Introduction to Machine Learning

A quick refresher course in probability theory

Third lecture, 23.02.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

1. **Hypothesis testing:**

   - I toss a coin ten times and get nine heads. How unlikely is that? Can we continue to believe that the coin is ***fair*** when it produces nine heads out of ten tosses?

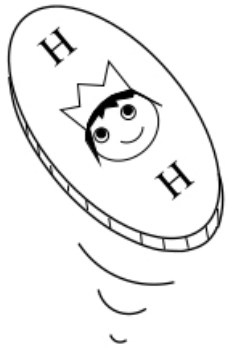
2. **Likelihood and estimation:**

   - Suppose we know that our random variable is (say) Binomial$(10, p)$, for some p, but we don't know the value of p. We will see how to ***estimate*** the value of p using maximum likelihood estimation.

Hypothesis testing

You have probably come across the idea of hypothesis tests, p-values, and significance in other courses. Common hypothesis tests include t-tests and chi-squared tests. However, hypothesis tests can be conducted in much simpler circumstances than these. The concept of the hypothesis test is at its easiest to understand with the Binomial distribution in the following example. All other hypothesis tests throughout statistics are based on the same idea.

Example: **Weird Coin?**



I toss a coin 10 times and get 9 heads. How weird is that?

What is 'weird'?

- Getting 9 heads out of 10 tosses: we'll call this *weird.*
- Getting 10 heads out of 10 tosses: *even more weird!*
- Getting 8 heads out of 10 tosses: *less weird.*
- Getting 1 head out of 10 tosses: *same as getting 9 tails out of 10 tosses: just as weird as 9 heads if the coin is fair.*
- Getting 0 heads out of 10 tosses: *same as getting 10 tails: more weird than 9 heads if the coin is fair.*

Hypothesis testing

Set of weird outcomes

If our coin is fair, the outcomes that are ***as weird or weirder*** than 9 heads are:

9 heads, 10 heads, 1 head, 0 heads.

So how weird is 9 heads or worse, if the coin is fair?

Define X =*#heads out of 10 tosses.*

Distribution of X, if the coin is fair: $X \sim \textit{Binomial}(n = 10, p = 0.5)$.

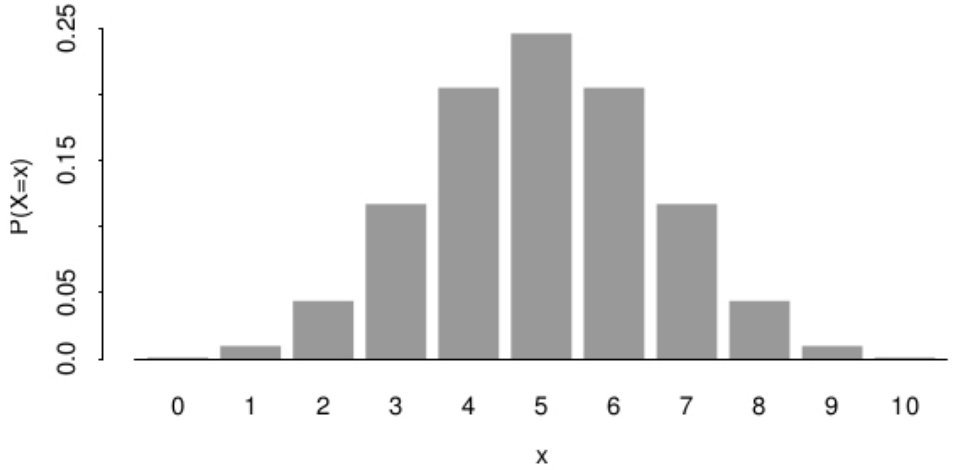
Probability of observing something at least as weird as 9 heads, if the coin is fair:

We can add the probabilities of all the outcomes that are ***at least as weird*** as 9 heads out of 10 tosses, assuming that the coin is fair.

$\mathbb{P}(X = 9)+\mathbb{P}(X = 10)+\mathbb{P}(X = 1)+\mathbb{P}(X = 0)$ *where* $X \sim \textit{Binomial}(10, 0.5)$.

Hypothesis testing

Probabilities for Binomial$(n = 10,\ p = 0.5)$



For $X \sim \text{Binomial}(10, 0.5)$, we have:

$$\begin{aligned}
\mathbb{P}(X = 9) + \mathbb{P}(X = 10) + \mathbb{P}(X = 1) + \mathbb{P}(X = 0) = & \\
& \binom{10}{9}(0.5)^9(0.5)^1 + \binom{10}{10}(0.5)^{10}(0.5)^0 + \\
& \binom{10}{1}(0.5)^1(0.5)^9 + \binom{10}{0}(0.5)^0(0.5)^{10} \\
= \quad & 0.00977 + 0.00098 + 0.00977 + 0.00098 \\
= \quad & 0.021.
\end{aligned}$$

Hypothesis testing

Is this weird?

Yes, it is quite weird. If we had a fair coin and tossed it 10 times, we would only expect to see something as extreme as 9 heads on about *2.1% of occasions.*

Is the coin fair?

Obviously, we can't say. It might be: after all, on 2.1% of occasions that you toss a fair coin 10 times, you do get something as weird as 9 heads or more.

However, 2.1% is a small probability, so it is still very unusual for a fair coin to produce something as weird as what we've seen. If the coin really was fair, it would be very unusual to get 9 heads or more.

We can deduce that, *EITHER we have observed a very unusual event with a fair coin, OR the coin is not fair.*

In fact, this gives us *some evidence that the coin is not fair.*

The value 2.1% *measures the strength of our evidence. The smaller this probability, the more evidence we have.*

Hypothesis testing

Formal hypothesis test

We now formalize the procedure above. Think of the steps:

- We have a question that we want to answer: *Is the coin fair?*

- There are two alternatives:
  1. *The coin is fair.*
  2. *The coin is not fair.*

- Our observed information is X, the number of heads out of 10 tosses. We write down the distribution of X *if the coin is fair:*
  $X \sim$ *Binomial*$(10, 0.5)$.

- We calculate the probability of observing something *AT LEAST AS EXTREME as our observation,* $X = 9$*, if the coin is fair: prob=0.021.*

- The probability is small (2.1%). We conclude that this is unlikely with a fair coin, so *we have observed some evidence that the coin is NOT fair.*

Null hypothesis and alternative hypothesis

We express the steps above as two competing hypotheses.

Null hypothesis: *the first alternative, that the coin IS fair.*

We expect to believe the null hypothesis unless we see convincing evidence that it is wrong.

Alternative hypothesis: *the second alternative, that the coin is NOT fair.*

In hypothesis testing, we often use this same formulation.

- The null hypothesis is *specific*.

  It specifies an exact distribution for our observation: $X \sim \textit{Binomial}(10, 0.5)$.

- The alternative hypothesis is *general*.

  It simply states that the null hypothesis is wrong. It does not say what the *right* answer is.

We use H_0 *and* H_1 to denote the null and alternative hypotheses respectively.

Null hypothesis and alternative hypothesis

The null hypothesis is H_0 : *the coin is fair.*
The alternative hypothesis is H_1 : *the coin is NOT fair.*

To set up the test, we write:

$$\textit{Number of heads, } X \sim \textit{Binomial}(10, p),$$

and

$$H_0 \ : \ p = 0.5$$
$$H_1 \ : \ p \neq 0.5.$$

Think of 'null hypothesis' as meaning the 'default': the hypothesis we will accept unless we have a good reason not to.

p-values

In the hypothesis-testing framework above, we always *measure evidence AGAINST the null hypothesis.*

That is, we believe that our coin is fair unless we see convincing evidence otherwise.

We measure the strength of evidence against H_0 using the *p-value.*

In the example above, the p-value was $p = 0.021$.

A p-value of 0.021 represents *quite strong evidence against the null hypothesis.*

It states that, if the null hypothesis is TRUE, we would only have *a 2.1% chance of observing something as extreme as 9 heads or tails.*

Some people might even see this as strong enough evidence to decide that the null hypothesis is not true, but this is generally an over-simplistic interpretation.

In general, the p-value is *the probability of observing something AT LEAST AS EXTREME AS OUR OBSERVATION, if H_0 is TRUE.*

This means that *SMALL p-values represent STRONG evidence against H_0.*

Small p-values mean Strong evidence.
Large p-values mean **Little** evidence.

p-values

Note: Be careful not to confuse the term p-value, which is 0.021 in our example, with the Binomial probability p. Our hypothesis test is designed to test whether the Binomial probability is $p = 0.5$. To test this, we calculate the p-value of 0.021 as a measure of the strength of evidence ***against*** the hypothesis that $p = 0.5$.

Interpreting the hypothesis test

There are different schools of thought about how a p-value should be interpreted.

- Most people agree that the p-value is a useful measure of the ***strength of evidence against the null hypothesis***. The smaller the p-value, the stronger the evidence against H_0.

- Some people go further and use an ***accept/reject framework.*** Under this framework, the null hypothesis H_0 should be *rejected* if the p-value is less than 0.05 (say), and *accepted* if the p-value is greater than 0.05.

- In this course we use the ***strength of evidence*** interpretation. The p-value measures how far out our observation lies in the tails of the distribution specified by H_0. We do not talk about accepting or rejecting H_0. This decision should usually be taken in the context of other scientific information.

  However, as a rule of thumb, we consider that p-values of 0.05 and less start to suggest that the null hypothesis is doubtful.

Statistical significance

You have probably encountered the idea of ***statistical significance*** in other courses.

Statistical significance refers to the p-value.

The result of a hypothesis test is ***significant at the 5% level*** if the p-value is *less than 0.05.*

This means that *the chance of seeing what we did see (9 heads), or more, is less than 5% if the null hypothesis is true.*

Saying the test is ***significant*** is a quick way of saying that there is evidence against the null hypothesis, usually at the 5% level.

Statistical significance

In the coin example, we can say that our test of $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$ *is significant at the 5% level, because the* p*-value is 0.021 which is* < 0.05.

This means:

- *we have some evidence that* $p \neq 0.5$.

It does ***not*** mean:

- the difference between p and 0.5 is *large*, or
- the difference between p and 0.5 is *important in practical terms.*

> Statistically significant means that *we have evidence, in OUR sample, that* p *is different from 0.5. It says NOTHING about the SIZE, or the IMPORTANCE, of the difference.*

"Substantial evidence of a difference", not "Evidence of a substantial difference."

Statistical significance

Beware!

The p-value gives the *probability of seeing something as weird as what we did see, if H_0 is true.*

This means that *5% of the time, we will get a p-value < 0.05 WHEN H_0 IS TRUE!!*

Similarly, about once in every thousand tests, we will get a p-value < 0.001, when H_0 is true!

A small p-value does NOT mean that H_0 is definitely wrong.

One-sided and two-sided tests

The test above is a *two-sided test.* This means that we considered it *just as weird to get 9 tails as 9 heads.*

If we had a good reason, ***before*** tossing the coin, to believe that the binomial probability could ***only*** be $= 0.5$ or > 0.5, i.e. that it would be ***impossible*** to have $p < 0.5$, then we could conduct a one-sided test: $H_0 : p = 0.5$ *versus* $H_1 : p > 0.5$.

This would have the effect of halving the resultant p-value.

Example: Presidents and deep-sea divers

Men in the class: would you like to have daughters? Then become a deep-sea diver, a fighter pilot, or a heavy smoker.

Would you prefer sons? Easy!
Just become a US president.

Numbers suggest that men in different professions tend to have more sons than daughters, or the reverse. Presidents have sons, fighter pilots have daughters. But is it real, or just chance? We can use hypothesis tests to decide.

The facts

- The 44 US presidents from George Washington to Barack Obama have had a total of 153 children, comprising 88 sons and only 65 daughters: a sex ratio of 1.4 sons for every daughter.
- Two studies of deep-sea divers revealed that the men had a total of 190 children, comprising 65 sons and 125 daughters: a sex ratio of 1.9 daughters for every son.

Example: Presidents and deep-sea divers

Could this happen by chance?

Is it possible that the men in each group *really had a 50-50 chance of producing sons and daughters?*

This is the same as the question in Section 2.2.

For the presidents: *If I tossed a coin 153 times and got only 65 heads, could I continue to believe that the coin was fair?*

For the divers: If I tossed a coin *190* times and got only *65* heads, could I continue to believe that the coin was fair?

Example: Presidents

Hypothesis test for the presidents

We set up the competing hypotheses as follows.

Let X be the number of daughters out of 153 presidential children.

Then $X \sim$ Binomial$(153, p)$, where p is the probability that each child is a daughter.

Null hypothesis: $H_0 : p = 0.5.$

Alternative hypothesis: $H_1 : p \neq 0.5.$

***p*-value:** *We need the probability of getting a result AT LEAST AS EXTREME as $X = 65$ daughters, if H_0 is true and p really is 0.5.*

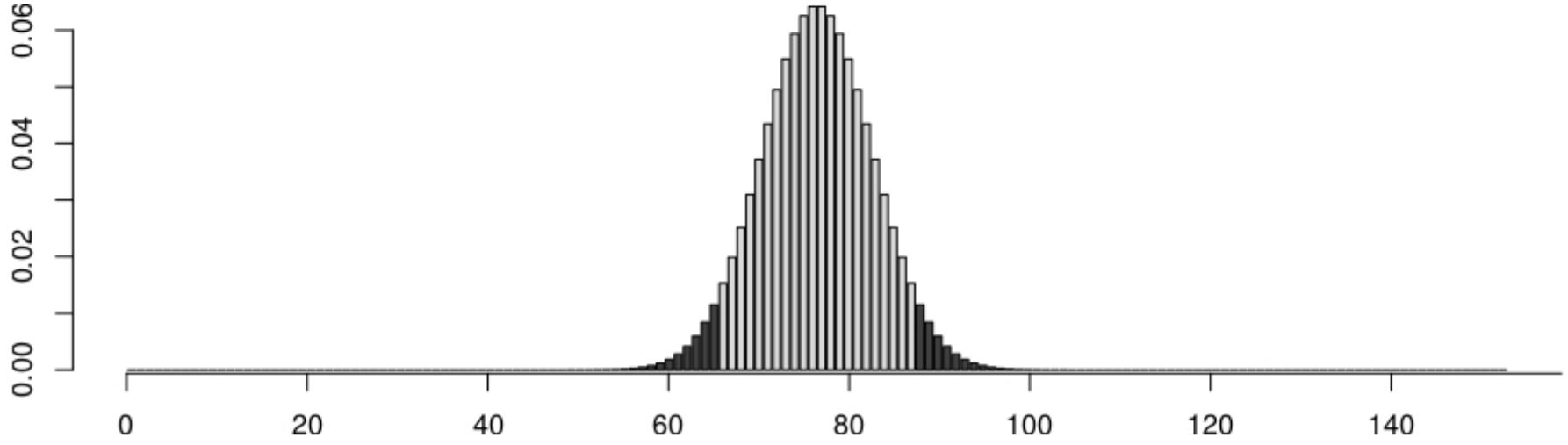
Example: Presidents

Which results are at least as extreme as $X = 65$?

$X = 0, 1, 2, \ldots, 65$, *for even fewer daughters.*

$X = (153 - 65), \ldots, 153$, *for too many daughters, because we would be just as surprised if we saw* ≤ 65 *sons, i.e.* $\geq (153 - 65) = 88$ *daughters.*

Probabilities for $X \sim \text{Binomial}(n = 153, p = 0.5)$



Example: Presidents

Calculating the p-value

The p-value for the president problem is given by

$\mathbb{P}(X \leq 65) + \mathbb{P}(X \geq 88)$ *where* $X \sim$ *Binomial*$(153, 0.5)$.

In principle, we could calculate this as

$$\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \ldots + \mathbb{P}(X = 65) + \mathbb{P}(X = 88) + \ldots + \mathbb{P}(X = 153)$$

$$= \binom{153}{0}(0.5)^0(0.5)^{153} + \binom{153}{1}(0.5)^1(0.5)^{152} + \ldots$$

This would take a lot of calculator time! Instead, we use a computer with a package such as R.

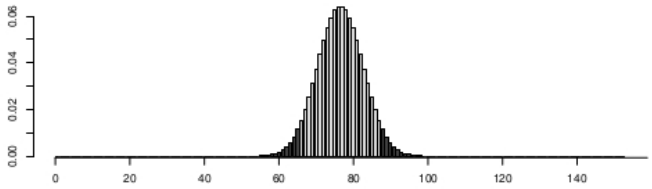
R command for the p-value

The R command for calculating the *lower-tail* p*-value for the Binomial*$(n = 153, p = 0.5)$ *distribution is*

```
pbinom(65, 153, 0.5).
```

Typing this in R gives:

```
> pbinom(65, 153, 0.5)
[1] 0.03748079
```



This gives us the *lower-tail* p*-value only:*
$\mathbb{P}(X \leq 65) = 0.0375$.

Example: Presidents

To get the overall p-value:

Multiply the lower-tail p-value by 2:

$$2 \times 0.0375 = 0.0750.$$

In *R*:

```
> 2 * pbinom(65, 153, 0.5)
[1] 0.07496158
```

This works because the upper-tail p-value, by definition, is always going to be the same as the lower-tail p-value. The upper tail gives us the probability of finding something *equally surprising* at the opposite end of the distribution.

Example: Presidents

Note: The R command `pbinom` is equivalent to the *cumulative distribution function* for the Binomial distribution:

$$\begin{aligned} \texttt{pbinom(65, 153, 0.5)} &= \mathbb{P}(X \leq 65) \qquad \textit{where } X \sim \textit{Binomial}(153, 0.5) \\ &= F_X(65) \qquad \textit{for } X \sim \textit{Binomial}(153, 0.5). \end{aligned}$$

The overall p-value in this example is $2 \times F_X(65)$.

Note: In the R command `pbinom(65, 153, 0.5)`, the order that you enter the numbers 65, 153, and 0.5 is important. If you enter them in a different order, you will get an error. An alternative is to use the longhand command `pbinom(q=65, size=153, prob=0.5)`, in which case you can enter the terms in any order.

Summary: are presidents more likely to have sons?

Back to our hypothesis test. Recall that X was the number of daughters out of 153 presidential children, and $X \sim \text{Binomial}(153, p)$, where p is the probability that each child is a daughter.

Null hypothesis: $H_0 : p = 0.5.$

Alternative hypothesis: $H_1 : p \neq 0.5.$

***p*-value:** $2 \times F_X(65) = 0.075.$

What does this mean?

The p-value of 0.075 means that, *if the presidents really were as likely to have daughters as sons, there would only be 7.5% chance of observing something as unusual as only 65 daughters out of the total 153 children.*

This is slightly unusual, but not very unusual.

Summary: are presidents more likely to have sons?

We conclude that *there is no real evidence that presidents are more likely to have sons than daughters. The observations are compatible with the possibility that there is no difference.*

Does this mean presidents are equally likely to have sons and daughters? *No: the observations are also compatible with the possibility that there is a difference. We just don't have enough evidence either way.*

Hypothesis test for the deep-sea divers

For the deep-sea divers, there were 190 children: 65 sons, and 125 daughters.

Let X be the *number of sons out of 190 diver children.*

Then $X \sim$ *Binomial*$(190, p)$, *where* p *is the probability that each child is a son.*

Note: We could just as easily formulate our hypotheses in terms of daughters instead of sons. Because `pbinom` is defined as a lower-tail probability, however, it is usually easiest to formulate them in terms of the *low* result (sons).

Null hypothesis: $H_0 : p = 0.5.$

Alternative hypothesis: $H_1 : p \neq 0.5.$

***p*-value:** *Probability of getting a result AT LEAST AS EXTREME as* $X = 65$ *sons, if* H_0 *is true and* p *really is 0.5.*

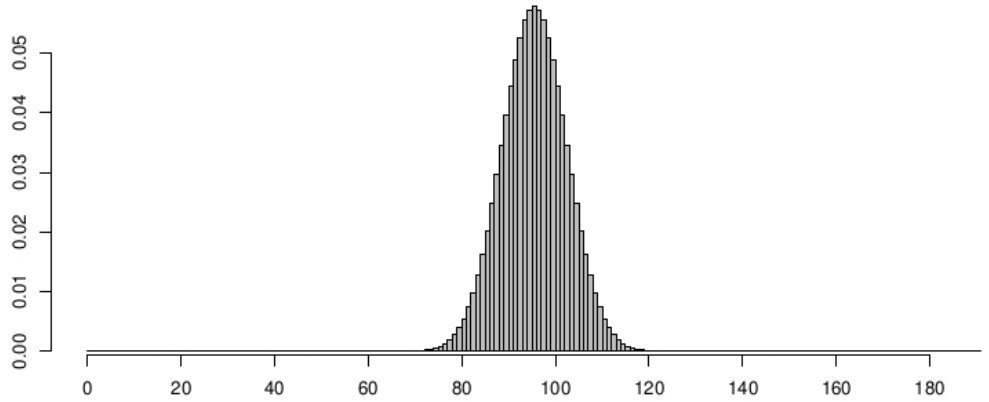
Results at least as extreme as $X = 65$ are:

$X = 0, 1, 2, \ldots, 65$, *for even fewer sons.*

$X = (190-65), \ldots, 190$, *for the equally surprising result in the opposite direction (too many sons).*

Hypothesis test for the deep-sea divers

Probabilities for $X \sim \text{Binomial}(n = 190, p = 0.5)$



R command for the p-value

p-*value* = 2×`pbinom(65, 190, 0.5)`.

Typing this in R gives:

```
> 2*pbinom(65, 190, 0.5)
[1] 1.603136e-05
```

This is 0.000016, or a little more than *one chance in 100 thousand.*

We conclude that *it is extremely unlikely that this observation could have occurred by chance, if the deep-sea divers had equal probabilities of having sons and daughters.*

We have *very strong evidence that deep-sea divers are more likely to have daughters than sons. The data are not really compatible with* H_0.

Likelihood and estimation

So far, the hypothesis tests have only told us whether the Binomial probability p *might be*, or *probably isn't*, equal to the value specified in the null hypothesis. They have told us nothing about the size, or potential importance, of the departure from H_0.

For example, for the deep-sea divers, we found that *it would be very unlikely to observe as many as 125 daughters out of 190 children if the chance of having a daughter really was* $p = 0.5$.

But what does this say about the *actual* value of p?

Remember the p-value for the test was 0.000016. Do you think that:

1. p could be as big as 0.8?

   No idea! The p-value does not tell us.

2. p could be as close to 0.5 as, say, 0.51?

   The test doesn't even tell us this much!
   If there was a huge sample size (number of children), we COULD get a p-value as small as 0.000016 even if the true probability was 0.51.

Common sense, however, gives us a hint. Because there were almost twice as many daughters as sons, my guess is that the probability of a having a daughter is something close to $p = 2/3$. We need some way of formalizing this.

Likelihood and estimation, example

I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 & \text{if the } i\text{th chosen ball is blue} \\ \\ 0 & \text{if the } i\text{th chosen ball is red} \end{cases}$$

Note that X_i's are i.i.d. and $X_i \sim Bernoulli(\frac{\theta}{3})$. After doing my experiment, I observe the following values for X_i's.

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1.$$

Thus, I observe 3 blue balls and 1 red balls.

1. For each possible value of θ, find the probability of the observed sample, $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$.
2. For which value of θ is the probability of the observed sample is the largest?

Likelihood and estimation, example

Since $X_i \sim Bernoulli(\frac{\theta}{3})$, we have

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3} & \text{for } x = 1 \\ \\ 1 - \frac{\theta}{3} & \text{for } x = 0 \end{cases}$$

Since X_i's are independent, the joint PMF of X_1, X_2, X_3, and X_4 can be written as

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

Therefore,

$$\begin{aligned} P_{X_1X_2X_3X_4}(1,0,1,1) &= \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} \\ &= \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right). \end{aligned}$$

Note that the joint PMF depends on θ, so we write it as $P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4; \theta)$. We obtain the values given in Table 8.1 for the probability of $(1, 0, 1, 1)$.

θ	$P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$
0	0
1	0.0247
2	0.0988
3	0

Table 8.1: Values of $P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$ for Example 8.1

The probability of observed sample for $\theta = 0$ and $\theta = 3$ is zero. This makes sense because our sample included both red and blue balls. From the table we see that the probability of the observed data is maximized for $\theta = 2$. This means that the observed data is most likely to occur for $\theta = 2$. For this reason, we may choose $\hat{\theta} = 2$ as our estimate of θ. This is called the maximum likelihood estimate (MLE) of θ.

https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php

Estimation

The process of using observations to suggest a value for a parameter is called *estimation.*

The value suggested is called the *estimate* of the parameter.

In the case of the deep-sea divers, we wish to estimate the probability p that the child of a diver is a daughter. The common-sense estimate to use is

$$p = \frac{\textit{number of daughters}}{\textit{total number of children}} = \frac{125}{190} = 0.658.$$

However, there are many situations where our common sense fails us. For example, what would we do if we had a regression-model situation (see Section 3.8) and wished to specify an alternative form for p, such as

$$p = \alpha + \beta \times (\text{diver age}).$$

How would we estimate the unknown intercept α and slope β, given known information on diver age and number of daughters and sons?

We need a general framework for estimation that can be applied to any situation. The most useful and general method of obtaining parameter estimates is the method of *maximum likelihood estimation.*

Likelihood

Likelihood is one of the most important concepts in statistics.
Return to the deep-sea diver example.

X is the *number of daughters out of 190 children.*

We know that $X \sim$ *Binomial*$(190, p)$,

and we wish to estimate the value of p.

The available data is the observed value of X: $\quad X = 125.$

Suppose for a moment that $p = 0.5$. What is the probability of observing $X = 125$?

When $X \sim$ *Binomial*$(190, 0.5)$,

$$\begin{aligned} \mathbb{P}(X = 125) &= \binom{190}{125} (0.5)^{125} (1 - 0.5)^{190-125} \\ &= 3.97 \times 10^{-6}. \end{aligned}$$

Not very likely!!

Likelihood

What about $p = 0.6$? What would be the probability of observing $X = 125$ if $p = 0.6$?

When $X \sim$ *Binomial*$(190, 0.6)$,

$$\begin{aligned}\mathbb{P}(X = 125) &= \binom{190}{125}(0.6)^{125}(1-0.6)^{190-125} \\ &= 0.016.\end{aligned}$$

This still looks quite unlikely, but it is almost 4000 times more likely than getting $X = 125$ *when* $p = 0.5$.

So far, we have discovered that *it would be thousands of times more likely to observe* $X = 125$ *if* $p = 0.6$ *than it would be if* $p = 0.5$.

This suggests that $p = 0.6$ *is a better estimate than* $p = 0.5$.

You can probably see where this is heading. If $p = 0.6$ is a better estimate than $p = 0.5$, what if we move p even closer to our common-sense estimate of 0.658?

When $X \sim$ *Binomial*$(190, 0.658)$,

$$\begin{aligned}\mathbb{P}(X = 125) &= \binom{190}{125}(0.658)^{125}(1-0.658)^{190-125} \\ &= 0.061.\end{aligned}$$

This is even more likely than for $p = 0.6$. *So* $p = 0.658$ *is the best estimate yet.*

Likelihood

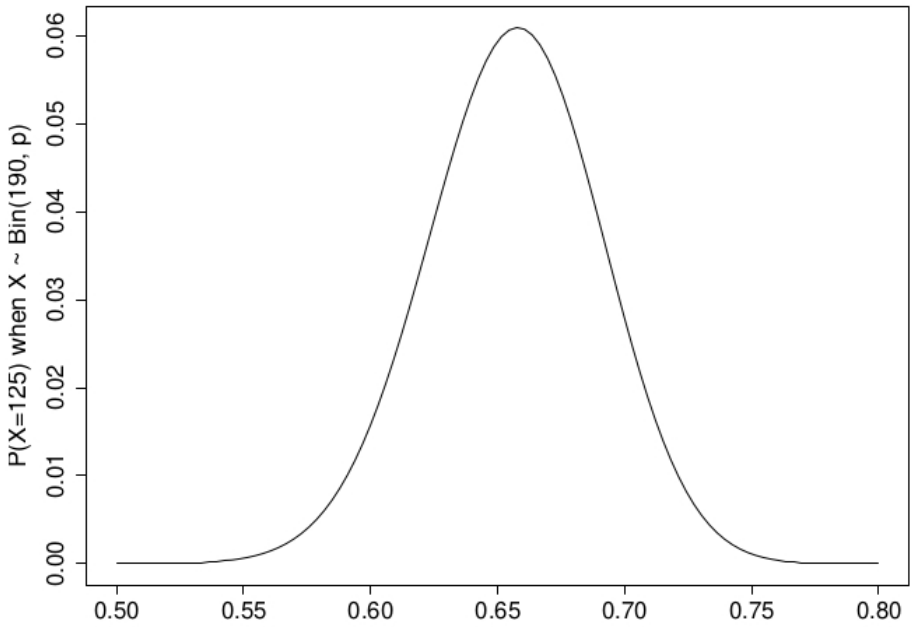
Can we do any better? What happens if we increase p a little more, say to $p = 0.7$?

When $X \sim$ *Binomial*$(190, 0.7)$,

$$\begin{aligned} \mathbb{P}(X = 125) &= \binom{190}{125} (0.7)^{125} (1 - 0.7)^{190-125} \\ &= 0.028. \end{aligned}$$

This has decreased from the result for $p = 0.658$, *so our observation of 125 is LESS likely under* $p = 0.7$ *than under* $p = 0.658$.

Overall, we can plot a graph showing ***how likely*** our observation of $X = 125$ is under each different value of p.



Likelihood

The graph reaches a *clear maximum. This is a value of* p *at which the observation* $X = 125$ *is MORE LIKELY than at any other value of* p.

This ***maximum likelihood*** value of p is our *maximum likelihood estimate.*

We can see that the maximum occurs somewhere close to our common-sense estimate of $p = 0.658$.

The likelihood function

Look at the graph we plotted overleaf:

Horizontal axis: *The unknown parameter,* p.

Vertical axis: *The probability of our observation,* $X = 125$, *under this value of* p.

This function is called the *likelihood function.*

It is a function of *the unknown parameter* p.

For our *fixed* observation $X = 125$, the likelihood function shows *how LIKELY the observation 125 is for every different value of* p.

The likelihood function is:

$$\begin{aligned} L(p) &= \mathbb{P}(X = 125) \text{ when } X \sim \text{Binomial}(190, p), \\ &= \binom{190}{125} p^{125} (1-p)^{190-125} \\ &= \binom{190}{125} p^{125} (1-p)^{65} \qquad \text{for } 0 < p < 1\,. \end{aligned}$$

The likelihood function

This function of p is the curve shown on the graph on page 55.

In general, if our observation were $X = x$ rather than $X = 125$, the likelihood function is *a function of* p *giving* $\mathbb{P}(X = x)$ *when* $X \sim$ *Binomial*$(190, p)$.

We write:

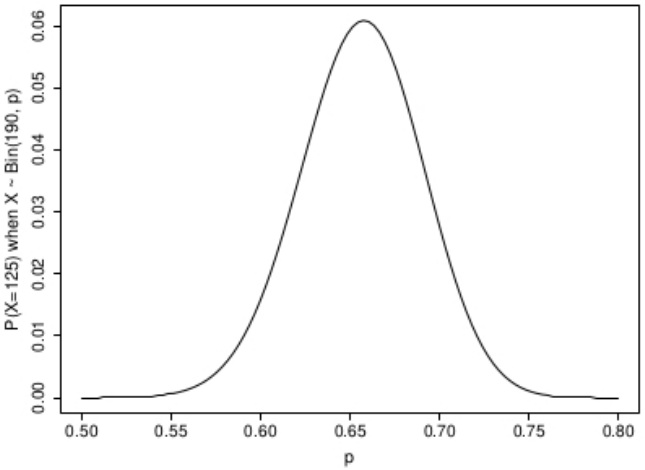
$$\begin{aligned} L(p\,;x) &= \mathbb{P}(X = x) \text{ when } X \sim \textit{Binomial}(190, p), \\ &= \binom{190}{x} p^x (1-p)^{190-x}\,. \end{aligned}$$

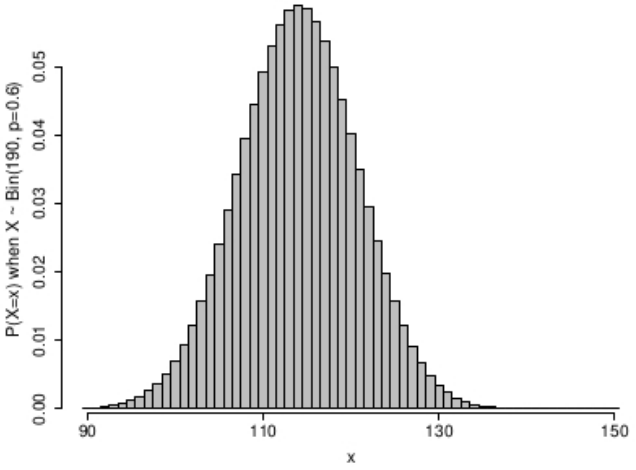
Difference between the likelihood function and the probability function

The likelihood function is *a probability of x, but it is a FUNCTION of p.*

The likelihood gives *the probability of a FIXED observation x, for every possible value of the parameter p.*

Compare this with the *probability function*, which is *the probability of every different value of x, for a FIXED value of p.*





Likelihood function, $L(p\,;x)$.
Function of p for fixed x.
Gives $\mathbb{P}(X = x)$ as p changes.
($x = 125$ here, but could be anything.)

Probability function, $f_X(x)$.
Function of x for fixed p.
Gives $\mathbb{P}(X = x)$ as x changes.
($p = 0.6$ here, but could be anything.)

Maximizing the likelihood

We have decided that a sensible parameter estimate for p is the maximum likelihood estimate: *the value of p at which the observation $X = 125$ is more likely than at any other value of p.*

We can find the maximum likelihood estimate using *calculus.*

The likelihood function is

$$L(p\,;125) = \binom{190}{125} p^{125}(1-p)^{65}.$$

Maximizing the likelihood

We wish to find the value of p that maximizes this expression.

To find the maximizing value of p, *differentiate the likelihood with respect to* p*:*

$$\frac{dL}{dp} = \binom{190}{125} \times \Big\{ 125 \times p^{124} \times (1-p)^{65} + p^{125} \times 65 \times (1-p)^{64} \times (-1) \Big\}$$

(Product Rule)

$$= \binom{190}{125} \times p^{124} \times (1-p)^{64} \Big\{ 125(1-p) - 65p \Big\}$$

$$= \binom{190}{125} p^{124}(1-p)^{64} \Big\{ 125 - 190p \Big\} \, .$$

The maximizing value of p occurs when

$$\frac{dL}{dp} = 0.$$

This gives:

$$\frac{dL}{dp} = \binom{190}{125} p^{124}(1-p)^{64} \Big\{ 125 - 190p \Big\} = 0$$

$$\Rightarrow \quad \Big\{ 125 - 190p \Big\} = 0$$

$$\Rightarrow \quad p = \frac{125}{190} = 0.658 \, .$$

Maximizing the likelihood

For the diver example, the maximum likelihood estimate of 125/190 is *the same as the common-sense estimate (page 53):*

$$p = \frac{\textit{number of daughters}}{\textit{total number of children}} = \frac{125}{190}.$$

This gives us confidence that the method of maximum likelihood is sensible.

The 'hat' notation for an estimate

It is conventional to write the estimated value of a parameter with a 'hat', like this: $\widehat{p}$.

For example,

$$\widehat{p} = \frac{125}{190}\,.$$

The correct notation for the maximization is:

$$\left.\frac{dL}{dp}\right|_{p=\widehat{p}} = 0 \quad \Rightarrow \quad \widehat{p} = \frac{125}{190}\,.$$

Summary of the maximum likelihood procedure

1. Write down the distribution of X in terms of the unknown parameter:

$$X \sim \textit{Binomial}(190, p).$$

2. Write down the observed value of X:

$$\textit{Observed data: } X = 125.$$

3. Write down the likelihood function for this observed value:

$$\begin{aligned} L(p\,;125) &= \mathbb{P}(X = 125) \textit{ when } X \sim \textit{Binomial}(190, p) \\ &= \binom{190}{125} p^{125}(1-p)^{65} \qquad \textit{for } 0 < p < 1. \end{aligned}$$

4. Differentiate the likelihood with respect to the parameter, and set to 0 for the maximum:

$$\frac{dL}{dp} = \binom{190}{125} p^{124}(1-p)^{64} \left\{ 125 - 190p \right\} = 0, \textit{ when } p = \widehat{p}.$$

This is the *Likelihood Equation.*

5. Solve for $\widehat{p}$: *From the graph, we can see that* $p = 0$ *and* $p = 1$ *are not maxima.*

$$\therefore \qquad \widehat{p} = \frac{125}{190}.$$

This is the *maximum likelihood estimate* (MLE) of p.

Summary of the maximum likelihood procedure

Verifying the maximum

Strictly speaking, when we find the maximum likelihood estimate using

$$\left.\frac{dL}{dp}\right|_{p=\widehat{p}} = 0,$$

we should verify that the result is a maximum (rather than a minimum) by showing that

$$\left.\frac{d^2L}{dp^2}\right|_{p=\widehat{p}} < 0\,.$$

In Stats 210, we will be relaxed about this. You will usually be told to assume that the MLE occurs in the interior of the parameter range. Where possible, it is always best to *plot the likelihood function, as on page 55.*

This confirms that the maximum likelihood estimate *exists and is unique.*

In particular, *care must be taken when the parameter has a restricted range like* $0 < p < 1$ *(see later).*

Estimators

For the example above, we had observation $X = 125$, and the maximum likelihood estimate of p was

$$\widehat{p} = \frac{125}{190}\,.$$

It is clear that we could follow through the same working with *any* value of X, which we can write as $X = x$, and we would obtain

$$\widehat{p} = \frac{x}{190}\,.$$

Exercise: Check this by maximizing the likelihood using x instead of 125.

This means that even *before* we have made our observation of X, we can provide *a RULE for calculating the maximum likelihood estimate once X is observed:*

Rule: *Let*

$$X \sim \textit{Binomial}(190,\ p).$$

Whatever value of X we observe, the maximum likelihood estimate of p will be

$$\widehat{p} = \frac{X}{190}\,.$$

Estimators

Note that this expression is now a *random variable: it depends on the random value of* X.

A random variable specifying how an estimate is calculated from an observation is called *an estimator.*

In the example above, *the maximum likelihood estimaTOR of* p *is*

$$\widehat{p} = \frac{X}{190}\,.$$

The maximum likelihood estimaTE of p*, once we have observed that* $X = x$*, is*

$$\widehat{p} = \frac{x}{190}\,.$$

General maximum likelihood estimator for Binomial(n, p)

Take *any* situation in which our observation X has the distribution

$$X \sim \textit{Binomial}(n, p),$$

where n is KNOWN and p is to be estimated.

We make a single observation $X = x$.

Follow the steps on page 59 to find the maximum likelihood estimator for p.

1. Write down the distribution of X in terms of the unknown parameter:

$$X \sim \textit{Binomial}(n, p).$$

(n is known.)

2. Write down the observed value of X:

Observed data: $X = x$.

3. Write down the likelihood function for this observed value:

$$\begin{aligned} L(p\,;x) &= \mathbb{P}(X = x) \textit{ when } X \sim \textit{Binomial}(n, p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \qquad \textit{for } 0 < p < 1. \end{aligned}$$

General maximum likelihood estimator for Binomial(n, p)

4. Differentiate the likelihood with respect to the parameter, and set to 0 for the maximum:

$$\frac{dL}{dp} = \binom{n}{x} p^{x-1}(1-p)^{n-x-1}\left\{ x - np \right\} = 0, \text{ when } p = \widehat{p}.$$

(Exercise)

5. Solve for $\widehat{p}$:

$$\widehat{p} = \frac{x}{n}\,.$$

This is the *maximum likelihood estimate of* p.

General maximum likelihood estimator for Binomial(n, p)

The maximum likelihood estimator of p is

$$\widehat{p} = \frac{X}{n}.$$

(Just replace the x in the MLE with an X, to convert from the estimate to the estimator.)

By deriving the general maximum likelihood estimator for *any* problem of this sort, we can plug in values of n and x to get an instant MLE for any Binomial(n, p) problem in which n is known.

Example: Recall the president problem in Section 2.3. Out of 153 children, 65 were daughters. Let p be the probability that a presidential child is a daughter. What is the maximum likelihood estimate of p?

Solution: *Plug in the numbers* $n = 153$, $x = 65$*:*

the maximum likelihood estimate is

$$\widehat{p} = \frac{x}{n} = \frac{65}{153} = 0.425\,.$$

Note: We showed in Section 2.3 that p *was not significantly different from* 0.5 *in this example.*
However, the MLE of p is definitely different from 0.5.
This comes back to the meaning of *significantly different* in the statistical sense.
Saying that p *is not significantly different from* 0.5 *just means that we can't DISTINGUISH any difference between* p *and* 0.5 *from routine sampling variability.*

Exercise 1

X_i is a random variable of having a girl in a 3-children family.

1) *Find the likelihood function when we have observed 4 family* $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$. *Hint*: $X_i \sim Binomial(3, \theta)$

2) *Find the maximum likelihood estimate of* θ

If $X_i \sim Binomial(3, \theta)$, then

$$P_{X_i}(x;\theta) = \binom{3}{x}\theta^x(1-\theta)^{3-x}$$

Thus,

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= P_{X_1}(x_1;\theta)P_{X_2}(x_2;\theta)P_{X_3}(x_3;\theta)P_{X_4}(x_4;\theta) \\ &= \binom{3}{x_1}\binom{3}{x_2}\binom{3}{x_3}\binom{3}{x_4}\theta^{x_1+x_2+x_3+x_4}(1-\theta)^{12-(x_1+x_2+x_3+x_4)}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$, we have

$$\begin{aligned} L(1, 3, 2, 2; \theta) &= \binom{3}{1}\binom{3}{3}\binom{3}{2}\binom{3}{2}\theta^8(1-\theta)^4 \\ &= 27 \qquad \theta^8(1-\theta)^4. \end{aligned}$$

Exercise 1

X_i is a random variable of having a girl in a 3-children family.
1) *Find the likelihood function when we have observed in 4 family* (x_1,x_2,x_3,x_4) = (1,3,2,2).
2) *Find the maximum likelihood estimate of* θ

In Example 8.8., we found the likelihood function as

$$L(1,3,2,2;\theta) = 27 \qquad \theta^8(1-\theta)^4.$$

To find the value of θ that maximizes the likelihood function, we can take the derivative and set it to zero. We have

$$\frac{dL(1,3,2,2;\theta)}{d\theta} = 27\big[\qquad 8\theta^7(1-\theta)^4 - 4\theta^8(1-\theta)^3\big].$$

Thus, we obtain

$$\hat{\theta}_{ML} = \frac{2}{3}.$$

Exercise 2

Xi~*Exponential*(θ) *and we have observed* (x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)

1) *Find the likelihood function.*

2) *Find the maximum likelihood estimate of* θ. *Hint*:

$$f_{X_i}(x;\theta) = \theta e^{-\theta x} u(x),$$

where $u(x)$ is the unit step function, i.e., $u(x) = 1$ for $x \geq 0$ and $u(x) = 0$ for $x < 0$. Thus, for $x_i \geq 0$, we can write

If $X_i \sim Exponential(\theta)$, then

$$f_{X_i}(x;\theta) = \theta e^{-\theta x} u(x),$$

where $u(x)$ is the unit step function, i.e., $u(x) = 1$ for $x \geq 0$ and $u(x) = 0$ for $x < 0$. Thus, for $x_i \geq 0$, we can write

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) f_{X_3}(x_3;\theta) f_{X_4}(x_4;\theta) \\ &= \theta^4 e^{-(x_1+x_2+x_3+x_4)\theta}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$, we have

$$L(1.23, 3.32, 1.98, 2.12; \theta) = \theta^4 e^{-8.65\theta}.$$

Exercise 2

Xi~*Exponential*(θ) *and we have observed* (x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)

1) *Find the likelihood function.*

2) *Find the maximum likelihood estimate of* θ.

In Example 8.8., we found the likelihood function as

$$L(1.23, 3.32, 1.98, 2.12; \theta) = \theta^4 e^{-8.65\theta}.$$

Here, it is easier to work with the log likelihood function, $\ln L(1.23, 3.32, 1.98, 2.12; \theta)$. Specifically,

$$\ln L(1.23, 3.32, 1.98, 2.12; \theta) = 4 \ln \theta - 8.65\theta.$$

By differentiating, we obtain

$$\frac{4}{\theta} - 8.65 = 0,$$

which results in

$$\hat{\theta}_{ML} = 0.46$$

It is worth noting that technically, we need to look at the second derivatives and endpoints to make sure that the values that we obtained above are the maximizing values. For this example, it turns out that the obtained values are indeed the maximizing values.

https://www.probabilitycourse.com/chapter8/8_2_3_max_likelihood_estimation.php

Maximum Log-likelihood Estimation

Example 1. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as:

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

Maximum Log-likelihood Estimation

Here are some standard terms we will use as we do statistics.

- Experiment: Flip the coin 100 times and count the number of heads.
- Data: The data is the result of the experiment. In this case it is '55 heads'.
- Parameter(s) of interest: We are interested in the value of the unknown parameter p.
- Likelihood, or likelihood function: this is $P(\text{data}\,|\,p)$. Note it is a function of both the data and the parameter p. In this case the likelihood is

$$P(55 \text{ heads}\,|\,p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Notes: **1.** The likelihood $P(\text{data}\,|\,p)$ changes as the parameter of interest p changes.

2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(\text{data}\,|\,p)$ for $P(p\,|\,\text{data})$. We know from our earlier work with Bayes' theorem that $P(\text{data}\,|\,p)$ and $P(p\,|\,\text{data})$ are usually very different.

Maximum Log-likelihood Estimation

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d}{dp}P(\text{data } |p) = \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Solving this for p we get

$$\begin{aligned} 55p^{54}(1-p)^{45} &= 45p^{55}(1-p)^{44} \\ 55(1-p) &= 45p \\ 55 &= 100p \\ \text{the MLE is } \hat{p} &= .55 \end{aligned}$$

Note: **1.** The MLE for p turned out to be exactly the fraction of heads we saw in our data.

2. The MLE is computed from the data. That is, it is a statistic.

3. Officially you should check that the critical point is indeed a maximum. You can do this with the second derivative test.

Maximum Log-likelihood Estimation

The **likelihood function** is defined as

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}), \tag{4.5}$$

and is regarded as a function of $\boldsymbol{\theta}$, for fixed $\mathbf{x}$, as opposed to the interpretation of the density function f. Quite often we use directly the **log-likelihood function**

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \mathbf{x}) = \log(L(\boldsymbol{\theta}; \mathbf{x})) = \log(f(\mathbf{x}; \boldsymbol{\theta})). \tag{4.6}$$

Example 2. Redo the previous example using log likelihood.

answer: We had the likelihood $P(55 \text{ heads} \mid p) = \binom{100}{55} p^{55}(1-p)^{45}$. Therefore the log likelihood is

$$\ln(P(55 \text{ heads} \mid p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\begin{aligned}
\frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right] \\
&= \frac{55}{p} - \frac{45}{1-p} = 0 \\
&\Rightarrow 55(1-p) = 45p \\
&\Rightarrow \hat{p} = .55
\end{aligned}$$

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

Maximum Log-likelihood Estimation

Example 6. Capture/recapture method

The capture/recapture method is a way to estimate the size of a population in the wild. The method assumes that each animal in the population is equally likely to be captured by a trap.

Suppose 10 animals are captured, tagged and released. A few months later, 20 animals are captured, examined, and released. 4 of these 20 are found to be tagged. Estimate the size of the wild population using the MLE for the probability that a wild animal is tagged.

answer: Our unknown parameter n is the number of animals in the wild. Our data is that 4 out of 20 recaptured animals were tagged (and that there are 10 tagged animals). The likelihood function is

$$P(\text{data} \mid n \text{ animals}) = \frac{\binom{n-10}{16}\binom{10}{4}}{\binom{n}{20}}$$

(The numerator is the number of ways to choose 16 animals from among the $n-10$ untagged ones times the number of was to choose 4 out of the 10 tagged animals. The denominator is the number of ways to choose 20 animals from the entire population of n.) We can use R to compute that the likelihood function is maximized when $n = 50$. This should make some sense. It says our best estimate is that the fraction of all animals that are tagged is 10/50 which equals the fraction of recaptured animals which are tagged.

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

Maximum Log-likelihood Estimation

Example 7. Hardy-Weinberg. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1-\theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Suppose we test a random sample of people and find that k_1 are AA, k_2 are Aa, and k_3 are aa. Find the MLE of θ.

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

Maximum Log-likelihood Estimation

answer: The likelihood function is given by

$$P(k_1, k_2, k_3 \mid \theta) = \binom{k_1 + k_2 + k_3}{k_1} \binom{k_2 + k_3}{k_2} \binom{k_3}{k_3} \theta^{2k_1} (2\theta(1-\theta))^{k_2} (1-\theta)^{2k_3}.$$

So the log likelihood is given by

$$\text{constant} + 2k_1 \ln(\theta) + k_2 \ln(\theta) + k_2 \ln(1-\theta) + 2k_3 \ln(1-\theta)$$

We set the derivative equal to zero:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1-\theta} = 0$$

Solving for θ, we find the MLE is

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3},$$

which is simply the fraction of A alleles among all the genes in the sampled population.

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

MLE for n observation, example 1

EXAMPLE Suppose a sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted λ, so that

$$f_X(x;\theta) \equiv f_X(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, ...$$

for some $\lambda > 0$. To estimate λ by maximum likelihood, proceed as follows.

STEP 1 Calculate the likelihood function $L(\lambda)$.

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \right\} = \frac{\lambda^{x_1+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

for $\lambda \in \Theta = R^+$.

STEP 2 Calculate the log-likelihood $\log L(\lambda)$.

$$\log L(\lambda) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

STEP 3 Differentiate $\log L(\lambda)$ with respect to λ, and equate the derivative to zero to find the m.l.e..

$$\frac{d}{d\lambda}\{\log L(\lambda)\} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - n = 0 \Rightarrow \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Thus the maximum likelihood estimate of λ is $\widehat{\lambda} = \bar{x}$

STEP 4 Check that the second derivative of $\log L(\lambda)$ with respect to λ is negative at $\lambda = \widehat{\lambda}$.

$$\frac{d^2}{d\lambda^2}\{\log L(\lambda)\} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \quad \text{at } \lambda = \widehat{\lambda}$$

MLE for n observation, example 1

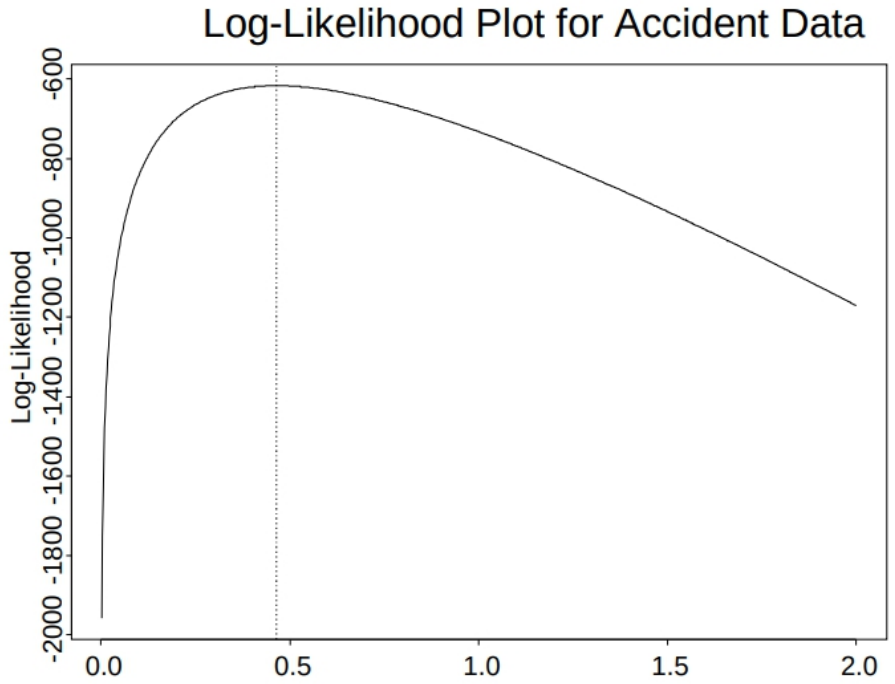
EXAMPLE: The following data are the observed frequencies of occurrence of domestic accidents: we have $n = 647$ data as follows

Number of accidents	Frequency
0	447
1	132
2	42
3	21
4	3
5	2

.

The estimate of λ if a Poisson model is assumed is

$$\widehat{\lambda}_{ML} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(447 \times 0) + (132 \times 1) + (42 \times 2) + (21 \times 3) + (3 \times 4) + (2 \times 5)}{647} = 0.465$$



https://www.ma.imperial.ac.uk/~das01/MyWeb/M2S1Previous/2004/Handouts/Mle.pdf

MLE for n observation, example 2

Example 3. Light bulbs

Suppose that the lifetime of *Badger* brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

MLE for n observation, example 2

answer: We need to be careful with our notation. With five different values it is best to use subscripts. Let X_j be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has pdf $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$. We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of the individual densities:

$$f(x_1, x_2, x_3, x_4, x_5 \,|\, \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

So the likelihood and log likelihood functions with this data are

$$f(2, 3, 1, 3, 4 \,|\, \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 \,|\, \lambda) = 5 \ln(\lambda) - 13\lambda$$

Finally we use calculus to find the MLE:

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \ \Rightarrow\ \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Note: **1.** In this example we used an uppercase letter for a random variable and the corresponding lowercase letter for the value it takes. This will be our usual practice.

2. The MLE for λ turned out to be the reciprocal of the sample mean $\bar{x}$, so $X \sim \exp(\hat{\lambda})$ satisfies $E(X) = \bar{x}$.

MLE for 2 parameters, example 1

```
# Plot the normal distribution function with mean = 0, std = 1
f <- function(x)(1/sqrt(2*pi*1)*exp(-x^2/2))
curve(f,from=-10, to=10)
```

Example 4. Normal distributions
Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a $\mathrm{N}(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Find the maximum likelihood estimate for the pair (μ, σ^2).

answer: Let's be precise and phrase this in terms of random variables and densities. Let uppercase $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$ random variables, and let lowercase x_i be the value X_i takes. The density for each X_i is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1, \ldots, x_n \,|\, \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1, \ldots, x_n$, the likelihood and log likelihood are

$$f(x_1, \ldots, x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1, \ldots, x_n|\mu, \sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

MLE for 2 parameters, example 1

Since $\ln(f(x_1,\ldots,x_n|\mu,\sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1,\ldots,x_n|\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \;\Rightarrow\; \sum_{i=1}^{n} x_i = n\mu \;\Rightarrow\; \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

$$\frac{\partial f(x_1,\ldots,x_n|\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^{n} \frac{1}{n}(x_i-\hat{\mu})^2 = \sum_{i=1}^{n} \frac{1}{n}(x_i-\overline{x})^2 && = \text{the variance of the data.} \end{aligned}$$