Math Review 2024: Probability and Statistics

Lilli Enders

August 5th, 2024

1 Learning Objectives

Welcome to our statistics and probability review session! Today, we're going to cover some of the fundamentals of probability and statistics, and really focus on developing skills that you can apply to your research (whatever that may be!). If you want to talk more about statistics after this session, feel free to reach out to me (lilli.enders@whoi.edu) or to check out the resources listed at the end of this document. Today we'll be covering:

- Events and sample spaces
- Probability and joint probability
- Probability density functions and cumulative density functions
- Methods to assess the character of data
- The normal distribution and central limit theorem
- Hypothesis testing
- Linear regression

The examples we'll look at are in the Jupyter Notebook I've provided (also on my github). Let's start our whirlwind tour of probability and stats!

2 *Probability Basics*

Probability is a tool that we use to understand **uncertainty.** We'll talk about probability in terms of **random variables**, which are processes in nature that we don't know the value of. Probability theory gives us a tool to quantify the uncertainty that is inherent to these processes, so that we can make meaningful conclusions about the nature of the process.

The steps required to calculate the chance of a particular event happening are probably intuitive to you already. Generally, we want to:

1. Define an *experiment*, where we make an exhaustive list of all possible outcomes
2. Determine the *relative likelihood* of each outcome
3. Compare the likelihood of all possible outcomes to determine the *probability* of each outcome.

In practice (at least in my experience), a lot of probability and statistics is learning how to frame your problem in the context of the (incredibly specific and often frustrating) terminology that statisticians use. Here we'll review some of that terminology, and get some practice putting data into framework that gives us a useful result.

2.1 Events and Sample Spaces

A *sample space*, which we'll denote by **S**, is a list of each of the possible outcomes of an experiment, where each item in the list is an *event*. An *event* is the simplest unit of the outcome, such that it can't be broken down into yet simpler outcomes (ex. rather than an event being "it rained on a Tuesday", we would have two events, "it rained" and "it was a Tuesday").

In some cases, defining a sample space is relatively straightforward. You've probably heard of sample spaces in the context of coin tosses. For three consecutive tosses of a fair coin, the events in **S = HHH, HHT, HTH, THH, TTH, THT, HTT, TTT**, where **H** = 'head' and **T** = 'tail'.

In most cases (i.e., in your research), it will probably be a little more difficult to define a sample space. We want to focus on two criteria for defining events within a

sample space: (1) events should be **mutually exclusive** and (2) events should be **collectively exhaustive**. Let's say, for example, that you're really interested in the probability that it will rain on a summer day in Falmouth because you really want to get an ice cream cone and it bums you out to eat ice cream in the rain. In this case, we might focus on setting thresholds for rainfall. We could define our sample space **S = 'it does not rain', 'it rains'**, where 'it does not rain' means that the rainfall is less than 0.1 cm, and 'it rains' means that it rains more than 0.1 cm. (note: This is a *discrete sample space*, because there are a fixed number of outcomes. We'll talk about *continuous sample spaces* a bit later.)

2.2 Probability of an Event

We assign probability to our events, with a probability between 0 (not gonna happen!) and 1 (absolutely gonna happen!). We'll use the notation that $\mathbf{P}(A)$ is the probability that event A occurs. Much of probability theory will surround the analysis of the probability of two or more events. We will use the notation $A \cap B$ to indicate the **intersection** of two events A and B (both events occurred, i.e., *A and B*) and the notation $A \cup B$ to indicate the **union** of two events A and B (one or two of the events occurred, i.e., *A or B*).

Some basic rules of probability are as follows.

1. Probabilities are always non-negative
2. When an experiment is conducted, one of the events in S *must* occur, so $\mathbf{P}(S) = 1$
3. Each event has a complement (which is often easier to compute). For an event A, the complementary event $\neg A$ is the collection of all the elements in S which are not contained in A, such that

$$\mathbf{P}(A) = 1 - P(\neg A) \tag{1}$$

4. If two events do not intersect (i.e., are **independent**)

$$A \cap B = 0 \tag{2}$$

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

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) \tag{3}$$

5. In general, the probability of observing one of two events A and B is

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \tag{4}$$

2.3 Conditional Probability

In our research, we often want to consider *conditional probability*, i.e., the probability that an event occurs, given that another event has occurred. We define the notation $\mathbf{P}(A|B)$ as the probability of A, given that B occurred. A and B do not need to be **mutually exclusive**.

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

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

$$\text{Joint Probability: } \mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A|B) \tag{6}$$

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

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B|A) \tag{7}$$

2.3.1 Independence

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

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \tag{8}$$

If an event A is **independent** from event B, then $\mathbf{P}(A|B) = \mathbf{P}(A)$.

2.3.2 Bayes' Theorem

When we're working with conditional probabilities, it's helpful to have a tool for incorporating prior information into our analysis. Bayes' Theorem allows us a way to revise existing predictions (i.e., update our probabilities) given new evidence. We can use the equations we just looked at for conditional probability to derive Bayes' Theorem:

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

Intuitively, we can consider Bayes' Theorem as the probability that two events will occur, divided by a normalizing factor. There is a whole field of analysis surrounding this fundamental idea called **Bayesian Analysis**.

2.3.3 Example: *Detecting Whales*

Let's imagine that we have a piece of equipment that detects whales acoustically. We know that:

- Our previous data collection suggests that there are whales in our study area 5% of the time
- Our equipment will detect whales 99% of the time **if there is a whale there** (if we miss a whale: Type II error, false negative).
- Our equipment will detect a whale 10% of the time **if there is no whale there** (if we detect a whale: Type I error, false positive).

We can summarize this information in terms of probabilities:

$$\mathbf{P} = \begin{cases} 0.95 \quad \text{no whale} \begin{cases} 0.9 \quad \text{no detection} \\ 0.1 \quad \text{detection} \end{cases} \\ 0.05 \quad \text{whale} \begin{cases} 0.01 \quad \text{no detection} \\ 0.99 \quad \text{detection} \end{cases} \end{cases} \tag{10}$$

We get a detection signal that tells us a whale is in our study area. What is the probability that its *actually* a whale? *Hint*: calculate the joint probability

$$\mathbf{P}(\text{whale}|\text{detection}) = \frac{\mathbf{P}(\text{whale} \cap \text{detection})}{\mathbf{P}(\text{detection})} \tag{11}$$

2.4 Random Variables

Now that we've locked down some of the fundamental terminology, it's easier to define exactly what a random variable is. Usually, we're not interested in the sample space S itself, but instead in an event in S that can be characterized by functions defined on S. If we think about our whale detection example, the function might be the number of times that we correctly detect a whale in our study region. These functions are referred to as **random variables**.

We'll denote random variables as X, and a particular value taken by the variable (i.e., a **realization of** X as x. Random variables are *random* because we can't predict their values in advance, and they are *variables* because their values depend upon which event in S takes place.

2.5 Probability Density Functions

2.5.1 *Probability Density Functions (PDFs)*

$p(X)$: Probability density functions take in an outcome for a given random variable, and return the probability of that outcome. PDFs map from a random value X to a likelihood, p. We can interpret the PDF as the probability that the value X occurs.

2.5.2 *Cumulative Density Functions (CDFs)*

$P_x(X)$: Cumulative density functions are defined for a random variable X such that $P_x(X)$ is the probability that $x \leq X$. We can interpret the CDF as the probability that a value is equal to or less than X. It is the integral of the PDF $p(X)$ up to that point, and the PDF is the derivative of the CDF.

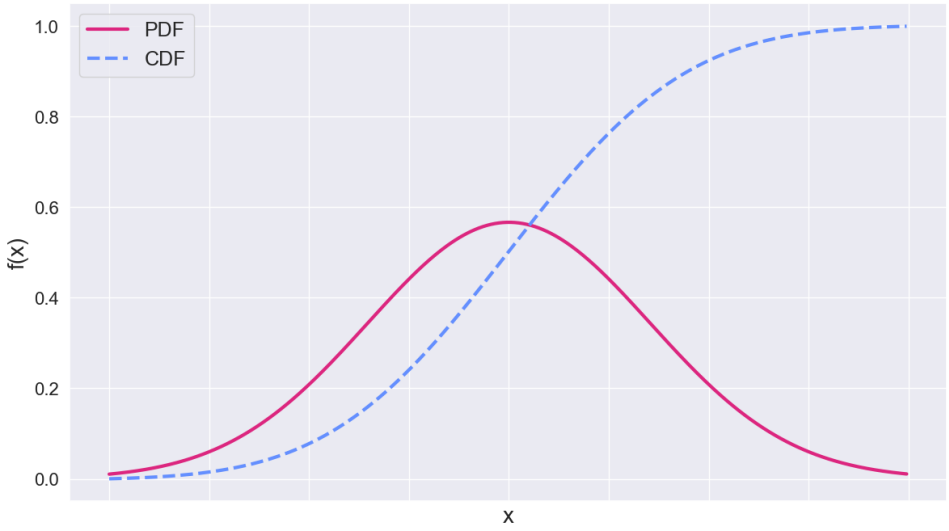

Figure 1: PDF and CDF example

3 *Statistics Basics*

Now that we're feeling comfortable with probability basics, lets dig into *statistics*. Statistics leverages the rules of probability to analyze data. Probability deals with predictions about the future, while statistics deals with evaluating information from the past (using probability).

3.1 Assessing the Character of Data

There are many quantitative descriptors that we can apply to our datasets in order to get a better understanding of their statistical characteristics. We can calculate many of these descriptors for both **continuous**, and **discrete** datasets.

Lets start by talking about the *central moments* of random variables. Most characteristics of a function can be described by the first four moments: ***mean, variance, skewness, and kurtosis***. These moments can be calculated in the continuous and discrete cases, and they tell us something about the location, scale, and shape of a random variable's distribution.

Note: I didn't include the discreet definition of kurtosis here because it's cumbersome and not particularly necessary for most cases, but it's easy to look up if you want to compute it yourself!

Descriptor	Continuous Formula	Meaning
Mean (μ)	$\int_{-\infty}^{\infty} X p_x(X) dX$	$\frac{\sum_i^N x_i}{N}$
Variance (σ^2)	$\int_{-\infty}^{\infty} (X - < x >)^2 p_x(X) dX$	$\frac{\sum_i^N (x_i - \bar{x})^2}{N}$
Skewness (γ_1)	$\int_{-\infty}^{\infty} (X - < x >)^3 p_x(X) dX$	$\frac{n}{(n-1)(n-2)} \frac{\sum_i^N (x_i - \bar{x})^3}{N}$
Kurtosis (γ_2)	$\int_{-\infty}^{\infty} (X - < x >)^4 p_x(X) dX - 3$	

Table 1: The continuous and discrete definitions of the first four central moments.

The ***mean***, also known as the *location parameter*, is the first moment. It represents the center, or expectation, of the random variable. The ***variance***, also known as the *scale parameter*, represents the spread in the data. We'll also hear about the ***standard deviation***, which is the square root of variance, and another measure of the spread of a distribution ($\sqrt{\sigma^2}$). The standard deviation is useful because it has the same units as the measurement. Whenever someone says that a value estimated

from a distribution is $a \pm b$, usually a refers to some type of mean, and b refers to some type of standard deviation.

The ***skewness*** and ***kurtosis*** are together referred to as the *shape parameters*. These moments are probably less intuitive to you than the mean and variance. ***Skewness*** gives us a measure of the *variance of the variance*. Symmetric distributions have low skewness values ($\gamma_1 = 0$), asymmetrical distributions have non-zero skewness ($\gamma_1 > 0$, *positively* or *right skewed*, $\gamma_1 < 0$, *negatively* or *left skewed*). ***Kurtosis*** is a measure of the "tailedness" or "peakedness" of the distribution. It can be used to give us a measure of extreme values in our dataset, but isn't something that most of us will reach for very often.

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 quartile. The space between the first and third quartiles is referred to as the interquartile range (IQR).

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)

Relating Distributions to Each Other In our research, we often want to assess the similarity between two datasets. We can do this using simple statistics, such as ***covariance*** and ***correlation***. For two random variable distributions, X and Y, the covariance is given by

$$\text{Covariance: } < (X - \mu_X)(Y - \mu_Y) > \tag{12}$$

where μ_X and μ_Y are the ***mean*** values of X and Y, respectively.

Correlation is a similar statistic, but it is scaled by the standard deviation of both distributions, σ_X and σ_Y. Oftentimes correlation will be used in time-series analysis to provide evidence for two variables being related to one another.

$$\text{Correlation: } \frac{< (X - \mu_X)(Y - \mu_Y) >}{\sigma_X \sigma_Y} \tag{13}$$

Covariances are not scaled, so they should not be compared between different sets of data. However, correlations are scaled, ranging from -1 to +1, with -1 indicating perfect anticorrelation, 0 indicating independence, and +1 indicating perfect correlation, and can be compared to each other. However, when assessing the significance of correlations, there are several important characteristics to keep track of, such as the auto-correlation, which can artificially inflate correlation significance.

3.1.1 Example: *Two Tidal Time Series*

Let's look at two tidal time series. The file I've provided has sea level data from two sites in Tracadie, New Brunswick. Let's think about the following questions:

- Based on physics/climatology, what do we expect about these two datasets?
- What are the general statistics of this dataset?
- How are they similar/how are they different?

As a starting point, I've plotted the **time series** and the **histogram** of the data from each site here. The rest of the solutions are in the attached Jupyter Notebook.

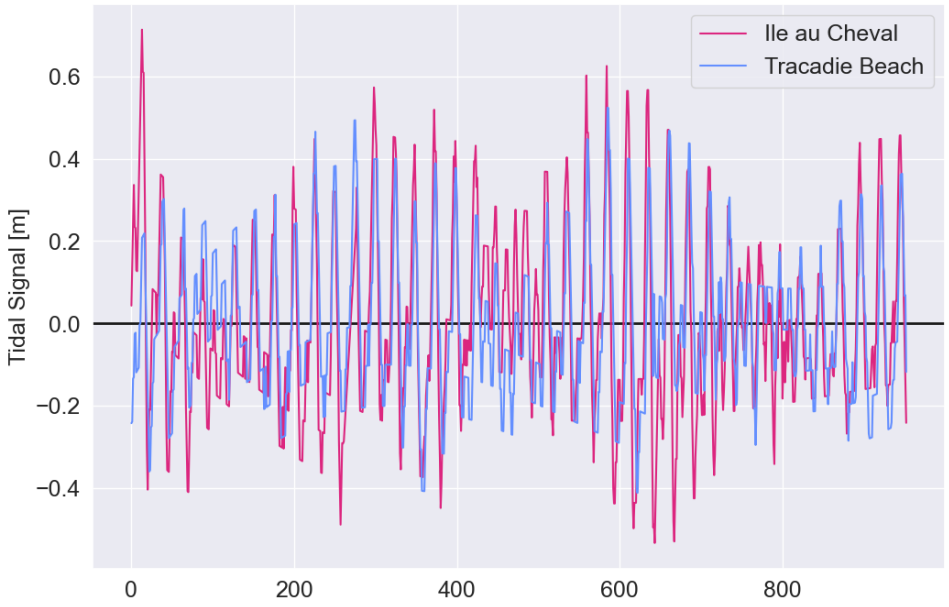

Figure 2: Time series of tidally forced sea level, in m, from two sites in Tracadie, New Brunswick

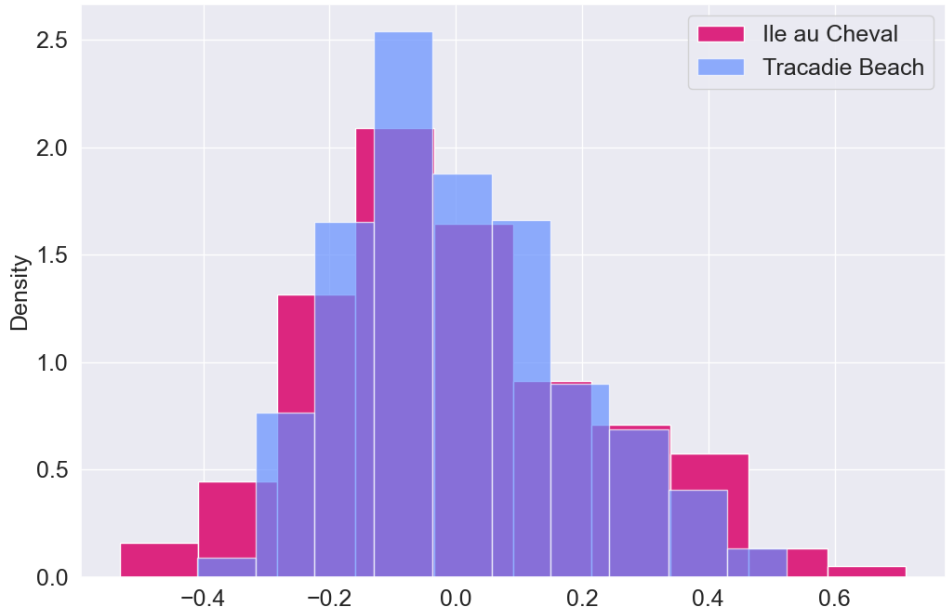

Figure 3: Histograms of tidally forced sea level, in m, from two sites in Tracadie, New Brunswick

3.2 The Normal Distribution and Central Limit Theorem

Normally distributed is a phrase that you'll hear a lot in statistics. Many statistical techniques, such as hypothesis testing, are primarily defined for normally distributed

variables. 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

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{14}$$

Here's my pitch for why the normal distribution is pretty cool:

- The normal distribution can be described using *only* the **mean** and the **standard deviation** (see ya never, skew and kurtosis!)
- The mean, median, and mode are all the same value in the normal distribution, so you know them without having to work too hard!
- There are three full standard deviation in either direction from the mean in a normal distribution. About 68% of the data is within 1 standard deviation (SD) in either direction from the mean, 95% within 2 SDs, and 99.7% within 3 SDs. If you've ever constructed a confidence interval, you know that this is really handy to know.

3.2.1 Central Limit Theorem

Although many statistical techniques are only defined for normally distributed data, many types of data are not normally distributed. However, because of the **Central Limit Theorem**, we can still work with non-normally distributed data. The Central Limit Theorem states that, for any random variable, with any distribution, the fluctuations between its sample mean and population mean will become normally distributed, when scaled by the number of samples. The Central Limit Theorem states that for any independent, identically distributed (IID) variables (this means that it doesn't matter what distribution these variables have, or even whether you know it, it just has to be reasonably the same), as the number of distributions N goes to infinity, a random variable represented by:

$$Z = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{15}$$

has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. This can be thought of intuitively like this: as we accumulate more and more data points, we will have more and more of the mean or intermediate values relative to the extreme

values. This is because each event has the same small probability of an extreme event occurring, hence the probability that extreme values dominate/skew the dataset is becoming smaller and smaller.

3.3 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 **one-sample** t-test, one of the more commonly used tests (note the connection to the Central Limit Theorem here: our data doesn't need to

be normally distributed, because were looking at the differences between population and sample mean, which we now know is normally distributed!).

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{16}$$

3. Calculate t, and then map it on to a Student's t-distribution (a generalized form of the standard normal distribution that accounts for degrees of freedom). Compute the probability of getting this t value if the null hypothesis were false (the p-value). This is done by integrating the Student's t-distribution PDF to the extrema. Compare this to your level of significance (commonly, you want $p < 0.05$ to reject the null hypothesis).

3.3.1 Example: *Mean of île au Cheval Dataset*

Using our tidal dataset, let's practice hypothesis testing. Imagine that we want to know whether the mean of the île au Cheval dataset is **significantly different** than -0.1m. Perform a hypothesis test deduce whether this is true.

3.4 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.

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

$$y = mx + b \tag{17}$$

Here, 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 ***known***, and we want to solve for m and b. So, for our datapoints y, the formula for least squares is:

$$error(regression) = \sum_i (y_i - (mx_i + b))^2 \tag{18}$$

Here, the error of the regression is really the **error of the two coefficients**, which we can calculate using differentiation. To get the slope, m, we take the derivative of the error with respect to the slope and set the resulting expression equal to zero. We can do the same thing with respect to the y-intercept to get the value of b.

$$\frac{dE}{db} = -2(\sum_i (y_i - (mx_i + b))) = -2 \text{ mean}(y) - 2m \text{ mean}(x) - 2b = 0 \tag{19}$$

$$b = \text{mean}(y) - m \text{ mean}(x) \tag{20}$$

Now for m:

$$\begin{aligned} \frac{dE}{db} &= \frac{dE}{dm}(\sum_i (y_i - (mx_i + b))^2) \\ &= \frac{dE}{dm}(\sum_i (y_i^2 - (m^2 x_i^2 + b^2 N - 2mx_i y_i - 2by_i + 2bmx_i)) \\ &= 2m\sum_i x_i^2 - 2\sum_i x_i y_i + 2b\sum_i x_i \end{aligned} \tag{21}$$

When we set this derivative equal to zero, we can solve for m:

$$\begin{aligned} 0 &= 2m\sum_i x_i^2 - 2\sum_i x_i y_i + 2b\sum_i x_i \\ 0 &= 2m\sum_i x_i^2 - 2\sum_i x_i y_i + 2\Big(\frac{\sum_i y_i}{N} - m\frac{\sum_i x_i}{N}\Big)\sum_i x_i \\ m\sum_i x_i^2 - m\frac{(\sum_i x_i)^2)}{N} &= \sum_i x_i y_i - \frac{\sum_i y_i}{N}\sum_i x_i \\ m &= \frac{\sum_i x_i y_i - \frac{\sum_i y_i}{N}\sum_i x_i}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2)}{N}} \end{aligned} \tag{22}$$

3.4.1 Example: *Fitting a Linear Trend to Lobster Landing Data*

Let's look at a new dataset, which has annual American Lobster landings (in pounds) from Maine. A quick look at the data shows us that the landings are increasing over time. Let's fit a linear trend to this data to quantify the change in landings over time.

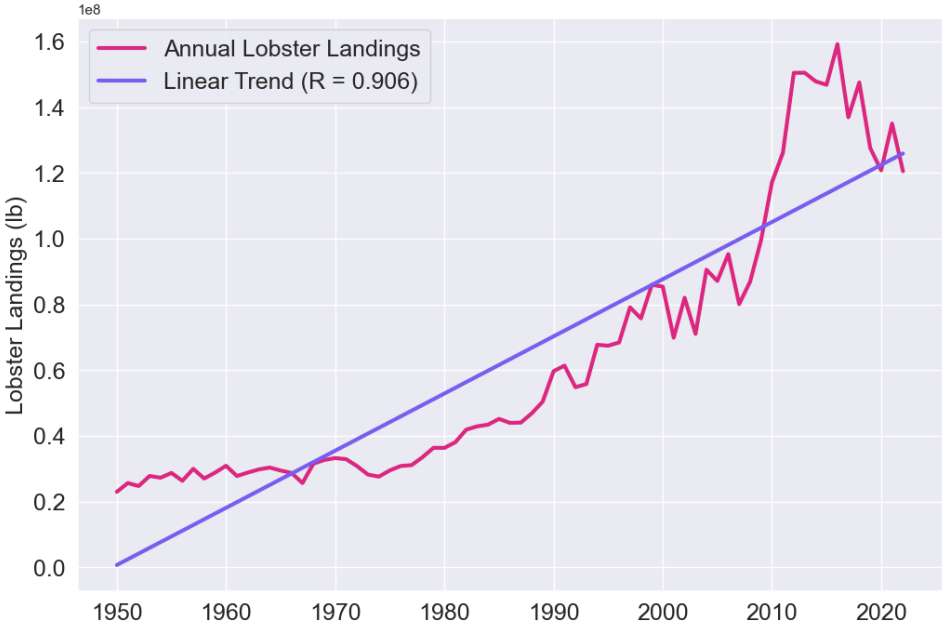

Figure 4: Annual lobster landings, with linear trend

4 Resources

In an hour and a half, we've really only touched the surface of probability and statistics. You'll likely return to these concepts and build on them as you dig into your own research. I hope this infrastructure will help you in building your stats toolbox! This material was adapted from previous teachers of this course, particularly Arianna Krinos and Brynn Hamilton.

Some of my favourite resources for continued statistics learning:

- Courses! At MIT: 18.05 (Introduction to Probability and Statistics), at WHOI: 12.805 (Dynamical Insights from Data/Data Analysis in Physical Oceanography)
- Statistical Analysis in Climate Research by von Storch and Zweirs
- Your research and your friends research!