Contents

1 Random variables

1.1 The binomial distribution

Definition. A random variable is essentially a random number that is generated from some process.

The binomial distribution is an important example. Essentially, we do something some number (n) of times and each time we do it we have a probability (p) of a success. We then want to find the probability that we will have a certain number of successes.

Example. In Minecraft 1.16, there is a probability of 4.7% to get ender pearls when you barter with a Piglin. Dream does 262 trades and got ender pearls 42 times. If he is not cheating, what is the probability that he actually got at least 42 ender pearl trades by sheer luck?

In the above case, $n = 262$ and $p = 4.7\% = 0.047$. We could say that X is the random variable that represents the number of pearl trades, and then the notation would be $X \sim B(262, 0.047)$. In general, $X \sim B(n, p)$ means X follows a binomial distribution with n trials and a probability p of success on each trial. We have the usual assumptions like that the trials are independent of eachother.

We will compute the probability that dream got exactly 42 pearl trades. We ought to compute the probability of at least 42 (since if there were way more trials the probability of getting any exact number of successes will be very small even if it is close to the expected number), but a computer could do this in principle by computing all the values and adding them up, and statistical tables do the job.

With that preamble out of the way, lets do the calculation. To work out the probability, we will find how many "ways" there are to get 42 pearl trades, and then calculate the probability of each way. To make this precise, we could have a sequence like

YNNNNNNNNYNNYYN...NNYNNN

where Y means a pearl trade and N means no pearl trade. There are 262 symbols in total, but the Y's could go in any 42 positions. From level 3.1, we know that this means there are exactly $\begin{pmatrix} 262 \\ 42 \end{pmatrix}$ ways to make this list. Of course, this is a very large number.

The probability of each way is the probability of getting an N on all 220 of the trials you get it on and a Y on all 42 of the trials we get it on. This is of course, astronomically small, but we multiply it by the very large number of ways. Since the probability of a Y is 0.047 and the probability of an N is $1 - 0.047 = 0.953$, our final formula for $P(X = 42)$, which is

shorthand for "the probability that X is 42", is

$$P\left(X=42\right)=\left(\begin{array}{c}262\\42\end{array}\right)\left(0.047\right)^{42}\left(1-0.047\right)^{262-42}$$

If we do the adding up thing to get $P\left(X\geq 42\right)$ we get 0.000000000432%.

This means it is more plausible that Dream was cheating.

The general formula for a binomial distribution probability p and n trials is as follows:

$$P\left(X=k\right)=\left(\begin{array}{c}n\\k\end{array}\right)\left(p\right)^{k}\left(1-p\right)^{n-k}$$

1.2 The poisson distribution

The poisson distribution essentially considers situations where an event happens randomly and at a specific rate, ie on average once every 10 minutes. It might be something like customers entering a store, or traffic accidents. We have the usual assumptions like that things are independent of eachother and all that.

If the average number of times an event will occur is λ, then the random variable X that measures the number of times the event occurs is said to have a Poisson distribution with mean λ. We write $X\sim Po\left(\lambda\right)$.

To find $P(X=n)$, we can derive the formula by trying to take the limit of a binomial distribution and making p small and n really large, but we will not do this here. The formula, which we will prove in level 6.3 says that

$$P(X=n)=\frac{e^{-\lambda}\lambda^{n}}{n!}$$

We can of course plug numbers into this formula and do calculations.

It is, in fact, standard to approximate a binomial distribution with n large and p small as a Poisson distribution with parameter $\lambda=np$.

If we want something like $P(X\leq n)$, we would often use statistical tables or a computer to avoid adding up a bunch of values.

Example. If I expect something to happen 3 times and it does not happen at all, the formula gives that this has a probability of e^{-3}, which is just under 5%, meaning I am kind of unlucky, as you would expect.

2 Measures of location and spread

2.1 Expected value

If X is a random variable, then $E[X]$ means the "Expected value" of X. This essentially means that if we keep taking instances of X and take their average, this is the number that it will approach. So in the case of the poisson distribution, this is exactly the parameter λ.

In the case of a binomial distribution with probability p, on each trial, if we consider a value of 1 to be taken if we get a success and 0 otherwise, we have an "expected value" of p since on average we expect to get 1 with a proportion p of the time. It is intuitively true that expectations add up, in the sense that $E[X+Y]=E[X]+E[Y]$, so if we take a sum over all n trials, we get a total expected value of np, which makes sense if you think about it.

Similarly, if c is a number, then $E[cX]=cE[X]$. Again, this is fairly obvious.

This is not to be confused with the median, which is the least value m such that $P(X\leq m)\geq 0.5$. These do not always coincide.

Now we want to discuss how to actually find the expectation of a random variable. We do this by taking a weighted sum. Lets say we have a random variable that takes the value 1 with probability 10%, 2 with probability 20%, 3 with probability 30%, and 4 with probability 40%. As a sanity check, the probabilities do indeed add up to 100%. The expectation would be the weighted sum $1*0.1+2*0.2+3*0.3+4*0.4$, which comes out to be equal to 3.

2.2 Variance and standard deviation

We want to measure how far our data is likely to go from its expected value, ie how spread out it tends to be. One idea is to consider the mean absolute deviation, which is

$$E\left[|X - E[X]|\right]$$

However, we will not do this. Things tend to be nicer if we consider the average *squared* distance from the mean. It's not immediately obvious why, but we will see as we go through maths. Specifically, we want to look at $E\left[(X - E[X])^2\right]$. This is called the variance of X. We call this $Var[X]$.

Proposition. If we have 2 random variables X and Y, and they are independent of eachother, then $Var[X] + Var[Y] = Var[X + Y]$.

We will not prove this here, we defer it to level 6.3. We note that for expectation it holds even if the variables are not independent, since we could take many trials of both variables and see what the sum tends to converge to. The above property is not true for the mean absolute deviation, even for independent variables, which I suppose is one justification to use variance instead.

Now note that if c is a number and we consider cX, then $Var[cX] = E\left[(cX - E[cX])^2\right] = E\left[c^2(X - E[X])^2\right] = c^2 Var[X]$. It is not $cVar[X]$, so don't make that mistake.

To make this intuitive property hold, we take the square root of the variance. This is called the standard deviation. It seems silly to take the "square root of the mean squared deviation", but that is what we do.

Proposition. $Var[X] = E[X^2] - E[X]^2$

Proof. We note that $E[X]$ is actually just ssome constant. Then

$$Var[X] = E\left[(X - E[X])^2\right] = E\left[X^2 - 2XE[X] + E[X]^2\right] = E\left[X^2\right] - 2E[X]E[X] + E[X]^2 = E\left[X^2\right] - E[X]^2$$

□

Now here are some results which I will quote. Again, see level 6.3 for the derivation.

- The variance of a binomial distribution with parameters n and p is $np(1-p)$

- The variance of a poisson distribution with mean λ is also λ. This actually comes out quite nicely from the "limit of the binomial distribution" process.

3 Continuous random variables

3.1 Density and properties

We want to consider random variables that can take arbitrary real values. For example, what if we want to generate a random number between 0 and 1. Here, the probability of landing on any specific number is 0, but we can meaningfully talk about the probability that we land, say, between 0.3 and 0.5 - this probability is 0.2.

If we have a continuous random variable X (Or in fact, any random variable), we can define the cumulative distribution function (cdf) as $F(t) = P(X \leq t)$, ie the probability that X is less than t. This will be a function that will be 0 or close to 0 for X very small or very large and negative, and 1 or close to 1 for X very large.

In the case where this F is something you can differentiate (except at possibly finitely many points that won't affect things), and we will assume the derivative is made up of a bunch of continuous parts (just to ensure you can integrate it), we will differentiate it to get $f(x)$. This is the probability density function (pdf). The intuition for this is that the probability of landing between x and $x + dx$ for small dx is about $f(x) * dx$.

In the "random number between 0 and 1" example above, the cumulative distribution function is as follows:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

And the probability density function is as follows:

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We can find $P(a \leq X \leq b)$ by integrating the density from a to b. In the above example, we would see that

$$P\left(0.3 \leq X \leq 0.5\right) = \int_{0.3}^{0.5} 1dx = 0.2$$

It would also be perfectly valid to consider

$$P\left(-1 \leq X \leq 0.4\right) = \int_{-1}^{0.4} f(x)dx = \int_{-1}^{0} 0dx + \int_{0}^{0.4} 1dx = 0.4$$

In general, if we are picking a random number from a to b for some a, b with $a < b$, and we do this in such a way that the density is constant, then this is called a uniform distribution.

We note that the density of the whole thing must equal 1, so this constant density (to ensure this) will always be $\frac{1}{b-a}$. As an example, if we want to pick a random number between 3 and 8, then the density would be

$$f(x) = \begin{cases} \frac{1}{5} = 0.2 & 3 \leq x \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

We could have other continuous random variables, as an example, we might have something with a density like

$$f(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The $\frac{3}{4}$ is to ensure the integral of the whole thing is 1. The graph of this density would look like Figure 1: It is essentially a random variable that takes values between -1 and 1, but is more rigged to be more likely to take values near 0.

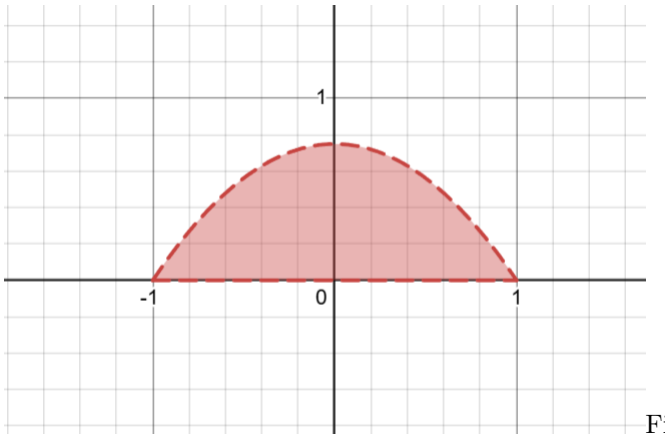

Figure 1

Now we want to discuss how to find the expectation of a continuous random variable. We do the analog of a weighted sum and just take a weighted integral. In the discrete case, we would take $\sum xP(X = x)$, and now we will take $\int_{-\infty}^{\infty} xf(x)dx$.

Example. Lets find the expected value of a uniform distribution that takes values from 0 to 1. We expect this to be $\frac{1}{2}$, of course. Since $f(x)$ is 0 for everything not between 0 and 1, we only need to consider the integral from 0 to 1. We have

$$\int_0^1 xf(x)dx = \int_0^1 xdx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

as expected.

Similarly, we can find the variance. We know that the variance is $E[X^2] - E[X]^2$, and we know how to calculate $E[X]$ and square it, but not how to calculate $E[X^2]$.

However, we can do this by just taking a weighted sum of the probability x takes a certain value times x^2 (in the discrete case). We can convert the weighted sum into an integral, and we have

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx$$

We get that in the case of a uniform distribution from 0 to 1, $E[X^2] = \frac{1}{3}$ if we go through the calculation and $E[X]^2 = \frac{1}{4}$, so if we subtract them we get that the variance is $\frac{1}{12}$. Since the variance will not be affected if we move the uniform distribution around and if we move the end points c apart the variance will multiply by c^2, we get the general formula that for a uniform distribution from a to b, its variance is $\frac{1}{12}(b-a)^2$, and its standard deviation is $\frac{1}{\sqrt{12}}(b-a)$.

Now we can generalize this: For any integrable function $g(x)$,

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

We will see in Level 6 this defined in terms of the Lebesgue integral, which will even allow us (by a clever change of how we view the "length" of parts of the number line) to use integrals to deal with the discrete case, and this will all make sense.

Definition. The median of a continuous random variable is the value of m such that

$$\int_{-\infty}^{m} f(x)dx = 0.5$$

Definition. The mode of a continuous random variable is defined as the value where f attains its maximum.

3.2 The normal distribution

Definition. The normal distribution with mean μ and variance σ^2 is the continuous random variable with density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

I have implicitly assumed a lot of stuff which I will not prove here, I will prove them in level 6.3. These are that

1. $\sqrt{2\pi}$ is the correct normalizing constant to make this an actual density

2. The variance of this if you compute it using an integral is actually σ. However, we will see that the mean is μ with a visual argument shortly.

If X is a random variable with a normal distribution with mean μ and variance σ^2 we write $X \sim N\left(\mu, \sigma^2\right)$

Definition. A standard normal is a normal distribution with $\mu = 0$ and $\sigma = 1$. This is often called a bell curve, and we will see later why this shape is universal in a sense and comes up a lot, since there is a deep mathematical reason. Figure 2 shows the graph of the probability density function of a standard normal.

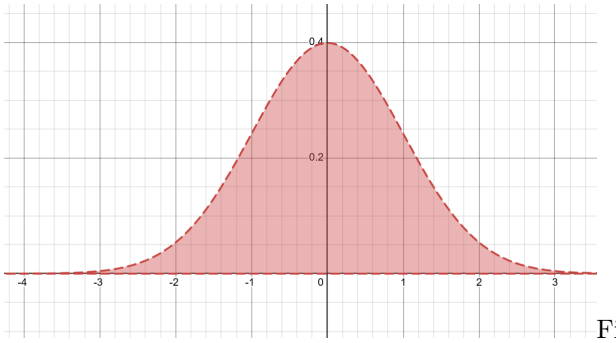

Figure 2

We note that $f(x) = f(-x)$, both from the graph and from the given equation. This symmetry property is useful, because we can deduce stuff like that $P(X > 0.3) = P(X < -0.3) = 1 - P(X > 0.3)$, for example. The normal distribution is so important that statistical tables usually include both

- $P(X < k)$ for many values of k, such as in intervals of 0.01, where X is a standard normal and

- For many p values, values of k such that $P(X < k) = p$

The general rule is that for a normal distribution, about $\frac{2}{3}$ of data is within 1 standard deviation of the mean, about 95% of data is within 2 standard deviations of the mean, and about 99.9% of data is within 3 standard deviations of the mean. This property holds for all values of μ and σ since mean and standard deviation scale the way you would expect as you shift and scale a distribution.

f we have a normal distribution with $\mu = 8$ and $\sigma = 3$ then about $\frac{2}{3}$ of the time we will be between 5 and 11. By the symmetry property, we see that in this case, $P(X < 5) = P(X > 11)$, and for the total probability to be 1, these values are both about $\frac{1}{6}$. We now see that $P(X < 11)$ is about $\frac{5}{6}$ in this case.

Example. One useful property of the normal distribution is that it can approximate many distributions. In level 5, we will see that this is a consequence of the central limit theorem, which we will prove in level 6. Specifically, here is what you need to know:

If we have a binomial distribution where np and $n(1-p)$ are large, often cited as n large and p close to $\frac{1}{2}$, we can approximate it as a normal distribution with $\mu = np, \sigma^2 = np(1-p)$. It is expected that the mean and variance should match, but not obvious why it should be a normal distribution - we will justify this in level 6.

If we have a Poisson distribution with parameter λ and λ is large (Heuristically, more than about 5 or 10 tends to be sufficient for this), then it can be well approximated by a normal distribution with $\mu = \sigma^2 = \lambda$.

If we do use normal approximations, we need to be careful: If $X \sim Po(100)$ and we want to use a normal approximation to find $P(95 \leq X \leq 105)$ then we are turning a distribution that takes integer values into a distribution that takes arbitrary real values. We therefore decide by convention that the cutoffs will be 94.5 and 105.5, so that the nearest integer will be between 95 and 105 inclusive.

Example. Consider the above example. For the poisson distribution, the actual value of $P(95 \leq X \leq 105)$ is about 0.41766. For a normal distribution with mean 100 and standard deviation 10, $P(95 \leq X \leq 105)$ is about 0.38292, but $P(94.5 \leq X \leq 105.5)$ is about 0.41768.

If X is a standard normal, and a statistical table that gives $P(X < k)$ for $k > 0$, and we want to find $P(-0.5 < X < 1)$ we can exploit symmetry. This is just

$$P(-0.5 < X < 1) = P(X < 1) - P(X < -0.5) = P(X < 1) - P(X > 0.5) = P(X < 1) - (1 - P(X < 0.5))$$

We can now calculate this using the table.

4 Hypothesis testing

The idea of hypothesis testing is that we want to determine, given some data, whether it is plausible. To demonstrate this, we will go back to the Dream example, where essentially we secretly did this.

First, we say that the "Null hypothesis", which is denoted H_0 is our prior assumption, in this case that Dream is not cheating, or that $p = 0.047$. The "Alternative hypothesis", which is denoted H_1, is that Dream is cheating, or that for dream $p > 0.047$.

We now pick a significance level. If we assume the Null hypothesis and calculate how lucky Dream was, we need to decide how lucky is too lucky. Often the significance level is 5%, 1% or 0.1% depending on how "sensitive" we want our test to be. We would also calculate how lucky it is that Dream got at LEAST as many ender pearls as he got instead of exactly, for reasons discussed earlier.

Note that if we get a p value of 0.1%, this does NOT mean there is a 99.9% chance Dream was cheating or a 0.1% chance he got lucky - This is very silly if you think about it, since this logic would mean that average luck would imply a 50% chance of cheating. It is just how lucky he would have had to get. If this p value is less than what you would assume prior is the probability he cheated, only then should you get suspicious.

Now, for the numbers $n = 262$ and the null hypothesis $p = 0.047$, the following table shows $P(X \geq k)$.

k	$P(X \geq k)$
17	11.39%
18	7.093%
19	4.208%
20	2.382%
21	1.287%
22	0.6648%
23	0.3287%
24	0.1557%
25	0.07078%
26	0.03090%
27	0.01297%
28	0.005237%

Definition. p-hacking is when you intentionally try to cheat to make your p value smaller. For example, many scientists might decide their significance level (ie threshold to reject the null hypothesis) is 1%, and then test 100 random hypotheses, so that 1 of them just so happens to pass, and this way they can publish a result. Another example would be that if you look at 1 million minecraft players and decide that your threshold is 0.1%, then on average you will find that 1000 of them are cheating. This is something to be careful of.

In the dream example, if you decide that 1% is your significance level, you will end up rejecting the null hypothesis if Dream gets 22 or more ender pearls. This actually makes the significance level of the test 0.6648%, since that is the actual probability you reject the null hypothesis and not 1%. This happens whenever your data is discrete.

We can of course do tests with poisson and normal distributions as well.

Definition. Critical values are the most extreme values under which you would not reject the null hypothesis

Note that we can do two tailed tests. What I mean is, in the above example, we did a one tailed test because our alternative hypothesis was that $p > 0.047$, but we can do a two tailed test in the sense that we are testing not if a parameter is more than what we expect, but if it is different in *either* direction.

Example. Suppose we have that the null hypothesis is X follows a standard normal distribution and we want to test at the 1% level of significance whether to accept the null hypothesis, where the alternative hypothesis is that the mean of X is either greater than or less than 0.

In this case, we need to have at most a 1% chance to reject the null hypothesis, so the convention is to get a 0.5% chance

on either side. So we would find the 0.5 percentile point and the 99.5 percentile point of the standard normal using some table, and those would be our critical values. These are about ± 2.5758.

5 Correlation and regression

5.1 Correlation

Suppose we have some joint random variables X and Y. We don't know if they are related to eachother, but we get some X values and the corresponding Y value and plot them on a scatter graph like this, like in figure 3

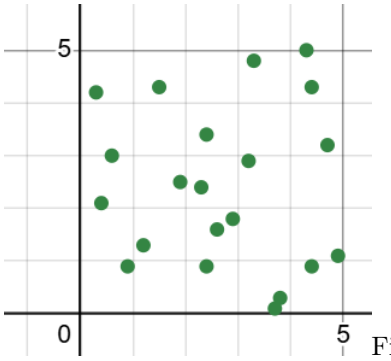

Figure 3

Given these data, if we have n data points, we define

- $\bar{x}$ is the mean of x, ie $\frac{1}{n}\sum_{i=1}^{n}$

- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, ie the sum of squared deviations.

- $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

- $S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$

Definition. We define the Product Moment Correlation Coefficient (PMCC) to be $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. We often call this r.

This takes values between -1 and 1. Values close to -1 indicate that x and y are inversely correlated (ie x is small when y is large), values close to 1 indicate that x and y are correlated, and values close to 0 indicate no clear trend. In level 6, we will see a really nice vector argument for why all of these assertions are true. Figure 4 shows some examples.

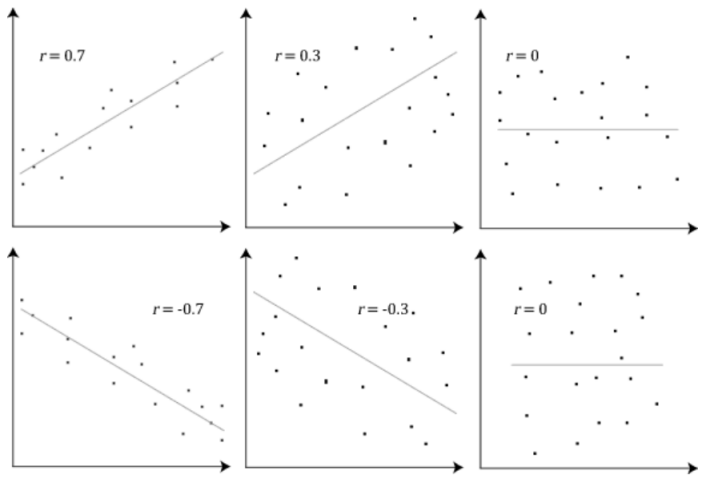

Figure 4

Of course, if we change y to $Ay + B$ or x to $Ax + B$ and $A > 0$ it is fairly easy to see from the formula that this will not affect the PMCC (ie, the numerator will be multiplied by A and the denominator will be multiplied by $\sqrt{A^2}$), which is a good sign.

There is another type of correlation coefficient we can use called the rank correlation coefficient.

Essentially, we take each pair (X,Y) and put them in a table. In each of the entries of the table we write the rank of X and Y from 1 to n, 1 meaning the highest value and n meaning the lowest. The rank correlation coefficient is then the PMCC between these ranks, which there is a formula for.

Example. The table might be as follows:

X	1	2	3	4	5	6	7	8
Y	2	3	1	6	4	7	8	5

Now, what we do is we calculate the differences in each column, d. The formula for the PMCC between these data, which we will derive in level 6.3, is

$$1 - 6\frac{\sum d^2}{n\left(n^2-1\right)}$$

For the above example, this would be

$$1 - 6\frac{(1-2)^2 + (2-3)^2 + (3-1)^2 + (4-6)^2 + (5-4)^2 + (6-7)^2 + (7-8)^2 + (8-5)^2}{8\left(8^2-1\right)}$$

$$= 1 - 6\frac{1+1+4+4+1+1+1+9}{504} = 1 - 6\frac{22}{504} \approx 0.738$$

It is possible to carry out a hypothesis test for correlation - There are tables that given a certain sample size, give values of r such that if the variables were uncorrelated and random, have a certain probability to occur by chance.

In the above case, if we were to do a two tailed test at the 5% significance level, the sample size is 8 so from a table you can find online or in a formula book, we will need the 2.5% and 97.5% percentage points of the PMCC, which are ± 0.7067. Since our rank correlation coefficient exceeds these critical values, we reject the null hypothesis in the above example and conclude that X and Y are positively correlated.

As the sample size gets larger, the critical values for the PMCC get closer to 0 which makes sense. Interestingly, one time given some data I found from the internet of about 600 Cambridge Maths students degree classification vs step grades, I got a PMCC of about 0.3 when converting all the data to numbers, which was significant.

The data in the image above, which was clearly uncorrelated as I put them in randomly, turned out to have a PMCC of -0.0498. This is not statistically significant what so ever for the sample size.

Note that correlation does not mean causation in the following sense. If X and Y are correlated and we are sure it is not a conincidence, one of the following things is true:

- X causes Y to change

- Y causes X to change

- Something else causes both X and Y to change.

It turns out that there is a strong correlation between shark attacks and ice cream sales depending on the time of year, but it is way more plausible that sharks attack more in the summer than that sharks really like ice cream or whatever your alternative proposal would be.

5.2 Regression

Regression is when given some data, we try to find a line that best fits the data. It turns out that a good way to do this is to try to minimize the sum of the squares of the errors shown in Figure 5.

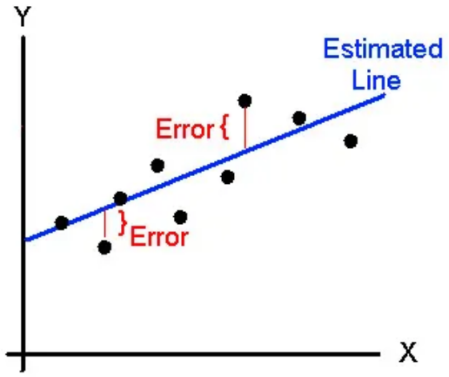

Figure 5

The formula for such a line is $y = a + bx$ where

$$b = \frac{s_{xy}}{s_{xx}}$$

and

$$a = \bar{y} - b\bar{x}$$

Indeed it is a good sanity check to see that $a + b\bar{x} = \bar{y}$ so the formula $y = a + bx$ is likely sensible, but the proof that it minimizes squares, while not too difficult, is tedious and deferred to level 6.3.

Given this minimal line, the sum of the errors squared is $S_{yy}\left(1 - r^2\right)$ where r is the PMCC. Again, the proof is deferred to level 6.3. This is called the residual sum of squares.