Chapter 2: Foundations of Statistical Inference

2.1 Introduction

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. The word ***inference*** means 'conclusions' or 'decisions'. Statistical inference is about drawing conclusions and making decisions based on observed data.

Data, or observations, typically arise from some *underlying process.* It is the underlying process we are interested in, not the observations themselves. Sometimes we call the underlying process the *population* or *mechanism* of interest.

The data are only a *sample* from this population or mechanism. We cannot possibly observe every outcome of the process, so we have to make do with the sample that we have observed.

The data give us *imperfect insight* into the population of interest. The role of statistical inference is *to use this imperfect data to draw conclusions about the population of interest, while simultaneously giving an honest reflection of the uncertainty in our conclusions.*

Example 1: Tossing a coin.

- ***Population:*** *all possible tosses of this coin.*
- ***Sample:*** *a small number of observed tosses, e.g. 10 observed tosses.*
- ***What do we want to make inference about?*** We might be interested in the probability of getting a Head on each toss. In particular, we might be interested in whether the coin is fair ($\mathbb{P}(\text{Head}) = 0.5$) or has been fiddled.

Example 2: Political polling: how many people will vote for the NZ Labour Party?

- ***Population:*** *all eligible voters in New Zealand.*
- ***Sample:*** *a random sample of voters, e.g. 1000.*
- ***What do we want to make inference about?*** We want to know the support for Labour among ***all*** voters, but this is too expensive to carry out except on election-night itself. Instead we aim to ***deduce*** the support for Labour by asking a smaller number of voters, while simultaneously reporting upon our uncertainty (margin of error).

In the next two chapters we meet several important concepts in statistical inference. We will illustrate them with ***discrete random variables***, then introduce ***continuous random variables*** in Chapter 4 and show how the same ideas still apply.

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.

3. **Expectation and variance of a random variable:**

   - The ***expectation*** of a random variable is the value it takes ***on average.***
   - The ***variance*** of a random variable measures how much the random variable ***varies about its average.***

   These are used to report how accurate and reliable our ***estimation procedure*** is. Does it give the right answer ***on average?*** How much does it ***vary*** about its average?

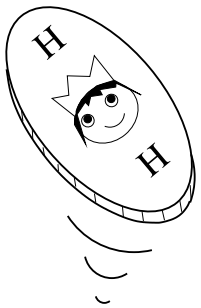
4. **Modelling:**

   - We have a situation in real life that we know is random. But what does the randomness ***look*** like? Is it highly variable, or little variability? Does it sometimes give results much *higher* than average, but never give results much *lower* (long-tailed distribution)? We will see how different probability distributions are suitable for different circumstances. Choosing a probability distribution to fit a situation is called ***modelling.***

2.2 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.*

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.*

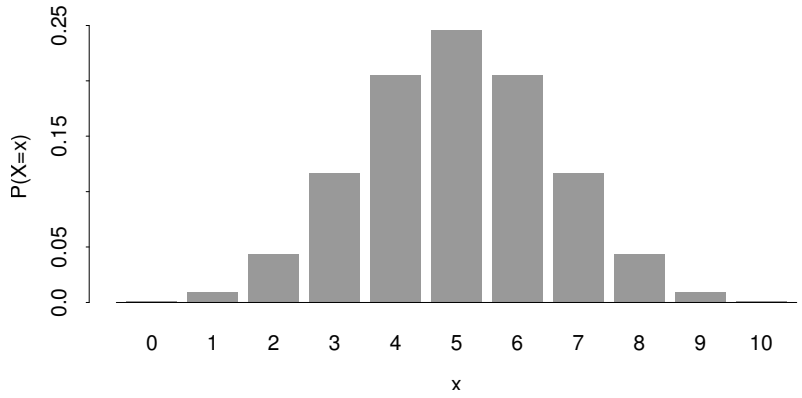
Distribution of X, if the coin is fair: $X \sim$ *Binomial*$(n = 10, p = 0.5)$.

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.

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

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



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

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

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.*

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: $X \sim$ *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.

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

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5. \end{aligned}$$

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.

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.

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."

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.

2.3 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.

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?

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: $H_1 : p \neq 0.5$.

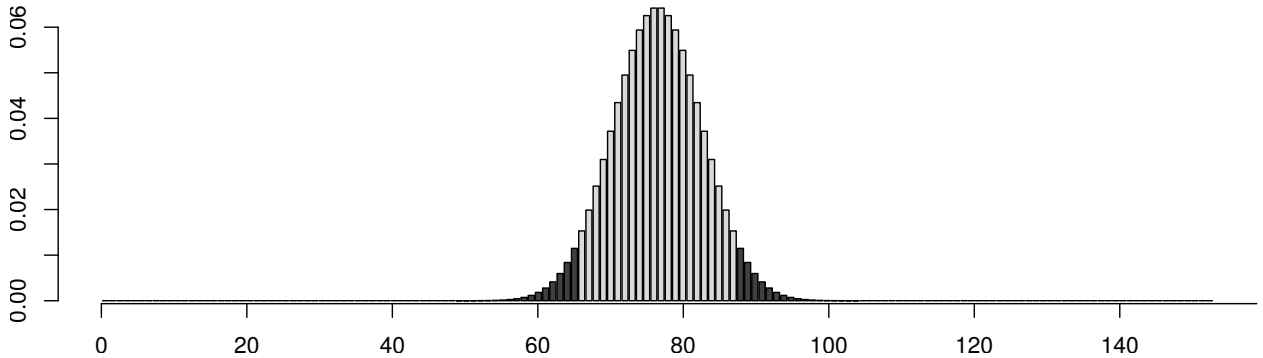
***p*-value:** *We need the probability of getting a result AT LEAST AS EXTREME as $X = 65$ daughters, if H_0 is true and p really is 0.5.*

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 $X \sim$ Binomial$(n = 153, p = 0.5)$



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.

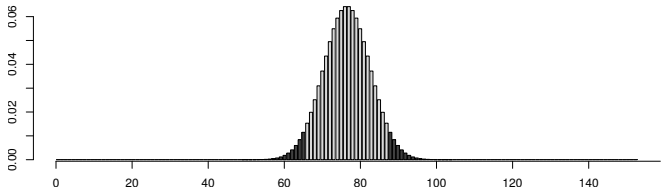
R command for the p-value

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:*
$\mathbb{P}(X \leq 65) = 0.0375$.

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.

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

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

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.

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: $H_1 : p \neq 0.5$.

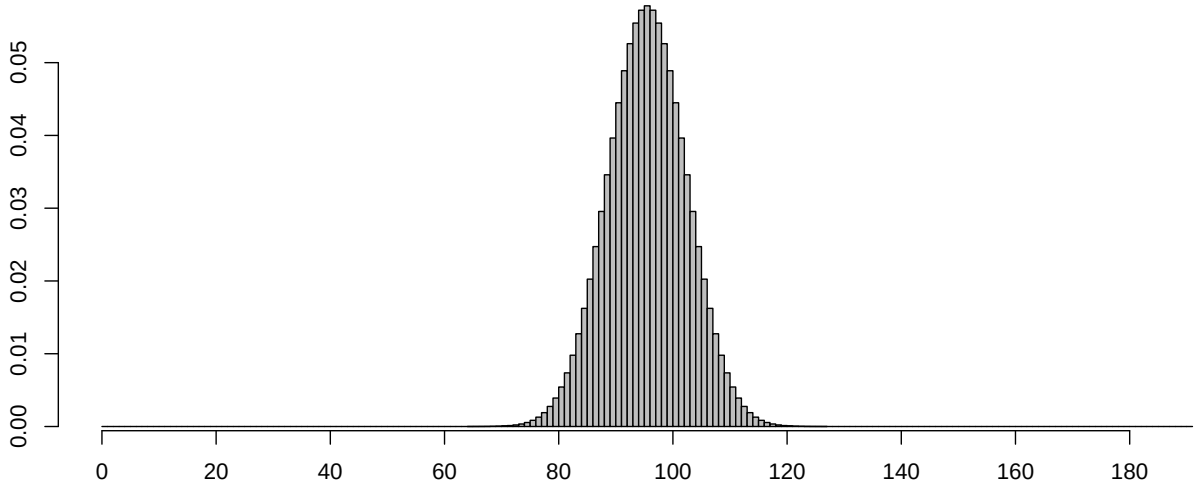
***p*-value:** *Probability of getting a result AT LEAST AS EXTREME as* $X = 65$ *sons, if* H_0 *is true and* p *really is 0.5.*

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).*

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.

What next?

p-values are often badly used in science and business. They are regularly treated as the end point of an analysis, after which no more work is needed. Many scientific journals insist that scientists quote a p-value with every set of results, and often only p-values less than 0.05 are regarded as 'interesting'. The outcome is that some scientists do every analysis they can think of until they finally come up with a p-value of 0.05 or less.

A good statistician will recommend a different attitude. *It is very rare in science for numbers and statistics to tell us the full story.*

Results like the p-value should be regarded as an investigative *starting point*, rather than the final conclusion. *Why* is the p-value small? What possible *mechanism* could there be for producing this result?

If you were a medical statistician and you gave me a p-value, I would ask you for a mechanism.

Don't accept that Drug A is better than Drug B *only* because the p-value says so: find a biochemist who can explain what Drug A does that Drug B doesn't. Don't accept that sun exposure is a cause of skin cancer on the basis of a p-value alone: find a mechanism by which skin is damaged by the sun.

Why might divers have daughters and presidents have sons?

Deep-sea divers are thought to have more daughters than sons because the underwater work at high atmospheric pressure lowers the level of the hormone testosterone in the men's blood, which is thought to make them more likely to conceive daughters. For the presidents, your guess is as good as mine ...

2.4 Example: Birthdays and sports professionals

Have you ever wondered what makes a professional sports player? Talent? Dedication? Good coaching?

Or is it just that they happen to have the right birthday. . . ?

The following text is taken from Malcolm Gladwell's book *Outliers*. It describes the play-by-play for the first goal scored in the 2007 finals of the Canadian ice hockey junior league for star players aged 17 to 19. The two teams are the Tigers and Giants. There's one slight difference . . . instead of the players' names, we're given their birthdays.

> *March 11 starts around one side of the Tigers' net, leaving the puck for his teammate January 4, who passes it to January 22, who flips it back to March 12, who shoots point-blank at the Tigers' goalie, April 27. April 27 blocks the shot, but it's rebounded by Giants' March 6. He shoots! Tigers defensemen February 9 and February 14 dive to block the puck while January 10 looks on helplessly. March 6 scores!*

Notice anything funny?

Here are some figures. There were 25 players in the Tigers squad, born between 1986 and 1990. Out of these 25 players, 14 of them were born in January, February, or March. Is it believable that this should happen by chance, or do we have evidence that there is a birthday-effect in becoming a star ice hockey player?

Hypothesis test

Let X be the number of the 25 players who are born from January to March.
We need to set up hypotheses of the following form:

Null hypothesis: H_0 : *there is no birthday effect.*

Alternative hypothesis: H_1 : *there is a birthday effect.*

What is the distribution of X under H_0 and under H_1?

Under H_0, there is no birthday effect. So the probability that each player has a birthday in Jan to March is about $1/4$.

(3 months out of a possible 12 months).

Thus the distribution of X under H_0 is $X \sim$ ***Binomial***$(25, 1/4)$.

Under H_1, there is a birthday effect, so $p \neq 1/4$.

Our formulation for the hypothesis test is therefore as follows.

Number of Jan to March players, $X \sim$ ***Binomial***$(25, p)$.

Null hypothesis: $H_0 : p = 0.25$.

Alternative hypothesis: $H_1 : p \neq 0.25$.

Our observation:

The observed proportion of players born from Jan to March is $14/25 = 0.56$.

This is *more than the* 0.25 *predicted by* H_0.

Is it sufficiently greater than 0.25 to provide evidence against H_0?

Just using our intuition, we can make a guess, but we might be wrong. The answer also depends on the sample size (25 in this case). We need the p-value to measure the evidence properly.

***p*-value:** *Probability of getting a result AT LEAST AS EXTREME as* $X = 14$ *Jan to March players, if* H_0 *is true and* p *really is 0.25.*

Results at least as extreme as $X = 14$ are:

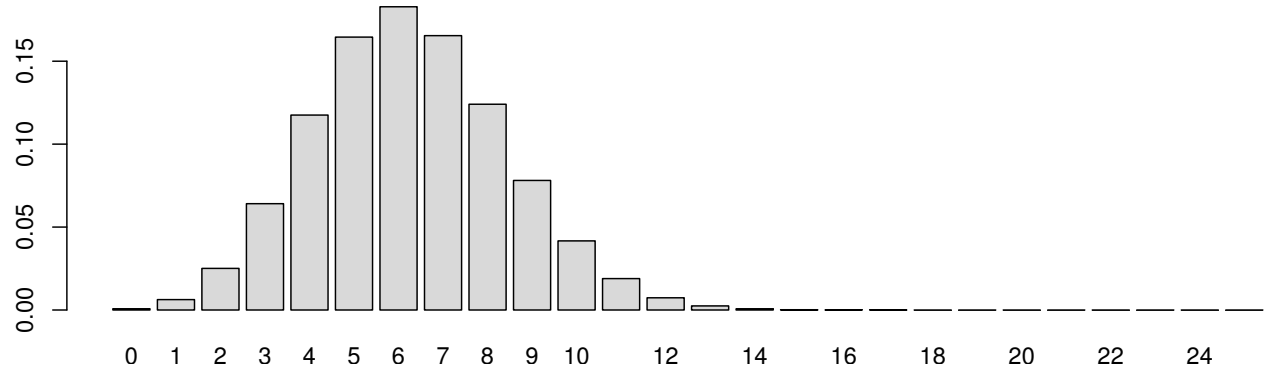
Upper tail: $X = 14, 15, \ldots, 25$*, for even more Jan to March players.*

Lower tail: an equal probability in the opposite direction, for too few Jan to March players.

Note: We do not need to calculate the values corresponding to our lower-tail p-value. It is more complicated in this example than in Section 2.3, because we do not have Binomial probability $p = 0.5$. In fact, the lower tail probability lies somewhere between 0 and 1 player, but it cannot be specified exactly.

We get round this problem for calculating the p-value by *just multiplying the upper-tail p-value by 2.*

Probabilities for $X \sim \text{Binomial}(n = 25, p = 0.25)$



R command for the p-value

We need *twice the UPPER-tail p-value:*

p-value $= 2 \times (1 - $`pbinom(13, 25, 0.25)`$)$.
(Recall $\mathbb{P}(X \geq 14) = 1 - \mathbb{P}(X \leq 13)$.*)*

Typing this in R gives:

```
2*(1-pbinom(13, 25, 0.25))
[1] 0.001831663
```

This p-value is *very small.*

It means that *if there really was no birthday effect, we would expect to see results as unusual as 14 out of 25 Jan to March players less than 2 in 1000 times.*

We conclude that *we have strong evidence that there is a birthday effect in this ice hockey team. Something beyond ordinary chance seems to be going on. The data are barely compatible with* H_0.

Why should there be a birthday effect?

These data are just one example of a much wider - and astonishingly strong - phenomenon. Professional sports players not just in ice hockey, but in soccer, baseball, and other sports have strong birthday clustering. Why?

It's because these sports select talented players for age-class star teams at young ages, about 10 years old. In ice hockey, the cut-off date for age-class teams is January 1st. A 10-year-old born in December is competing against players who are nearly a year older, born in January, February, and March. The age difference makes a big difference in terms of size, speed, and physical coordination. Most of the 'talented' players at this age are simply older and bigger. But there then follow years in which they get the best coaching and the most practice. By the time they reach 17, these players really are the best.

2.5 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.

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

and we wish to estimate the value of p.

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

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!!

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.*

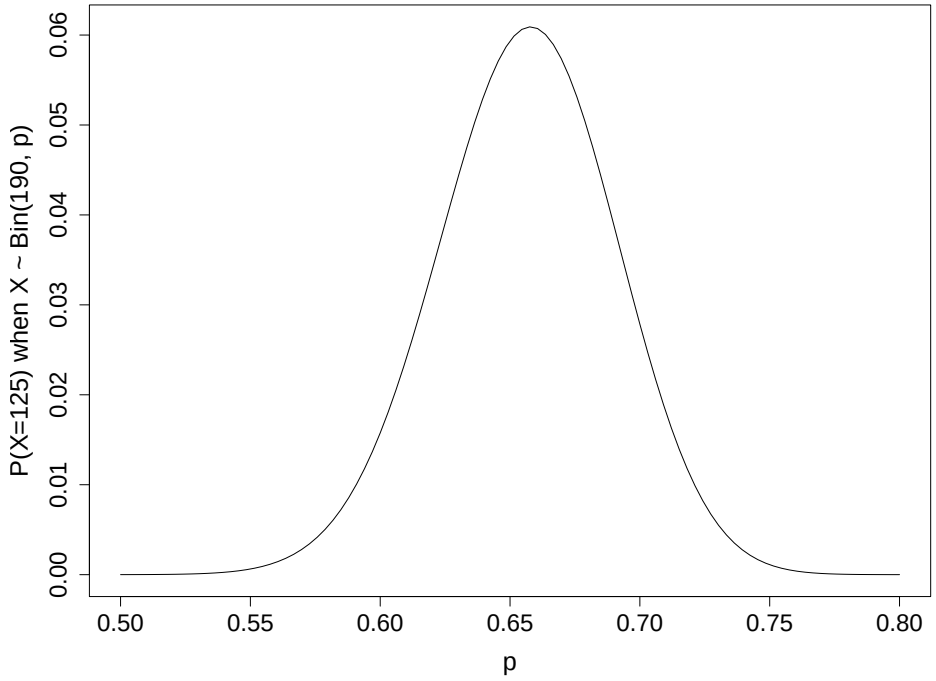
Can we do any better? What happens if we increase p a little more, say to $p = 0.7$?

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.



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:

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

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

We write:

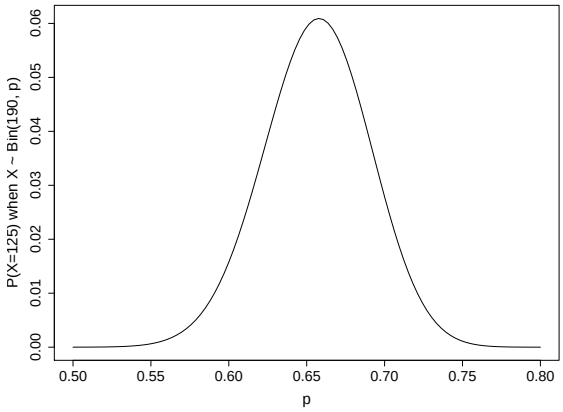
$$\begin{aligned} L(p\,;x) &= \mathbb{P}(X = x) \text{ when } X \sim \text{Binomial}(190, p), \\ &= \binom{190}{x} p^{x} (1-p)^{190-x}\,. \end{aligned}$$

Difference between the likelihood function and the probability function

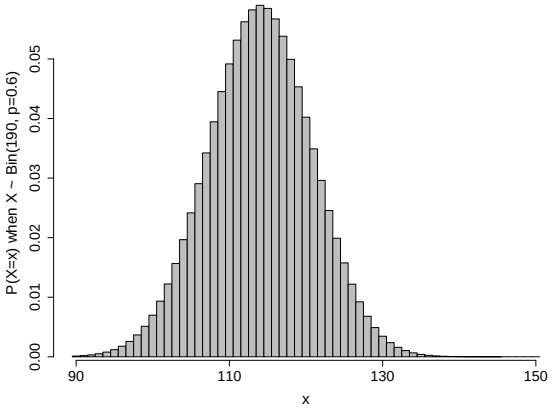
The likelihood function is *a probability of x, but it is a FUNCTION of p.*

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

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

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*:*

$$\begin{aligned}
\frac{dL}{dp} &= \binom{190}{125} \times \Big\{ 125 \times p^{124} \times (1-p)^{65} + p^{125} \times 65 \times (1-p)^{64} \times (-1) \Big\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \textit{(Product Rule)} \\
&= \binom{190}{125} \times p^{124} \times (1-p)^{64} \Big\{ 125(1-p) - 65p \Big\} \\
&= \binom{190}{125} p^{124} (1-p)^{64} \Big\{ 125 - 190p \Big\} \,.
\end{aligned}$$

The maximizing value of p occurs when

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

This gives:

$$\begin{aligned}
\frac{dL}{dp} = \binom{190}{125} p^{124} (1-p)^{64} \Big\{ 125 - 190p \Big\} &= 0 \\
\Rightarrow \quad \Big\{ 125 - 190p \Big\} &= 0 \\
\Rightarrow \quad p &= \frac{125}{190} = 0.658 \,.
\end{aligned}$$

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:

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

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.

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

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

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:

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

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

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

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

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

   (Exercise)

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

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

This is the *maximum likelihood estimate of 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.*

We expect that p probably IS different from 0.5, just by a little. The maximum likelihood estimate gives us the *'best'* estimate of p.

Note: We have only considered the class of problems for which $X \sim$ Binomial(n, p) and n is KNOWN. If n is not known, we have a harder problem: we have two parameters, and one of them (n) should only take discrete values $1, 2, 3, \ldots$.
We will not consider problems of this type in Stats 210.

2.6 Random numbers and histograms

We often wish to generate random numbers from a given distribution. Statistical packages like R have custom-made commands for doing this.

To generate (say) 100 random numbers from the Binomial$(n = 190, p = 0.6)$ distribution in R, we use:

```
rbinom(100, 190, 0.6)
```

or in long-hand,

```
rbinom(n=100, size=190, prob=0.6)
```

Caution: the R inputs `n` and `size` are the opposite to what you might expect: `n` gives the required sample size, and `size` gives the Binomial parameter n!

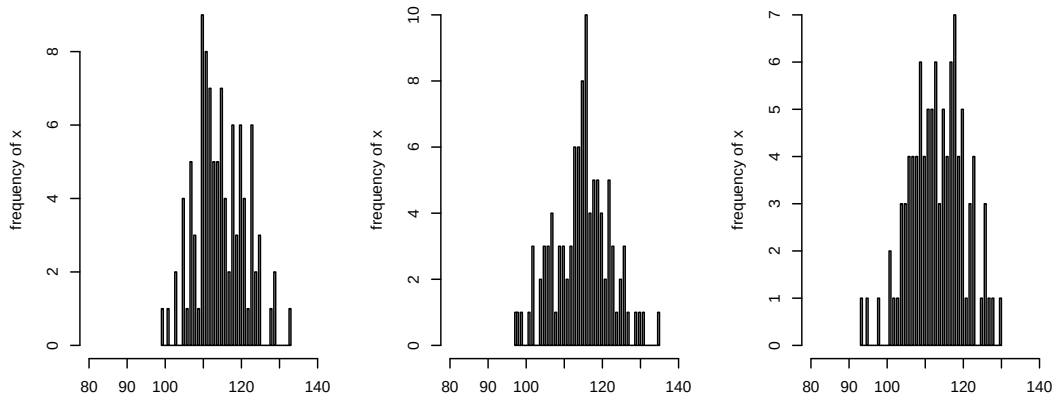
Histograms

The usual graph used to visualise a set of random numbers is the *histogram.*

The height of each bar of the histogram shows how many of the random numbers fall into the interval represented by the bar.

For example, if each histogram bar covers an interval of length 5, and if 24 of the random numbers fall between 105 and 110, then the height of the histogram bar for the interval (105, 110) would be *24.*

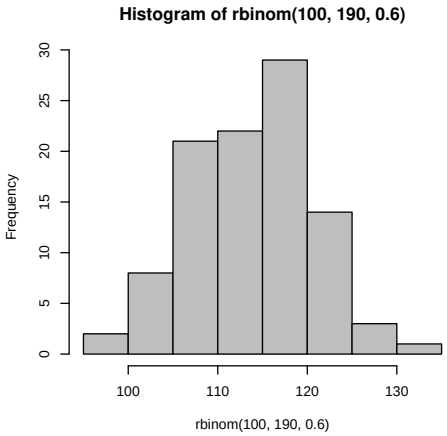
Here are histograms from applying the command `rbinom(100, 190, 0.6)` three different times.



Each graph shows *100 random numbers from the Binomial*$(n = 190, p = 0.6)$ *distribution.*

Note: The histograms above have been specially adjusted so that each histogram bar covers an interval of just one integer. For example, the height of the bar plotted at $x = 109$ shows *how many of the 100 random numbers are equal to 109.*

Usually, histogram bars would cover a larger interval, and the histogram would be smoother. For example, on the right is a histogram using the default settings in *R*, obtained from the command `hist(rbinom(100, 190, 0.6))`.



Each histogram bar covers an interval of *5 integers.*

In all the histograms above, the sum of the heights of all the bars is 100, because there are 100 observations.

Histograms as the sample size increases

Histograms are useful because *they show the approximate shape of the underlying probability function.*

They are also useful for exploring the effect of increasing sample size.

All the histograms below have bars covering an interval of *1 integer.*
They show how the histogram becomes smoother and less erratic as sample size increases.

Eventually, with a large enough sample size, *the histogram starts to look identical to the probability function.*

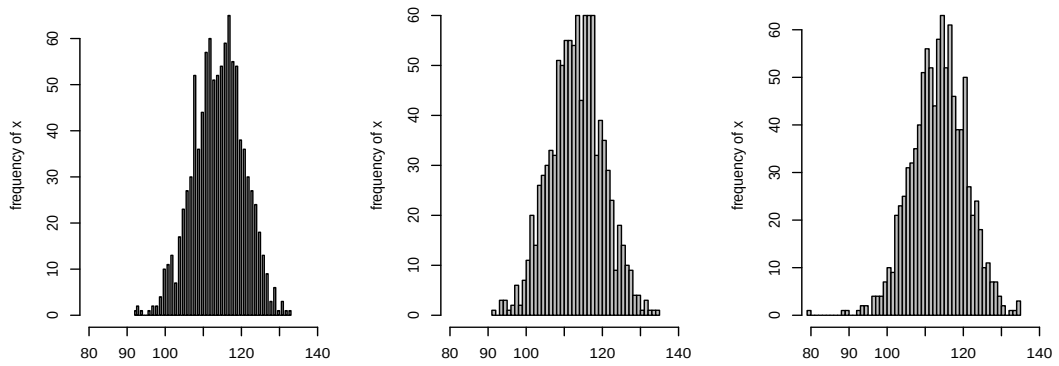
Note: difference between a histogram and the probability function

The histogram plots OBSERVED FREQUENCIES of a set of random numbers.

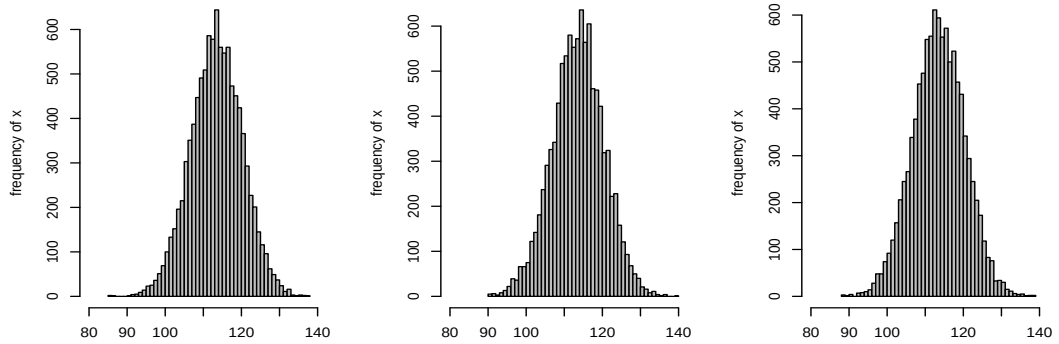
The probability function plots EXACT PROBABILITIES for the distribution.

The histogram *should have the same shape as the probability function, especially as the sample size gets large.*

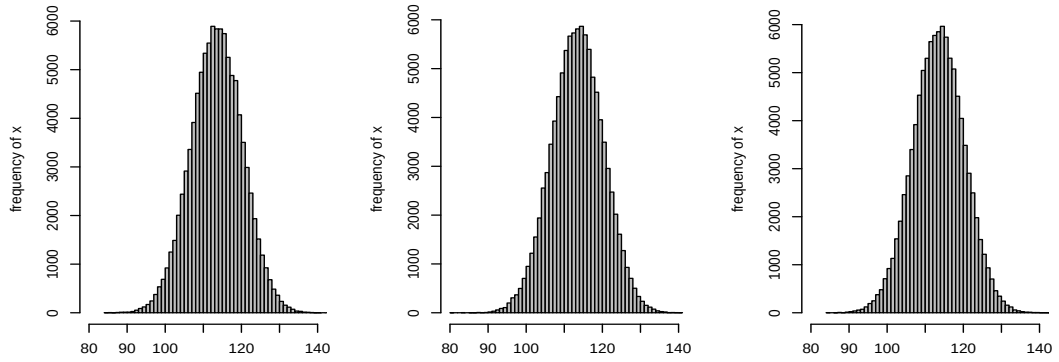
Sample size 1000: `rbinom(1000, 190, 0.6)`



Sample size 10,000: `rbinom(10000, 190, 0.6)`



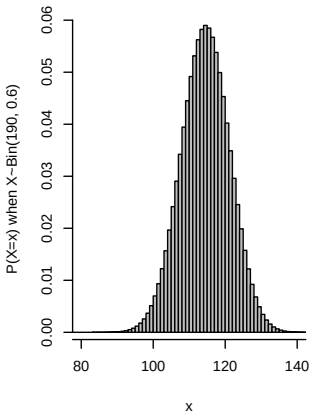
Sample size 100,000: `rbinom(100000, 190, 0.6)`



Probability function for Binomial(190, 0.6):

The probability function is
fixed and exact.

The histograms become stable in shape and approach the shape of the probability function as sample size gets large.



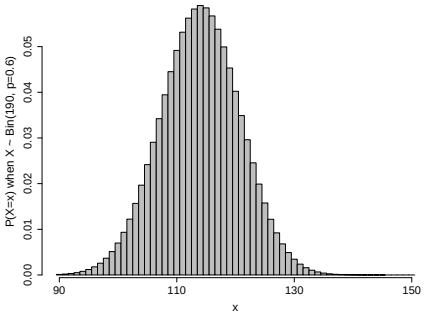
2.7 Expectation

Given a random variable X that measures something, we often want to know *what is the* **average** *value of* X*?*

For example, here are 30 random observations taken from the distribution $X \sim \text{Binomial}(n = 190, p = 0.6)$:

***R* command:** `rbinom(30, 190, 0.6)`

```
116 116 117 122 111 112 114 120 112 102
125 116  97 105 108 117 118 111 116 121
107 113 120 114 114 124 116 118 119 120
```



The average, or *mean*, of the ***first ten*** values is:

$$\frac{116 + 116 + \ldots + 112 + 102}{10} = 114.2.$$

The mean of the ***first twenty*** values is:

$$\frac{116 + 116 + \ldots + 116 + 121}{20} = 113.8.$$

The mean of the ***first thirty*** values is:

$$\frac{116 + 116 + \ldots + 119 + 120}{30} = 114.7.$$

The answers all seem to be close to *114*. What would happen if we took the average of hundreds of values?

100 values from Binomial(190, 0.6):

R command: `mean(rbinom(100, 190, 0.6))`
Result: `114.86`

Note: You will get a different result every time you run this command.

1000 values from Binomial(190, 0.6):

R command: `mean(rbinom(1000, 190, 0.6))`
Result: `114.02`

1 million values from Binomial(190, 0.6):

R command: `mean(rbinom(1000000, 190, 0.6))`
Result: `114.0001`

The average seems to be *converging to the value 114.*

The larger the sample size, *the closer the average seems to get to 114.*

If we kept going for larger and larger sample sizes, we would keep getting answers closer and closer to 114. This is because *114 is the DISTRIBUTION MEAN: the mean value that we would get if we were able to draw an infinite sample from the Binomial(190, 0.6) distribution.*

This distribution mean is called the *expectation, or expected value, of the Binomial(190, 0.6) distribution.*

It is a *FIXED property of the Binomial(190, 0.6) distribution.* This means it is a *fixed constant: there is nothing random about it.*

Definition: The **expected value**, also called the **expectation** or **mean**, of a discrete random variable X, *can be written as either* $\mathbb{E}(X)$, *or E(X), or* μ_X, *and is given by*

$$\mu_X = \mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)\,.$$

The expected value is a measure of the centre, or average, of the set of values that X can take, weighted according to the probability of each value.

If we took a very large sample of random numbers from the distribution of X, *their average would be approximately equal to* μ_X.

Example: Let $X \sim \text{Binomial}(n = 190, p = 0.6)$. What is $\mathbb{E}(X)$?

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{190} x\binom{190}{x}(0.6)^x(0.4)^{190-x}.\end{aligned}$$

Although it is not obvious, the answer to this sum is $n \times p = 190 \times 0.6 = 114$. We will see why in Section 2.10.

Explanation of the formula for expectation

We will move away from the Binomial distribution for a moment, and use a simpler example.

Let the random variable X be defined as $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

X takes only the values 1 and -1. What is the 'average' value of X?

Using $\frac{1+(-1)}{2} = 0$ *would not be useful, because it ignores the fact that usually* $X = 1$*, and only occasionally is* $X = -1$*.*

Instead, think of observing X many times, say 100 times.

Roughly *90* of these 100 times will have $X = 1$.
Roughly *10* of these 100 times will have $X = -1$

The average of the 100 values will be roughly

$$\begin{aligned}&\frac{90 \times 1 + 10 \times (-1)}{100}, \\ &= 0.9 \times 1 + 0.1 \times (-1) \\ (\ &= 0.8. \ )\end{aligned}$$

We could repeat this for any sample size.

As the sample gets large, the average of the sample will get ever closer to

$$0.9 \times 1 + 0.1 \times (-1).$$

This is why the distribution mean is given by

$$\mathbb{E}(X) = \mathbb{P}(X = 1) \times 1 + \mathbb{P}(X = -1) \times (-1),$$

or in general,

$$\mathbb{E}(X) = \sum_x \mathbb{P}(X = x) \times x.$$

> $\mathbb{E}(X)$ *is a fixed constant giving the average value we would get from a large sample of* X*.*

Linear property of expectation

Expectation is a ***linear*** operator:

Theorem 2.7: *Let* a *and* b *be constants. Then*

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Proof:

Immediate from the definition of expectation.

$$\begin{aligned} \mathbb{E}(aX + b) &= \sum_x (ax + b) f_X(x) \\ &= a \sum_x x f_X(x) + b \sum_x f_X(x) \\ &= a\,\mathbb{E}(X) + b \times 1. \end{aligned}$$

□

Example: finding expectation from the probability function

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$. Write down the probability function of X and find $\mathbb{E}(X)$.

We have:

$$\mathbb{P}(X = x) = \binom{3}{x}(0.2)^x(0.8)^{3-x} \text{ for } x = 0, 1, 2, 3.$$

x	*0*	*1*	*2*	*3*
$f_X(x) = \mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Then

$$\begin{aligned}\mathbb{E}(X) = \sum_{x=0}^{3} x f_X(x) &= 0 \times 0.512 + 1 \times 0.384 + 2 \times 0.096 + 3 \times 0.008 \\ &= 0.6.\end{aligned}$$

Note: We have: $\mathbb{E}(X) = 0.6 = 3 \times 0.2$ *for* $X \sim$ *Binomial*$(3, 0.2)$.
We will prove in Section 2.10 that whenever $X \sim \text{Binomial}(n, p)$, then $\mathbb{E}(X) = np$.

Example 2: Let Y be Bernoulli(p) (Section 1.2). That is,

$$Y = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability } 1 - p. \end{cases}$$

Find $\mathbb{E}(Y)$.

y	*0*	*1*
$\mathbb{P}(Y = y)$	$1 - p$	p

$$\mathbb{E}(Y) = 0 \times (1 - p) + 1 \times p = p.$$

Expectation of a sum of random variables: $\mathbb{E}(X+Y)$

For ANY random variables $X_1, X_2, \ldots, X_n$,

$$\mathbb{E}\left(X_1 + X_2 + \ldots + X_n\right) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n).$$

In particular, $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ *for ANY* X *and* Y.

This result holds for ***any*** random variables $X_1, \ldots, X_n$. *It does NOT require* $X_1, \ldots, X_n$ *to be independent.*

We can summarize this important result by saying:

The expectation of a sum
is the sum of the expectations – ALWAYS.

The proof requires multivariate methods, to be studied in later courses.

Note: We can combine the result above with the linear property of expectation. For any constants $a_1, \ldots, a_n$, we have:

$$\mathbb{E}\left(a_1X_1 + a_2X_2 + \ldots + a_nX_n\right) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n).$$

Expectation of a product of random variables: $\mathbb{E}(XY)$

There are two cases when finding the expectation of a product:

1. **General case:**

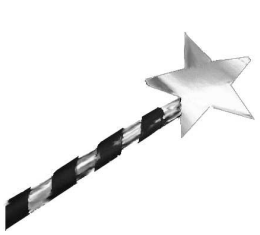
   For general X *and* Y, $\quad \mathbb{E}(XY)$ *is NOT equal to* $\mathbb{E}(X)\mathbb{E}(Y)$.

   We have to find $\mathbb{E}(XY)$ either using their joint probability function (see later), or using their covariance (see later).

2. **Special case:** when X and Y are *INDEPENDENT:*

   When X *and* Y *are INDEPENDENT,* $\quad \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

2.8 Variable transformations



We often wish to *transform* random variables through a function. For example, given the random variable X, possible transformations of X include:

$$X^2\,, \qquad \sqrt{X}\,, \qquad 4X^3\,, \qquad \ldots$$

We often summarize all possible variable transformations by referring to $Y = g(X)$ *for some function* g.

For discrete random variables, it is very easy to find the probability function for $Y = g(X)$, given that the probability function for X is known. Simply ***change all the values and keep the probabilities the same.***

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$, and let $Y = X^2$. Find the probability function of Y.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Thus *the probability function for* $Y = X^2$ *is:*

y	0^2	1^2	2^2	3^2
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

This is because Y *takes the value* 0^2 *whenever* X *takes the value* 0*, and so on.*

Thus the probability that $Y = 0^2$ is *the same as the probability that* $X = 0$.

Overall, we would write the probability function of $Y = X^2$ as:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

To transform a discrete random variable, ***transform the values and leave the probabilities alone.***

Example 2: Mr Chance hires out giant helium balloons for advertising. His balloons come in three sizes: heights 2m, 3m, and 4m. 50% of Mr Chance's customers choose to hire the cheapest 2m balloon, while 30% hire the 3m balloon and 20% hire the 4m balloon.

The amount of helium gas in cubic metres required to fill the balloons is $h^3/2$, where h is the height of the balloon. Find the probability function of Y, the amount of helium gas required for a randomly chosen customer.

Let X be the height of balloon ordered by a random customer. The probability function of X is:

height, x (m)	*2*	*3*	*4*
$\mathbb{P}(X = x)$	*0.5*	*0.3*	*0.2*

Let Y be the amount of gas required: $Y = X^3/2$.
The probability function of Y is:

gas, y (m^3)	*4*	*13.5*	*32*
$\mathbb{P}(Y = y)$	*0.5*	*0.3*	*0.2*

Transform the values, and leave the probabilities alone.

Expected value of a transformed random variable

We can find the expectation of a transformed random variable just like any other random variable. For example, in Example 1 we had $X \sim \text{Binomial}(3, 0.2)$, and $Y = X^2$.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

and for $Y = X^2$:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

Thus the expectation of $Y = X^2$ is:

$$\begin{aligned}\mathbb{E}(Y) = \mathbb{E}(X^2) &= 0 \times 0.512 + 1 \times 0.384 + 4 \times 0.096 + 9 \times 0.008 \\ &= 0.84.\end{aligned}$$

Note: $\mathbb{E}(X^2)$ *is NOT the same as* $\{\mathbb{E}(X)\}^2$. *Check that* $\{\mathbb{E}(X)\}^2 = 0.36$.

To make the calculation quicker, we could cut out the middle step of writing down the probability function of Y. Because we transform the values and keep the probabilities the same, we have:

$$\mathbb{E}(X^2) = 0^2 \times 0.512 + 1^2 \times 0.384 + 2^2 \times 0.096 + 3^2 \times 0.008.$$

If we write $g(X) = X^2$, this becomes:

$$\mathbb{E}\{g(X)\} = \mathbb{E}(X^2) = g(0) \times 0.512 + g(1) \times 0.384 + g(2) \times 0.096 + g(3) \times 0.008.$$

Clearly the same arguments can be extended to any function $g(X)$ and any discrete random variable X:

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x).}$$

Transform the values, and leave the probabilities alone.

Definition: For any function g and discrete random variable X, the expected value of $g(X)$ is given by

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x) = \sum_x g(x)f_X(x).}$$

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Gas required was $X^3/2$ from page 74.
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\begin{aligned}\mathbb{E}\left(\frac{X^3}{2}\right) &= \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x) \\ &= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2 \\ &= 12.45 \ \textit{m}^3 \textit{ gas.}\end{aligned}$$

(b) Mr Chance charges \$400$\times h$ to hire a balloon of height h. What is his expected earning per customer?

Expected earning is $\mathbb{E}(400X)$.

$$\begin{aligned}\mathbb{E}(400X) &= 400 \times \mathbb{E}(X) \qquad \textit{(expectation is linear)} \\ &= 400 \times (2 \times 0.5 + 3 \times 0.3 + 4 \times 0.2) \\ &= 400 \times 2.7 \\ &= \$1080 \textit{ per customer.}\end{aligned}$$

(c) How much does Mr Chance expect to earn in total from his next 5 customers?

Let $Z_1, \ldots, Z_5$ be the earnings from the next 5 customers. Each Z_i has $\mathbb{E}(Z_i) = 1080$ by part (b). The total expected earning is

$$\begin{aligned}\mathbb{E}(Z_1 + Z_2 + \ldots + Z_5) &= \mathbb{E}(Z_1) + \mathbb{E}(Z_2) + \ldots + \mathbb{E}(Z_5) \\ &= 5 \times 1080 \\ &= \$5400.\end{aligned}$$

Getting the expectation. . .

Suppose $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then $3/4$ of the time, X takes value 3, and $1/4$ of the time, X takes value 8.

So $\mathbb{E}(X) = \frac{3}{4} \times 3 + \frac{1}{4} \times 8.$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

What about $\mathbb{E}(\sqrt{X})$?

$$\sqrt{X} = \begin{cases} \sqrt{3} & \text{with probability } 3/4, \\ \sqrt{8} & \text{with probability } 1/4. \end{cases}$$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

$$\mathbb{E}(\sqrt{X}) = \frac{3}{4} \times \sqrt{3} + \frac{1}{4} \times \sqrt{8}.$$

Common mistakes

i) $\mathbb{E}(\sqrt{X}) = \sqrt{\mathbb{E}X} = \sqrt{\frac{3}{4} \times 3 + \frac{1}{4} \times 8}$

Wrong!

ii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4} \times 3} + \sqrt{\frac{1}{4} \times 8}$

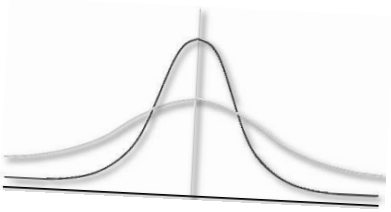
Wrong!

iii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4} \times 3} + \sqrt{\frac{1}{4} \times 8}$

Wrong!

$= \sqrt{\frac{3}{4} \times \sqrt{3}} + \sqrt{\frac{1}{4} \times \sqrt{8}}$

2.9 Variance



Example: Mrs Tractor runs the Rational Bank of Remuera. Every day she hopes to fill her cash machine with enough cash to see the well-heeled citizens of Remuera through the day. She knows that the expected amount of money withdrawn each day is \$50,000. How much money should she load in the machine? \$50,000?

No: \$50,000 is the average, near the centre of the distribution. About half the time, the money required will be GREATER than the average.

How much money should Mrs Tractor put in the machine if she wants to be 99% certain that there will be enough for the day's transactions?

Answer: it depends how much the amount withdrawn *varies above and below its mean.*

For questions like this, we need the study of *variance.*

Variance is the *average squared distance of a random variable from its own mean.*

Definition: The **variance** of a random variable X *is written as either* $\mathrm{Var}(X)$ *or* σ_X^2, *and is given by*

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}\left[(X-\mathbb{E}X)^2\right].$$

Similarly, the variance of a function of X is

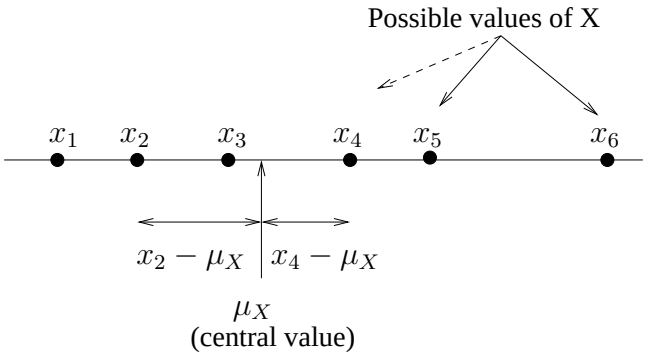
$$\mathrm{Var}(g(X)) = \mathbb{E}\left[\Big(g(X)-\mathbb{E}(g(X))\Big)^2\right].$$

Note: The variance *is the square of the standard deviation of X, so*

$$sd(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

Variance as the average squared distance from the mean

The variance is a measure of how *spread out* are the values that X can take. It is the *average squared distance between a value of X and the central (mean) value, μ_X.*



$$\textit{Var}(X) = \underbrace{\mathbb{E}}_{(2)}\ \underbrace{[(X-\mu_X)^2]}_{(1)}$$

(1) Take distance from observed values of X to the central point, μ_X. Square it to balance positive and negative distances.

(2) Then take the average over all values X can take: ie. if we observed X many times, find what would be the average squared distance between X and μ_X.

Note: The mean, μ_X, and the variance, σ_X^2, of X are just *numbers: there is nothing random or variable about them.*

Example: Let $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then

$$\begin{aligned} \mathbb{E}(X) = \mu_X &= 3 \times \frac{3}{4} + 8 \times \frac{1}{4} = 4.25 \\ \textit{Var}(X) = \sigma_X^2 &= \frac{3}{4} \times (3 - 4.25)^2 + \frac{1}{4} \times (8 - 4.25)^2 \\ &= 4.6875. \end{aligned}$$

When we observe X, we get either *3 or 8: this is random.*
But μ_X *is fixed at 4.25, and* σ_X^2 *is fixed at 4.6875, regardless of the outcome of X.*

For a discrete random variable,

$$\textit{Var}(X) = \mathbb{E}\left[(X-\mu_X)^2\right] = \sum_x (x-\mu_X)^2 f_X(x) = \sum_x (x-\mu_X)^2 \mathbb{P}(X=x).$$

This uses the definition of the expected value of a function of X:

$$\textit{Var}(X) = \mathbb{E}(g(X)) \quad \textit{where} \quad g(X) = (X-\mu_X)^2.$$

Theorem 2.9A: (important)

$$\textit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu_X^2$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu_X)^2\right] \quad \text{by definition} \\
&= \mathbb{E}[\underbrace{X^2}_{\text{r.v.}} - 2\underbrace{X}_{\text{r.v.}}\underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}}] \\
&= \mathbb{E}(X^2) - 2\mu_X\mathbb{E}(X) + \mu_X^2 \quad \text{by Thm 2.7} \\
&= \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= \mathbb{E}(X^2) - \mu_X^2 \,. \qquad \square
\end{aligned}$$

Note: $\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = \sum_x x^2\, \mathbb{P}(X=x)$. This is <u>not</u> the same as $(\mathbb{E}X)^2$:

e.g. $$X = \begin{cases} 3 & \text{with probability } 0.75, \\ 8 & \text{with probability } 0.25. \end{cases}$$

Then $\mu_X = \mathbb{E}X = 4.25$, so $\mu_X^2 = (\mathbb{E}X)^2 = (4.25)^2 = 18.0625$.
But

$$\mathbb{E}(X^2) = \left(3^2 \times \frac{3}{4} + 8^2 \times \frac{1}{4}\right) = 22.75.$$

Thus $\mathbb{E}(X^2) \neq (\mathbb{E}X)^2$ *in general.*

Theorem 2.9B: If a and b are constants and $g(x)$ is a function, then

i) $\mathit{Var}(aX + b) = a^2\,\mathit{Var}(X)$.

ii) $\mathit{Var}(a\,g(X) + b) = a^2\,\mathit{Var}\{g(X)\}$.

Proof:

(part (i))

$$\begin{aligned}
\mathit{Var}(aX+b) &= \mathbb{E}\Big[\{(aX+b) - \mathbb{E}(aX+b)\}^2\Big] \\
&= \mathbb{E}\Big[\{aX + b - a\mathbb{E}(X) - b\}^2\Big] \quad \textit{by Thm 2.7} \\
&= \mathbb{E}\Big[\{aX - a\mathbb{E}(X)\}^2\Big] \\
&= \mathbb{E}\Big[a^2\{X - \mathbb{E}(X)\}^2\Big] \\
&= a^2\mathbb{E}\Big[\{X - \mathbb{E}(X)\}^2\Big] \quad \textit{by Thm 2.7} \\
&= a^2\,\mathit{Var}(X)\,.
\end{aligned}$$

Part (ii) follows similarly.

Note: These are very different from the corresponding expressions for expectations (Theorem 2.7). Variances are more difficult to manipulate than expectations.

Example: finding expectation and variance from the probability function

Recall Mr Chance's balloons from page 74. The random variable Y is the amount of gas required by a randomly chosen customer. The probability function of Y is:

gas, y (m^3)	4	13.5	32
$\mathbb{P}(Y = y)$	0.5	0.3	0.2

Find $\mathrm{Var}(Y)$.

We know that $\mathbb{E}(Y) = \mu_Y = 12.45$ *from page 76.*

First method: *use* $Var(Y) = \mathbb{E}[(Y - \mu_Y)^2]$*:*

$$\begin{aligned} Var(Y) &= (4 - 12.45)^2 \times 0.5 + (13.5 - 12.45)^2 \times 0.3 + (32 - 12.45)^2 \times 0.2 \\ &= 112.47. \end{aligned}$$

Second method: *use* $\mathbb{E}(Y^2) - \mu_Y^2$*: (usually easier)*

$$\begin{aligned} \mathbb{E}(Y^2) &= 4^2 \times 0.5 + 13.5^2 \times 0.3 + 32^2 \times 0.2 \\ &= 267.475. \end{aligned}$$

So $Var(Y) = 267.475 - (12.45)^2 = 112.47$ *as before.*

Variance of a sum of random variables: $\mathbf{Var}(\boldsymbol{X + Y})$

There are two cases when finding the variance of a sum:

1. **General case:**

> *For general* X *and* Y*,*
> $Var(X + Y)$ *is NOT equal to* $Var(X) + Var(Y)$.

We have to find $\mathrm{Var}(X + Y)$ using their covariance (see later courses).

2. **Special case:** when X and Y are *INDEPENDENT:*

> *When* X *and* Y *are INDEPENDENT,*
> $Var(X + Y) = Var(X) + Var(Y)$.

2.10 Mean and Variance of the Binomial(n, p) distribution

Let $X \sim \text{Binomial}(n, p)$. We have mentioned several times that $\mathbb{E}(X) = np$. We now prove this and the additional result for $\text{Var}(X)$.

If $X \sim \text{Binomial}(n, p)$, then:

$$\boxed{\begin{array}{c}\mathbb{E}(X) = \mu_X = np \\ \textit{Var}(X) = \sigma_X^2 = np(1-p).\end{array}}$$

We often write $q = 1 - p$, so Var$(X) = npq$.

Easy proof: X as a sum of Bernoulli random variables

If $X \sim \text{Binomial}(n, p)$, then X is the *number of successes out of n independent trials, each with $\mathbb{P}(\text{success}) = p$.*

This means that we can write:

$$X = Y_1 + Y_2 + \ldots + Y_n,$$

where each

$$Y_i = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability } 1-p. \end{cases}$$

That is, Y_i *counts as a 1 if trial i is a success, and as a 0 if trial i is a failure.*

Overall, $Y_1 + \ldots + Y_n$ *is the total number of successes out of n independent trials, which is the same as X.*

Note: Each Y_i is a Bernoulli(p) random variable (Section 1.2).

Now if $X = Y_1 + Y_2 + \ldots + Y_n$, and $Y_1, \ldots, Y_n$ are independent, then:

$$\mathbb{E}(X) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \ldots + \mathbb{E}(Y_n) \quad \textit{(does NOT require independence),}$$

$$\textit{Var}(X) = \textit{Var}(Y_1) + \textit{Var}(Y_2) + \ldots + \textit{Var}(Y_n) \quad \textit{(DOES require independence).}$$

The probability function of each Y_i is:

y	0	1
$\mathbb{P}(Y_i = y)$	$1-p$	p

Thus,

$$\mathbb{E}(Y_i) = 0 \times (1-p) + 1 \times p = p.$$

Also,

$$\mathbb{E}(Y_i^2) = 0^2 \times (1-p) + 1^2 \times p = p.$$

So

$$\begin{aligned} \textit{Var}(Y_i) &= \mathbb{E}(Y_i^2) - (\mathbb{E}Y_i)^2 \\ &= p - p^2 \\ &= p(1-p). \end{aligned}$$

Therefore:

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \ldots + \mathbb{E}(Y_n) \\ &= p + p + \ldots + p \\ &= n \times p. \end{aligned}$$

And:

$$\begin{aligned} \textit{Var}(X) &= \textit{Var}(Y_1) + \textit{Var}(Y_2) + \ldots + \textit{Var}(Y_n) \\ &= n \times p(1-p). \end{aligned}$$

Thus we have proved that $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$. $\square$

Hard proof: for mathematicians (non-examinable)

We show below how the Binomial mean and variance formulae can be derived directly from the probability function.

$$\mathbb{E}(X) = \sum_{x=0}^{n} x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} x \Big(\frac{n!}{(n-x)!x!} \Big) p^x (1-p)^{n-x}$$

But $\dfrac{x}{x!} = \dfrac{1}{(x-1)!}$ and also the first term $x f_X(x)$ is 0 when $x = 0$.

So, continuing,

$$\mathbb{E}(X) = \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!} p^x (1-p)^{n-x}$$

Next: make n's into $(n-1)$'s, x's into $(x-1)$'s, wherever possible: e.g.

$$\begin{aligned} n - x &= (n-1) - (x-1), \quad p^x = p \cdot p^{x-1} \\ n! &= n(n-1)! \ \textit{etc.} \end{aligned}$$

This gives,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=1}^{n} \frac{n(n-1)!}{[(n-1)-(x-1)]!(x-1)!} p \cdot p^{(x-1)} (1-p)^{(n-1)-(x-1)} \\ &= \underbrace{np}_{\text{what we want}} \quad \underbrace{\sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)}}_{\text{need to show this sum = 1}} \end{aligned}$$

Finally we let $y = x - 1$ and let $m = n - 1$.
When $x = 1, y = 0$; and when $x = n,\ y = n - 1 = m$.

So

$$\begin{aligned} \mathbb{E}(X) &= np \sum_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y} \\ &= np(p + (1-p))^m \quad \text{(Binomial Theorem)} \end{aligned}$$

$$\mathbb{E}(X) = np, \quad \text{as required.}$$

For $\mathrm{Var}(X)$, use the same ideas again.

For $\mathbb{E}(X)$, we used $\frac{x}{x!} = \frac{1}{(x-1)!}$; so instead of finding $\mathbb{E}(X^2)$, it will be easier to find $\mathbb{E}[X(X-1)] = \mathbb{E}(X^2) - \mathbb{E}(X)$ because then we will be able to cancel $\frac{x(x-1)}{x!} = \frac{1}{(x-2)!}$.

Here goes:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n} \frac{x(x-1)n(n-1)(n-2)!}{[(n-2)-(x-2)]!(x-2)!x(x-1)}p^2p^{(x-2)}(1-p)^{(n-2)-(x-2)}\end{aligned}$$

First two terms ($x=0$ and $x=1$) are 0 due to the $x(x-1)$ in the numerator. Thus

$$\begin{aligned}\mathbb{E}[X(X-1)] &= p^2n(n-1)\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}(1-p)^{(n-2)-(x-2)} \\ &= n(n-1)p^2\underbrace{\sum_{y=0}^{m}\binom{m}{y}p^y(1-p)^{m-y}}_{\text{sum=1 by Binomial Theorem}} \quad \text{if } \begin{cases} m = n-2, \\ y = x-2. \end{cases}\end{aligned}$$

$$\text{So} \quad \mathbb{E}[X(X-1)] = n(n-1)p^2\,.$$

$$\begin{aligned}\text{Thus} \quad \mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X) + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p). \qquad \square\end{aligned}$$

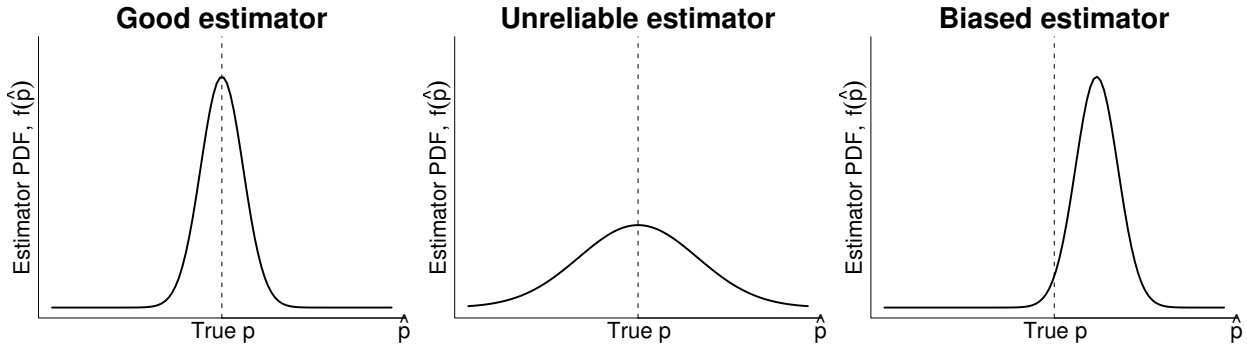
Note the steps: take out $x(x-1)$ and replace n by $(n-2)$, x by $(x-2)$ wherever possible.

2.11 Mean and Variance of Estimators

Perhaps the most important application of mean and variance is in the context of ***estimators:***

- An estimator is a *random variable.*
- It has a ***mean*** and a ***variance.***
- The ***mean*** tells us how ***accurate*** the estimator is: in particular, does it get the right answer on average?
- The ***variance*** tells us how ***reliable*** the estimator is. If the variance is high, it has high spread and we can get estimates a long way from the true answer. Because we don't know what the right answer is, we don't know whether our particular estimate is a good one (close to the true answer) or a bad one (a long way from the true answer). So an estimator with high variance is ***unreliable:*** sometimes it gives bad answers, sometimes it gives good answers; and we don't know which we've got.
  An unreliable estimator is like a friend who often tells lies. Once you find them out, you can't believe ANYTHING they say!

Good estimators and bad estimators



Example: MLE of the Binomial p parameter

In Section 2.5 we derived the MLE for the Binomial parameter p.

Reminder: $X \sim \text{Binomial}(n, p)$, where n is KNOWN and p is to be estimated.

Make a single observation $X = x$.

The maximum likelihood estimator of p is $\quad \widehat{p} = \dfrac{X}{n}$.

Why do we convert the estimate x/n to the estimator X/n?

The estimator has a ***mean*** and a ***variance***, and this means *we can study its properties.* Is it accurate? Is it reliable?

Estimator Mean, $\mathbb{E}(\widehat{p})$

$$\begin{aligned}
\mathbb{E}(\widehat{p}) = \mathbb{E}\left(\frac{X}{n}\right) &= \frac{1}{n}\mathbb{E}(X) \\
&= \frac{1}{n} \times np \qquad \textit{because } \mathbb{E}(X) = np \textit{ when } X \sim \textit{Binomial}(n,p) \\
\therefore \quad \mathbb{E}(\widehat{p}) &= p.
\end{aligned}$$

So this estimator ***gets the right answer on average*** — it is ***unbiased.***

Definition: If $\widehat{p}$ is an estimator of the parameter p, then $\widehat{p}$ is ***unbiased*** if $\mathbb{E}(\widehat{p}) = p$.

That is, an unbiased estimator *gets the right answer on average.*

If $\mathbb{E}(\widehat{p}) \neq p$, then $\widehat{p}$ is said to be a *biased estimator.*

If an estimator has a large bias, we probably don't want to use it. However, even if the estimator is *unbiased*, we still need to look at its ***variance*** to decide how *reliable* it is.

Estimator Variance, $\mathrm{Var}(\widehat{p})$

We have:

$$\begin{aligned}
\textit{Var}(\widehat{p}) &= \textit{Var}\left(\frac{X}{n}\right) \\
&= \frac{1}{n^2}\textit{Var}(X) \\
&= \frac{1}{n^2} \times np(1-p) \qquad \textit{because Var}(X) = np(1-p) \textit{ for } X \sim \textit{Bin}(n,p) \\
\therefore \quad \textit{Var}(\widehat{p}) &= \frac{p(1-p)}{n}. \qquad (\star)
\end{aligned}$$

To decide how reliable our estimator $\widehat{p}$ is, we would like to calculate the value of $\mathrm{Var}(\widehat{p})$. But $\mathrm{Var}(\widehat{p}) = p(1-p)/n$, and *we do not know the true value of p, so we cannot calculate the exact Var$(\widehat{p})$.*

Instead, *we have to ESTIMATE* $\mathrm{Var}(\widehat{p})$ *by replacing the unknown* p *in equation* $(\star)$ *by* $\widehat{p}$.

We call our ***estimated variance*** $\widehat{\mathrm{Var}}(\widehat{p})$*:* $\boxed{\widehat{\mathrm{Var}}(\widehat{p}) = \dfrac{\widehat{p}(1-\widehat{p})}{n}.}$

The ***standard error of*** $\widehat{\boldsymbol{p}}$ is defined as: $\mathrm{se}(\widehat{p}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{p})}$.

The ***margin of error*** associated with $\widehat{p}$ is defined as:

$$\boxed{\text{Margin of error} = 1.96 \times \mathrm{se}(\widehat{p}) = 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.}$$

The expression $\widehat{p} \pm 1.96 \times \mathrm{se}(\widehat{p})$ gives an approximate *95% confidence interval for* p *under the Normal approximation.*

This is because the Central Limit Theorem guarantees that $\widehat{p}$ will be approximately Normally distributed when n is large. We will study the Central Limit Theorem and this result in Chapter 5, Section 5.3.

Example: For the deep-sea diver example in Section 2.3, we had $X \sim \text{Binomial}(190, p)$ with observation $X = 125$ daughters out of 190 children.

So

$$\widehat{p} = \frac{X}{n} = \frac{125}{190} = 0.658, \qquad \Rightarrow \qquad \mathrm{se}(\widehat{p}) = \sqrt{\frac{0.658 \times (1-0.658)}{190}} = 0.034.$$

For our final answer, we should therefore quote:

$$\widehat{p} = 0.658 \pm 1.96 \times 0.034 = 0.658 \pm 0.067 \qquad \textit{or} \quad \widehat{p} = 0.658 \quad (0.591, 0.725).$$

Our estimate is fairly precise, although not extremely precise. We are pretty sure that the daughter probability is somewhere between 0.59 and 0.73.

Why do we use the MLE instead of some other estimator?

The MLE is a sensible estimator to use, but we could think of other sensible estimators too. The reason why the MLE is so highly preferred is because it has ***excellent general properties.*** Under mild conditions, and with a large enough sample size, any MLE will be (i) unbiased, (ii) Normally distributed, and (iii) have the minimal possible variance of all estimators. Wow!

Comments about p, $\widehat{p}$, $\mathbb{E}(\widehat{p})$, $\mathbf{Var}(\widehat{p})$, and $\widehat{\mathbf{Var}}(\widehat{p})$

It might seem difficult at first to get to grips with what these quantities are and what they represent. Here are some comments and notes.

- p is a parameter: an unknown but fixed ***number*** that we wish to estimate.
- $\widehat{p}$ is a ***random variable:*** for example, $\widehat{p} = \frac{X}{n}$. It is a particular type of random variable that generates estimates of p, so it is called an ***estimator***.
- $\mathbb{E}(\widehat{p})$ is a number that tells us whether or not our estimator is unbiased. We are mostly interested in $\mathbb{E}(\widehat{p})$ in an abstract sense: for example, if $\mathbb{E}(\widehat{p}) = p$, no matter what p is, then our estimator is unbiased and we are happy. If $\mathbb{E}(\widehat{p}) \neq p$, we want to know how badly wrong it is and whether we should devise a correction factor.

  For example, if we discovered that $\mathbb{E}(\widehat{p}) = \left(\frac{n}{n+1}\right) p$, then we could create a different estimator $\widehat{q}$ such that $\widehat{q} = \left(\frac{n+1}{n}\right) \widehat{p}$. Then we would have,

  $$\mathbb{E}(\widehat{q}) = \left(\tfrac{n+1}{n}\right) \mathbb{E}(\widehat{p}) = \left(\tfrac{n+1}{n}\right) \times \left(\tfrac{n}{n+1}\right) p = p.$$

  So our new estimator $\widehat{q}$ is unbiased for p, but on the downside it also has higher variance than $\widehat{p}$, because $\mathrm{Var}(\widehat{q}) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}(\widehat{p})$. So we might or might not prefer to use $\widehat{q}$ instead of $\widehat{p}$. As the sample size n grows large, we might prefer to accept a tiny bias with the lower variance and use $\widehat{p}$.
- $\mathrm{Var}(\widehat{p})$ is a number that tells us about the ***reliability*** of our estimator. Unlike $\mathbb{E}(\widehat{p})$ which we care about more as an abstract property, we would like to know the actual numeric value of $\mathrm{Var}(\widehat{p})$ so we can calculate confidence intervals. Confidence intervals quantify our estimator reliability and should be included with our final report.

  Unfortunately, we find that $\mathrm{Var}(\widehat{p})$ depends upon the unknown value p: for example, $\mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}$, so we can't calculate it because we don't know what p is. This is why we use $\widehat{\mathrm{Var}}(\widehat{p})$ described next.
- $\widehat{\mathrm{Var}}(\widehat{p})$ is our best attempt at getting a value for $\mathrm{Var}(\widehat{p})$. We just take the expression for $\mathrm{Var}(\widehat{p})$ and substitute $\widehat{p}$ for the unknown p everywhere. This means that $\widehat{\mathrm{Var}}(\widehat{p})$ is an ***estimator*** for $\mathrm{Var}(\widehat{p})$.

  For example, if $\mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}$, then $\widehat{\mathrm{Var}}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n}$.

  Because $\widehat{\mathrm{Var}}(\widehat{p})$ is a function of the random variable $\widehat{p}$, $\widehat{\mathrm{Var}}(\widehat{p})$ is also a ***random variable.*** Typically, we use it only for calculating its numerical value and transforming this into a standard error and a confidence interval as described on the previous page.