ST101 Unit 3: Estimation

Contents

Estimation

This section is all about estimators. We will introduce the Maximum Likelihood Estimator and the Laplacian Estimator. We will see how to use these techniques to derive probabilities from observed data, such as coin flips.

The subtitle for this section might be “To Fake or Not to Fake?”. The answer may surprise you!

Probability of Heads

Suppose we flip a coin six times and get the outcome H-T-T-H-T-H. Based solely on this data, we would say that the probability of heads, P(H), for this coin is the number of heads we observed divided by the number of times we flipped the coin:

$$P(H) = \frac{3}{6} = 0.5$$

Now suppose we take a different coin and flip it five times. We observe H-H-T-H-H. based on this data we would assume that P(H) is:

$$P(H) = \frac{4}{5} = 0.8$$

We call these the empirical frequencies. They are just the ratio of the number of outcomes in which the event occurs to the total number of trials. If we took a coin and flipped it seven times and we saw T-T-T-T-T-T-T we would assume that the probability of heads was:

$$P(H) = \frac{0}{7} = 0$$

Identify the Estimator Quiz

Which of the following formulae captures the method we used to calculate the empirical frequencies above?

A	B
$\sum_i X_i$	$\prod_i X_i$
C	D
$\frac{1}{N}\sum_i X_i$	$\frac{1}{N}\prod_i X_i$

We call this the Maximum Likelihood Estimator, and its value will always be between 0 and 1, which is a valid probability. It is a really good way to estimate the underlying probability that may have produced a given data set.

What happens when we have more than two outcomes.

Let's say that we have a six-sided die. This therefore has six possible outcomes. We throw the die ten times, and get the following outcomes:

1-6-6-3-2-6-5-4-6-2

Now, N = 10 so:

$$P(1) = \frac{1}{10} \times 1 = 0.1$$

$$P(2) = \frac{1}{10} \times 2 = 0.2$$

$$P(3) = \frac{1}{10} \times 1 = 0.1$$

$$P(4) = \frac{1}{10} \times 1 = 0.1$$

$$P(5) = \frac{1}{10} \times 1 = 0.1$$

$$P(1) = \frac{1}{10} \times 4 = 0.4$$

The sum of the probabilities is 1, which is what we would expect since these are all the possible outcomes.

Likelihood

The estimation problem that we are trying to overcome is, given some data, what is the probability, P, that would give rise to the observed data?

In earlier units, we have seen how to estimate the data that we would observe given the probability of some event occurring. We can record a '1' if the event happens, and a '0' if it does not.

Likelihood Quiz

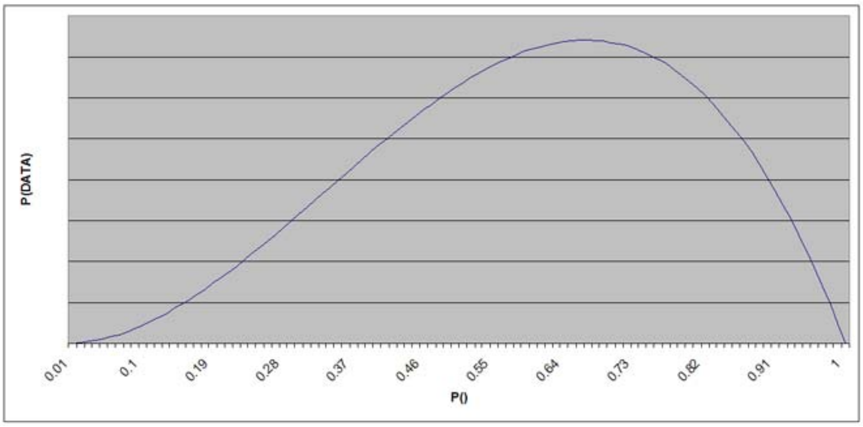
Suppose we see the following data:

1-0-1

For each value of P() what is the value of P(DATA)?

1. P() = ½
2. P() = ⅓
3. P() = ⅔
4. P() = 1

If we plot the probabilities on a graph, with P() on the x-axis and P(DATA) on the y-axis, we get:



The point at P() = 0.75 maximises the likelihood of the data. This is called the Maximum Likelihood Estimator, MLE.

Weakness

Suppose we flip a coin exactly once. It comes up heads. Using the Maximum Likelihood Estimator we would end up with a probability of heads:

$$P(H) = \frac{1}{1} = 1$$

Similarly, if we had seen tails, MLE would have given us:

$$P(H) = \frac{0}{1} = 0$$

Weakness Quiz

Does this mean that, from a single coin flip, the maximum likelihood estimator will always assume a loaded or biased coin?

Weakness Quiz 2

Suppose we make 111 coin flips. Will the maximum likelihood estimator always assume that the coin is loaded or biased?

Faking It

There is a solution. We can fake it!

More precisely, we can add fake data to the original data to create a larger pool of data points. In the case of a coin-flip, there are two possible outcomes – heads or tails – so we would add two fake data points to our data, one heads and one tails.

If we had just one result, we would see the following:

1	MLE = $P(H) = \frac{1}{1} = 1$
1-0-1	MLE = $P(H) = \frac{2}{3} = 0.667$

Fake Data Quiz

Calculate the MLE, with and without an extra fake data-point, for the following sequences of results:

- 1
- 0-0-1
- 1-0-0-1
- 1-1

Fake Data Summary

There are a couple of things that we should note here.

In general, adding the fake data points pulls everything towards 0.5, ‘smoothing’ the estimate. However, the first and last examples in the previous quiz showed that, the more data we get, the more we are willing to move away from 0.5. In the limiting case, as we see infinitely many ‘heads’, P(H) will indeed approach to 1.

In general, adding fake data gives better estimates in practice. The reason for this is that it is really quite reckless to suppose that a coin will always come up heads after just a single coin-flip. It is better to say that we have some evidence that heads may be more likely, but that we’re not yet convinced. The “not quite convinced” is the same as having a prior. These priors are called Dirichlet Priors.

More importantly, the technique of adding fake data to improve the estimator is called a Laplacian Estimator. When we have plenty of data, the Laplacian Estimator gives about the same result as the Maximum Likelihood Estimator, but when data is scarce, the Laplacian Estimator usually works much, much better.

Dice Example

Suppose we roll a die and get the results: 1-2-3-2.

What is the Maximum Likelihood Estimate and the Laplacian Estimate for each of the following?

- P(1)
- P(2)
- P(3)

Summary

In this section we introduced the Maximum Likelihood Estimator, and derived its mathematical formula:

$$MLE = \frac{1}{N}\sum X_i$$

This is a really simple formula, called the **empirical count**.

We also discussed the Laplacian Estimator that added k fake data points, one for each possible outcome, giving us the slightly more complicated formula:

$$LE = \frac{1}{N+k}(1+\sum X_i)$$

In both cases, N is the number of experiments, and k is the number of outcomes.

We the identified cases where the Laplacian Estimator gives much better results than the Maximum Likelihood Estimator. Specifically, cases where there isn't much data.

Averages

In this section, we will teach you about the three Ms in statistics: The 'mean', the 'median', and the 'mode'. These terms are really important to know. They are useful for looking at data, and are often confused. Let's begin with the mean.

Here is a list of house prices:

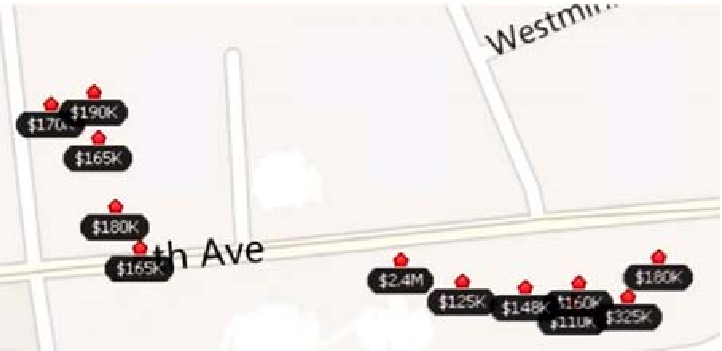
House prices
$190 k
$170 k
$165 k
$180 k
$165 k

The **mean** calculates the average of these prices using the formula we saw in the last section:

$$mean = \frac{1}{N}\sum X_i$$

In this case, the sum of the house prices is $870k, and there are 5 prices, so N = 5 and the mean is $174k. But why is the mean useful?

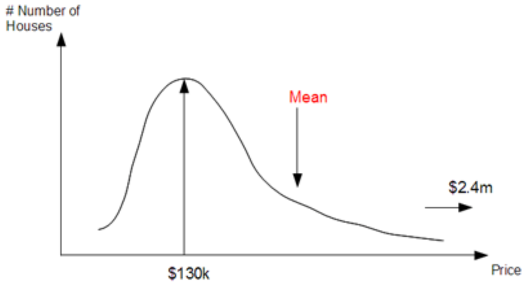
These prices are actually taken from a small area in Pittsburgh, Pennsylvania:



The prices in our area range from $165k to $190k. Our mean value of $174k really characterises this neighbourhood.

In a second neighbourhood on our map, most house prices are a little cheaper - $110k, $125k, $148k, and $160k. However, there are two outliers - $325k and $2.4 million. These outliers distort the mean house price for the second neighbourhood which is $492k.

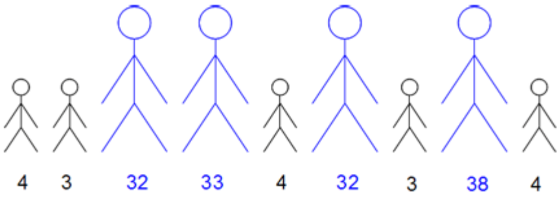
Clearly we ought to be suspicious of this number. Most of the homes are in the $100k range, but the mean doesn't reflect this. It gives the impression that the average house price in the neighbourhood is $492k, but that isn't a good description of reality. If we plotted the prices on a graph, we would get something like:



We have a peak around $130k, but there is a really long tail that goes all the way out to $2.4 million. The effect of this is to drag the mean away from the peak at $130k towards the outliers. It has a really strong impact on the mean.

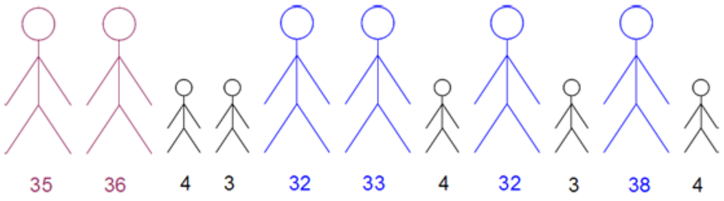
This is where the second M, the '**median**', comes into its own. The median is a different statistic that attempts to find the 'typical' value in a list, by identifying the one in the middle. To find the median, we just sort our list and pick the value in the middle. In this case, the median price is $160k, and if we look back at our map, this is a really good characterisation of house prices in this neighbourhood, and certainly a much better characterisation that that given by the mean value of $492k.

There is another limitation of the mean that may not be overcome by using the median. To illustrate this, we will consider a child's birthday party. This party has a number of children and some of their parents. We will say there are five children and four adults with ages as shown:



The mean age of the group is 17, although clearly this isn't very informative. There aren't any teenagers at this party. This is another case where the mean is giving us misleading statistics.

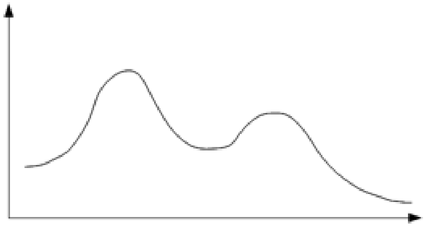
The median age for this group is 4. Is this meaningful? Well, not really. Suppose another two parents aged 35 and 36 turn up:



The median is now 32! A small change to the people at the party shifted the median from 4 to 32.

The '**mode**' is the value that occurs most often in a list. We already met modes when we looked at bar charts and talked about finding the most frequent bar. In this case, because the ages are discrete it is easy to see that the most frequent age at this party is 4, and that it was unchanged by the arrival of the two new parents.

So, the mode is a very useful statistic in cases where the data is **multi-modal**. Multi-modal distributions have curves with more than one peak:



This particular example is actually a **bi-modal** curve since it has exactly two peaks. The mode is the value of the highest peak.

Obviously, although the mean will always be determined precisely, sometimes there may be ties for the median or the mode. If we have an even number of elements, there will be two elements at the centre, and either could be chosen as the median. Equally, two elements could appear in the list the same number of times, and either one could be the mode. In these cases, we just assume that ties are broken at random. We can pick either number.

Three Averages Quiz

Calculate the mean, median and mode for the data:

5, 9, 100, 9, 97, 6, 9, 98, 9

Three Averages Quiz 2

Calculate the mean, median and mode for the data:

3, 9, 3, 8, 2, 9, 1, 9, 2, 4

Variance

This section is all about variance, and its close cousin standard deviation.

Consider a college student who has five close friends and five close family members. Their ages are as follows:

Friends:	17, 19, 18, 17, 19
Family:	7, 38, 4, 23, 18

The mean age of both groups is 18, but the friends are clustered very close to 18 which the family are much more widely distributed. In neither case does the mean, the median or the mode capture the spread of the data.

To capture the spread of the data we need to calculate the **variance**. To do this, we first normalise the data by subtracting the mean from each value:

Friends:	17	19	18	17	19
(normalised)	-1	1	0	-1	1

Family:	7	38	4	23	18
(normalised)	-11	20	-14	5	0

We should note that the mean of both normalised sequences is 0. We can see that the normalised values for the ages of the friends is much closer to zero than those for the family members, which shows that the spread of the family members' ages is much larger.

Now we calculate the squares of the normalised values:

Friends:	17	19	18	17	19
(normalised)	-1	1	0	-1	1
	1	1	0	1	1

Family:	7	38	4	23	18
(normalised)	-11	20	-14	5	0
	121	400	196	25	0

We now calculate the variance by adding the squared values and dividing by the number of values:

$$\text{variance} = \frac{1}{N}\sum (X_i - \mu)^2 \qquad (\mu \text{ is the mean})$$

Variance Quiz

Calculate the variance for both the friends and the family group.

Measuring Spread

The variance is a measure of how far the data is spread. It is really small if all the data is clustered close to the mean, and can be really large if data points occur a long way from the mean.

The variance, by its very nature computes in the quadratic. It is the average quadratic deviation from the mean:

$$\text{variance} = \frac{1}{N}\sum (X_i - \mu)^2$$

We can take the square root of the variance to obtain the standard deviation (σ) which is not a quadratic:

$$\text{standard deviation} = \sigma = \sqrt{\text{var}\,iance}$$

For our group of friends, the standard deviation is $\sqrt{0.8} = 0.894$, and for the family members it is $\sqrt{148.4} = 12.182$. The standard deviation provides a measure of the amount by which we might expect the age of an average member of the group to deviate from the mean.

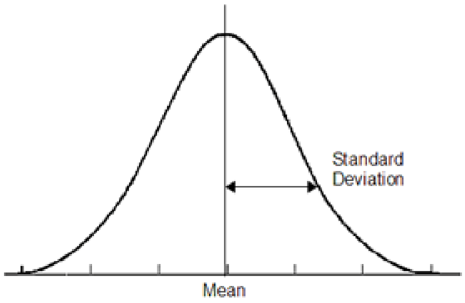
Standard Deviation Quiz

Calculate the mean, the variance, and the standard deviation for the following data sequences:

- 3, 4, 5, 6, 7
- 8, 9, 10, 11, 12
- 15, 20, 25, 30, 35
- 3, 3, 3, 3, 3
- 4

Formulae

Plotting the data onto a graph may give something like this:



The formulae that we used to calculate the values are:

$$\mu = \frac{1}{N}\sum_i X_i$$

$$\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$$

$$\sigma = \sqrt{\sigma^2}$$

A problem with these formulae is that they require two passes through the data.

First we have to go through the data to compute the mean. We do this by summing all the data and dividing by the total number of data items. For this we need to maintain two things:

1. the total number of data items, N (which we increment each time we see a new item)
2. the sum of all X_i, $\sum_i X_i$

Once we have done this, we have a value for the mean, μ, and now we have to go through and compute the variation for which we need to maintain $\sum (X_i - \mu)^2$.

It turns out that we can use a trick whereby, instead of maintaining $\sum (X_i - \mu)^2$, we can instead maintain $\sum X_i^2$, and so calculate the variation and standard deviation in a single pass.

We have $\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$

So, $\sigma^2 = \frac{1}{N}\sum (X_i - \mu)(X_i - \mu) \;=\; \frac{1}{N}\sum \left[X_i^2 - 2X_i\mu + \mu^2\right]$

$$\sigma^2 = \frac{1}{N}\sum X_i^2 - \frac{2\mu}{N}\sum X_i + \mu^2$$

Now, $\mu = \frac{1}{N}\sum_i X_i$

so, $\sigma^2 = \frac{1}{N}\sum X_i^2 - 2\mu^2 + \mu^2 \;\; = \;\; \frac{1}{N}\sum X_i^2 - \mu^2$

$$\sigma^2 = \frac{1}{N}\sum X_i^2 - \frac{1}{N^2}\left(\sum X_i\right)^2$$

Using this formula, the counters or statistics $\sum X_i$, $\sum X_i^2$, and N are all we need to calculate the variance in a single pass through the data.

Alternative Formula Quiz 1

Calculate $\sum X_i$, $\sum X_i^2$, and N for the data sequence:

3, 4, 5, 6, 7

Alternative Formula Quiz 2

If we plug these results into the formulae above, what values do we get for μ and σ^2?

We have covered a lot in this section. We introduced **variance**, which is the spread of the data squared, and then when on to explain **standard deviation**, which is the same, but without the square. We also now have a way to compute the values in a single pass through the data using only these running counters:

$\sum X_i$, $\sum X_i^2$, N

Raise Quiz

Suppose that we are considering giving all our employees a raise. We want to think about what the effect of the raise would be on the mean and standard deviation of the distribution of salaries within the company.

We are considering two types of raises:

1. A fixed amount of $1000
2. A relative raise of 20%

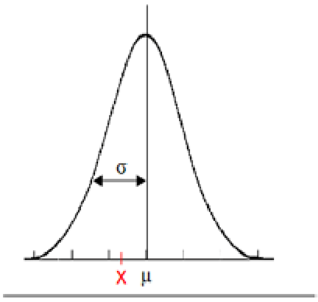
These will change the mean and standard deviation to new values which we will call μ' and σ'. The change will be either multiplicative or additive.

What are the multiplicative or additive factors in each case.

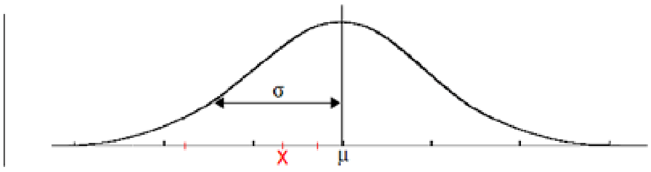
Standard Scores

We now want to introduce the concept of a standard score.

The basic idea is, that for any Gaussian, no matter what the mean and covariance are, we can state how far in or out a point, x, is. Let's think about an example. Suppose we have the point x on this Gaussian curve:



We can locate the corresponding point on a second Gaussian relative to the mean and the standard deviation, even if that curve is actually much wider with a different mean and standard deviation:



Mathematically, the standard score, Z, is defined as:

$$Z = \frac{X - \mu}{\sigma}$$

Standard Score Quiz

We have the data set:

3, 4, 5, 6, 7

The mean for this data set is 5, and the standard deviation is $\sqrt{2}$.

What is the standard score for x = 2, relative to the Gaussian that fits our data set?

Standard Score Quiz

What happens to Z if we change the new data point to x = 5?

Programming Estimators (Optional)

This section is completely optional, but it will provide the opportunity to program the techniques that we have been looking at throughout this unit.

Programming Mean Quiz

Create a Python function to return the mean for a list of data.

Programming Median Quiz

Write a Python function to return the median value of a list of data. Assume all the lists have an odd number of elements.

Programming Mode Quiz

Write a Python function to return the mode for a data set. Assume that if there are multiple modes you may return any one of them.

Programming Variance Quiz

Write a Python function to calculate the variance for a set of floating-point values. You have already written the code to calculate the mean, so you should use it.

Programming Standard Deviation Quiz

For the last programming quiz, we want you to calculate a Python function to calculate the standard deviation for a set of floating-point values.

Answers

Identify the Estimator Quiz

C

Likelihood Quiz

1. P(DATA) = ½ x ½ x ½ = 0.125
2. P(DATA) = ⅓ x ⅔ x ⅓ = 0.074
3. P(DATA) = ⅔ x ⅓ x ⅔ = 0.148
4. P(DATA) = 1 x 0 x 1 = 0

Weakness Quiz

Yes.

Weakness Quiz 2

Yes.

A fair coin always has P(H) = 0.5

For this to be true over N flips, the number of heads must be exactly N/2. This is just not possible for an odd number of observations like 111, so the coin will always look slightly biased.

Fake Data Quiz

Data	**MLE**	**MLE with extra data point**
1	P(H) = 1	P(H) = 0.667
0-0-1	P(H) = 0.333	P(H) = 0.4
1-0-0-1	P(H) = 0.5	P(H) = 0.5
1-1	P(H) = 1	P(H) = 0.75

Dice Example

For the MLE, the data sequence is: 1-2-3-2

For the Laplacian Estimator the data sequence becomes: 1-2-3-2-1-2-3-4-5-6

	MLE	Laplace
P(1)	$\frac{1}{4} = 0.25$	$\frac{2}{10} = 0.2$
P(2)	$\frac{2}{4} = 0.5$	$\frac{3}{10} = 0.3$
P(3)	$\frac{1}{4} = 0.25$	$\frac{2}{10} = 0.2$

Three Averages Quiz

Mean = 38
Median = 9
Mode = 9

Three Averages Quiz 2

Mean = 5
Median = 3 or 4
Mode = 9

Variance Quiz

Friends:	variance = 4/5 = 0.8
Family:	variance = 742/5 = 148.4

Standard Deviation Quiz

Data	Mean	Variance	Std Deviation
3, 4, 5, 6, 7	5	2	1.414
8, 9, 10, 11, 12	10	2	1.414
15, 20, 25, 30, 35	25	50	7.071
3, 3, 3, 3, 3	3	0	0
4	4	0	0

Alternative Formula Quiz 1

N = 5

$\sum X_i = 25$

$\sum X_i^2 = 135$

Alternative Formula Quiz 2

μ = 5

σ2 = 2

Raise Quiz

Fixed raise of $1000	Relative raise 20%
μ' = μ + $1000	μ' = 1.2 x μ
$\sigma^2 = \frac{1}{N}\sum\left(X_i + 1000 - (\mu + 1000)\right)^2$ σ' = σ	$\sigma^2 = \frac{1}{N}\sum\left(1.2 \times X_i - 1.2 \times \mu\right)^2$ σ' = 1.2 x σ

Standard Score Quiz

Z = -2.121

Standard Score Quiz 2

Z = 0

Programming Mean Quiz

```
def mean(data):
   return sum(data)/len(data)
```

Programming Median Quiz

```
def median(data):
   sdata = sorted(data)
   return sdata[int(len(data)/2)]
```

Programming Mode Quiz

```
def mode(data):
    mode = 0
    for item in data:
        if data.count(item) > data.count(mode):
            mode = item
    return mode
```

Programming Variance Quiz

```
def mean(data):
    return sum(data)/len(data)

def square(a):
    return a * a

def variance(data):
    ndata = []
    mu = mean(data)
    for i in data:
        ndata.append(square(i - mu))
    return sum(ndata)/len(ndata)
```

Programming Standard Deviation Quiz

```
def mean(data):
    return sum(data)/len(data)

def variance(data):
    mu=mean(data)
    return mean([(x-mu)**2 for x in data])

def stddev(data):
    return sqrt(variance(data))
```