partial \vec{n}} = \frac{\partial(\vec{n}^T\vec{n})}{\partial \vec{n}} = 2\vec{n} = 2(\vec{y} - \mathbf{E}\vec{c}) \tag{26}$$

Now we can multiply together the results of Eqs. 25 and 26, finding that

$$\frac{\partial J}{\partial \vec{c}} = -2\mathbf{E}^T\vec{n} = -2\mathbf{E}^T(\vec{y} - \mathbf{E}\vec{c}) \tag{27}$$

Now, recall that the scalar J will be minimized when this derivative is equal to zero. So now we can solve for our coefficent vector $\vec{c}$.

$$0 = -2\mathbf{E}^T(\vec{y} - \mathbf{E}\vec{c}) = \mathbf{E}^T(\vec{y} - \mathbf{E}\vec{c}) \tag{28}$$

$$\mathbf{E}^T\vec{y} = \mathbf{E}^T\mathbf{E}\vec{c} \tag{29}$$

In linear algebra, we take the inverse instead of multiplying, so

$$(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\vec{y} = \vec{c} \tag{30}$$

Although the calculus might be a little less intuitive, solving this via linear algebra can make it much easier to extend this calculation to propagate error, and it makes the relation clearer for higher order regression. It also is very fast and computationally efficient.

Let's quickly calculate the trend of our Tampa Bay temperature dataset. I show this calculation in the attached Jupyter Notebooks. In one very short, efficient line of code, I calculate both the slope and y-intercept in one fell swoop. Here's the result

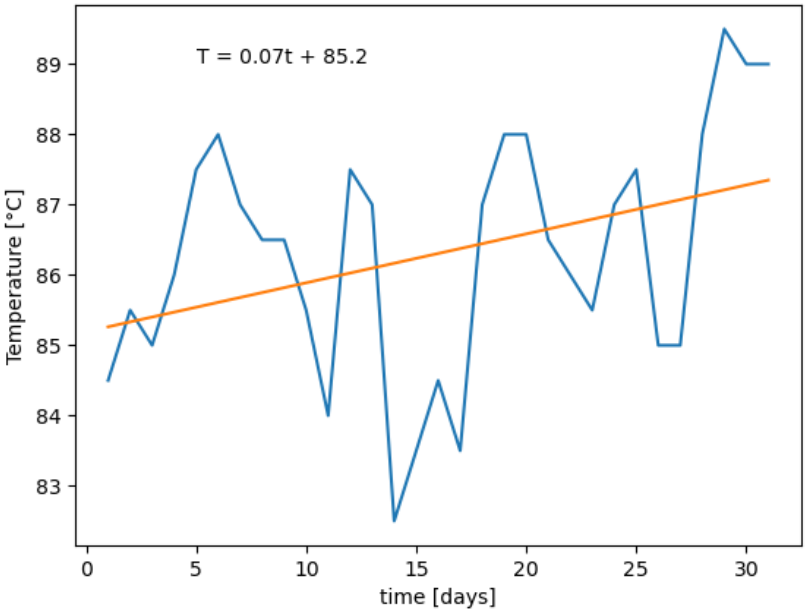

Figure 7: Result of computing the linear regression of for the June average daily temperatures in Tampa Bay

4 Acknowledgements

This review builds on materials from prior math review courses, taught by

- Arianna Krinos
- The 2019 Instructor Who Didn't Write Their Name on Their Notes
- Jeffrey Mei
- Hannah Mark

It also builds on

- class notes from 12.805 (Data Analysis in Physical Oceanography, taught by Jake Gebbie and Tom Farrar)
- 2022 Geophysical Fluid Dynamics Lecture series taught by Peter Schmid and Laure Zanna
- MIT OCW Course 6.041 taught by John Tsitsiklis
- a lot of erratic Googling