# Hypothesis testing and statistical modelling

Matthew Daws

UCLan

Jan 2020

# A biased coin?

*I suspect a coin I have is biased. I toss it 12 times, and get 9 heads and 3 tails.*

Before we begin, let's just think: do we think this data does support the idea that the coin is biased in favour of heads?

# A biased coin?

> *I suspect a coin I have is biased. I toss it 12 times, and get 9 heads and 3 tails.*

Before we begin, let's just think: do we think this data does support the idea that the coin is biased in favour of heads?

# Coin example: the probabilities

We will assume that each toss of the coin is identical and independent.

So once we know the real chance of getting a head, $p$, we know everything about the probabilities.

If we toss the coin $N$ times then how can we get $n$ heads (and so $N - n$ tails)? We need a bit of combinatorics: there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

ways to get $n$ heads is some order.

The probability of getting $n$ heads is $p^n$, and of getting $N - n$ tails is $(1-p)^{N-n}$. So in total the probability is

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

This is the "binomial distribution".

# Coin example: the probabilities

We will assume that each toss of the coin is identical and independent. So once we know the real chance of getting a head, $p$, we know everything about the probabilities.

If we toss the coin $N$ times then how can we get $n$ heads (and so $N - n$ tails)? We need a bit of combinatorics: there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

ways to get $n$ heads is some order.

The probability of getting $n$ heads is $p^n$, and of getting $N - n$ tails is $(1-p)^{N-n}$. So in total the probability is

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

This is the "binomial distribution".

# Coin example: the probabilities

We will assume that each toss of the coin is identical and independent. So once we know the real chance of getting a head, $p$, we know everything about the probabilities.

If we toss the coin $N$ times then how can we get $n$ heads (and so $N - n$ tails)? We need a bit of combinatorics: there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

ways to get $n$ heads is some order.

The probability of getting $n$ heads is $p^n$, and of getting $N - n$ tails is $(1-p)^{N-n}$. So in total the probability is

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

This is the “binomial distribution”.

# Coin example: the probabilities

We will assume that each toss of the coin is identical and independent. So once we know the real chance of getting a head, $p$, we know everything about the probabilities.

If we toss the coin $N$ times then how can we get $n$ heads (and so $N - n$ tails)? We need a bit of combinatorics: there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

ways to get $n$ heads is some order.

The probability of getting $n$ heads is $p^n$, and of getting $N - n$ tails is $(1-p)^{N-n}$. So in total the probability is

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

This is the “binomial distribution”.

# Coin example: the probabilities

We will assume that each toss of the coin is identical and independent. So once we know the real chance of getting a head, $p$, we know everything about the probabilities.

If we toss the coin $N$ times then how can we get $n$ heads (and so $N - n$ tails)? We need a bit of combinatorics: there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

ways to get $n$ heads is some order.

The probability of getting $n$ heads is $p^n$, and of getting $N - n$ tails is $(1-p)^{N-n}$. So in total the probability is

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

This is the "binomial distribution".

# Likelihoods

Notice that

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

depends on $p$. As we don't know the value of $p$, it is better to include it in the notation, and write $f(n|p)$.

But we have our data: we know the value of $n$. And we don't know the value of $p$.

So let's turn the notation around, and define the *likelihood* of $p$ as

$$\text{lik}(p) = f(n|p).$$

# Likelihoods

Notice that

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

depends on $p$. As we don't know the value of $p$, it is better to include it in the notation, and write $f(n|p)$.

But we have our data: we know the value of $n$. And we don't know the value of $p$.

So let's turn the notation around, and define the *likelihood* of $p$ as

$$\mathrm{lik}(p) = f(n|p).$$

# Likelihoods

Notice that

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

depends on $p$. As we don't know the value of $p$, it is better to include it in the notation, and write $f(n|p)$.

But we have our data: we know the value of $n$. And we don't know the value of $p$.

So let's turn the notation around, and define the *likelihood* of $p$ as

$$\text{lik}(p) = f(n|p).$$

# Estimation

What value of $p$ is "most likely", given our data?

A common way to answer this is via "maximum likelihood estimation".
We estimate $p$ as

$$\hat{p} = \text{argmax}_p \, \text{lik}(p).$$

That is, our estimate $\hat{p}$ is the value of $p$ which gives the greatest likelihood.

For practical calculations, looking at $\log \text{lik}(p)$ is usually easier.

# Estimation

What value of $p$ is "most likely", given our data?
A common way to answer this is via "maximum likelihood estimation".
We estimate $p$ as

$$\hat{p} = \operatorname{argmax}_p \operatorname{lik}(p).$$

That is, our estimate $\hat{p}$ is the value of $p$ which gives the greatest likelihood.
For practical calculations, looking at $\log \operatorname{lik}(p)$ is usually easier.

# Estimation

What value of $p$ is "most likely", given our data?
A common way to answer this is via "maximum likelihood estimation".
We estimate $p$ as

$$\hat{p} = \operatorname{argmax}_p \operatorname{lik}(p).$$

That is, our estimate $\hat{p}$ is the value of $p$ which gives the greatest likelihood.

For practical calculations, looking at $\log \operatorname{lik}(p)$ is usually easier.

# Estimation

What value of $p$ is "most likely", given our data?
A common way to answer this is via "maximum likelihood estimation".
We estimate $p$ as

$$\hat{p} = \text{argmax}_p \, \text{lik}(p).$$

That is, our estimate $\hat{p}$ is the value of $p$ which gives the greatest likelihood.
For practical calculations, looking at $\log \text{lik}(p)$ is usually easier.

# For the coin example

We have that

$$\log \text{lik}(p) = \log f(n|p) = \log \binom{N}{n} p^n (1-p)^{N-n}$$
$$= \log \binom{N}{n} + n\log(p) + (N-n)\log(1-p).$$

Maximising over $p$:

$$\frac{d}{dp}\log \text{lik}(p) = \frac{n}{p} - \frac{N-n}{1-p} = \frac{n(1-p)-p(N-n)}{p(1-p)} = \frac{n-pN}{p(1-p)}.$$

The turning point is at $n - pN = 0$ so $\hat{p} = n/N$
[Maths hat on: we should check that this really is the maximum. It is!]

# For the coin example

We have that

$$\log \text{lik}(p) = \log f(n|p) = \log \binom{N}{n} p^n (1-p)^{N-n}$$
$$= \log \binom{N}{n} + n\log(p) + (N-n)\log(1-p).$$

Maximising over $p$:

$$\frac{d}{dp}\log \text{lik}(p) = \frac{n}{p} - \frac{N-n}{1-p} = \frac{n(1-p)-p(N-n)}{p(1-p)} = \frac{n-pN}{p(1-p)}.$$

The turning point is at $n - pN = 0$ so $\hat{p} = n/N$
[Maths hat on: we should check that this really is the maximum. It is!]

# For the coin example

We have that

$$\log \text{lik}(p) = \log f(n|p) = \log \binom{N}{n} p^n (1-p)^{N-n}$$
$$= \log \binom{N}{n} + n\log(p) + (N-n)\log(1-p).$$

Maximising over $p$:

$$\frac{d}{dp}\log \text{lik}(p) = \frac{n}{p} - \frac{N-n}{1-p} = \frac{n(1-p)-p(N-n)}{p(1-p)} = \frac{n-pN}{p(1-p)}.$$

The turning point is at $n - pN = 0$ so $\hat{p} = n/N$
[Maths hat on: we should check that this really is the maximum. It is!]

# For the coin example

We have that

$$\log \text{lik}(p) = \log f(n|p) = \log \binom{N}{n} p^n (1-p)^{N-n}$$
$$= \log \binom{N}{n} + n\log(p) + (N-n)\log(1-p).$$

Maximising over $p$:

$$\frac{d}{dp} \log \text{lik}(p) = \frac{n}{p} - \frac{N-n}{1-p} = \frac{n(1-p) - p(N-n)}{p(1-p)} = \frac{n-pN}{p(1-p)}.$$

The turning point is at $n - pN = 0$ so $\hat{p} = n/N$
[Maths hat on: we should check that this really is the maximum. It is!]

# For the coin example

We have that

$$\log \text{lik}(p) = \log f(n|p) = \log \binom{N}{n} p^n (1-p)^{N-n}$$
$$= \log \binom{N}{n} + n\log(p) + (N-n)\log(1-p).$$

Maximising over $p$:

$$\frac{d}{dp}\log \text{lik}(p) = \frac{n}{p} - \frac{N-n}{1-p} = \frac{n(1-p)-p(N-n)}{p(1-p)} = \frac{n-pN}{p(1-p)}.$$

The turning point is at $n - pN = 0$ so $\hat{p} = n/N$
[Maths hat on: we should check that this really is the maximum. It is!]

# Mean, standard deviation etc.

All the standard formulaes we know for means, standard deviations etc. can be justified using maximum likelihood, or small extensions.

So this gives a simple idea which can unify a lot of elementary statistics.

# Hypothesis testing

The classical statistical approach is (Neyman–Pearson) hypothesis testing. We formulate two hypotheses, which are asymmetric:

**Definition**

$H_0$ is the "null hypothesis" which is the "status quo".
$H_1$ is the "alternative hypothesis" which is the sort of departure from $H_0$ which interests us.

- $H_0$ and $H_1$ are sometimes mutually exclusive, but need not cover all possible outcomes.
- But sometimes $H_1$ will simply be "anything is possible", against some more specific $H_0$.
- Our aim is to perform a "statistical test" with the aim of (maybe) "rejecting" $H_0$ in favour of $H_1$.

# Hypothesis testing

The classical statistical approach is (Neyman–Pearson) hypothesis testing. We formulate two hypotheses, which are asymmetric:

**Definition**

$H_0$ is the "null hypothesis" which is the "status quo".
$H_1$ is the "alternative hypothesis" which is the sort of departure from $H_0$ which interests us.

- $H_0$ and $H_1$ are sometimes mutually exclusive, but need not cover all possible outcomes.
- But sometimes $H_1$ will simply be "anything is possible", against some more specific $H_0$.
- Our aim is to perform a "statistical test" with the aim of (maybe) "rejecting" $H_0$ in favour of $H_1$.

# Hypothesis testing

The classical statistical approach is (Neyman–Pearson) hypothesis testing. We formulate two hypotheses, which are asymmetric:

**Definition**

$H_0$ is the "null hypothesis" which is the "status quo".
$H_1$ is the "alternative hypothesis" which is the sort of departure from $H_0$ which interests us.

- $H_0$ and $H_1$ are sometimes mutually exclusive, but need not cover all possible outcomes.
- But sometimes $H_1$ will simply be "anything is possible", against some more specific $H_0$.
- Our aim is to perform a "statistical test" with the aim of (maybe) "rejecting" $H_0$ in favour of $H_1$.

# Hypothesis testing

The classical statistical approach is (Neyman–Pearson) hypothesis testing. We formulate two hypotheses, which are asymmetric:

**Definition**

$H_0$ is the "null hypothesis" which is the "status quo".
$H_1$ is the "alternative hypothesis" which is the sort of departure from $H_0$ which interests us.

- $H_0$ and $H_1$ are sometimes mutually exclusive, but need not cover all possible outcomes.
- But sometimes $H_1$ will simply be "anything is possible", against some more specific $H_0$.
- Our aim is to perform a "statistical test" with the aim of (maybe) "rejecting" $H_0$ in favour of $H_1$.

# Coin example: the setup

We are interested in what is the probability of getting a head. Let's call this $p$. We shall then test

$$H_0 : p = 1/2 \quad \text{against} \quad H_1 : p \neq 1/2.$$

As we suspect that heads are more likely, we could instead test

$$H_0 : p = 1/2 \quad \text{against} \quad H_1 : p > 1/2.$$

# Coin example: the setup

We are interested in what is the probability of getting a head. Let's call this $p$. We shall then test

$$H_0 : p = 1/2 \quad \text{against} \quad H_1 : p \neq 1/2.$$

As we suspect that heads are more likely, we could instead test

$$H_0 : p = 1/2 \quad \text{against} \quad H_1 : p > 1/2.$$

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.

The probability of a type I error is also called the "size" or "significance level" of the test.

We typically construct tests by fixing a "size" we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.

The probability of a type I error is also called the “size” or “significance level” of the test.

We typically construct tests by fixing a “size” we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.

The probability of a type I error is also called the “size” or “significance level” of the test.

We typically construct tests by fixing a “size” we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.

The probability of a type I error is also called the "size" or "significance level" of the test.

We typically construct tests by fixing a "size" we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.
The probability of a type I error is also called the "size" or "significance level" of the test.
We typically construct tests by fixing a "size" we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# Type I and II errors

There are two possible errors we can make when performing a hypothesis test:

- $H_0$ could be rejected when it is true (a type I error);
- $H_0$ could be accepted when it is false (a type II error).

As we think of $H_0$ as the conservative / safe choice, we regard type I errors as more serious than type II errors.
The probability of a type I error is also called the "size" or "significance level" of the test.
We typically construct tests by fixing a "size" we are happy with (e.g. 5%) and then finding the test which the smallest type II error.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a “statistic” $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is “large”. Here “large” will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are “rejecting $H_0$” and not “accepting $H_1$”.

The probability which occurs in (2) is the “$p$-value”. It’s the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# $p$-values

In practise, almost all tests involve computing a "statistic" $Z$ (some value from the data) and rejecting $H_0$ if $Z$ is "large". Here "large" will depend on the size of the test.

The interpretation is the following:

1. We assume $H_0$ is true.
2. If $H_0$ is true, then $Z$ has a very small probability of being large.
3. If with our data $Z$ does turn out to be large, then we think: that was very unlikely if $H_0$ were true, so we have evidence to reject $H_0$.

Notice that $H_1$ did not appear. We use $H_1$ in the construction of the test, but it is worth remembering that ultimately we are "rejecting $H_0$" and not "accepting $H_1$".

The probability which occurs in (2) is the "$p$-value". It's the probability, assuming $H_0$ is true, of seeing data, as, or more, extreme, than the data we have.

# Coin example: $p$-values

It seems intuitively obvious (and can be Mathematically justified) that when testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ we should reject $H_0$ if the observed number of heads (the value $n$) is very small or very large.

In our example, we had $N = 12$, and we can draw up a table of the chance of getting $n$ heads, assuming $H_0$ holds:

| $n$ | probability |
|---|---|
| 0 | 0.000244141 |
| 1 | 0.002929688 |
| 2 | 0.016113281 |
| 3 | 0.053710938 |
| 4 | 0.120849609 |
| 5 | 0.193359375 |
| 6 | 0.225585938 |
| 7 | 0.193359375 |
| 8 | 0.120849609 |
| 9 | 0.053710938 |
| 10 | 0.016113281 |
| 11 | 0.002929688 |
| 12 | 0.000244141 |

- We observed $n = 9$. Thus "extreme or more" would be $n = 9, 10, 11$ or $12$, or also $n = 3, 2, 1$ or $0$.
- The total probability of these is 14.6%.
- So we do not reject $H_0$ at the 5% level.

# Coin example: $p$-values

It seems intuitively obvious (and can be Mathematically justified) that when testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ we should reject $H_0$ if the observed number of heads (the value $n$) is very small or very large. In our example, we had $N = 12$, and we can draw up a table of the chance of getting $n$ heads, assuming $H_0$ holds:

| $n$ | probability |
|---|---|
| 0 | 0.000244141 |
| 1 | 0.002929688 |
| 2 | 0.016113281 |
| 3 | 0.053710938 |
| 4 | 0.120849609 |
| 5 | 0.193359375 |
| 6 | 0.225585938 |
| 7 | 0.193359375 |
| 8 | 0.120849609 |
| 9 | 0.053710938 |
| 10 | 0.016113281 |
| 11 | 0.002929688 |
| 12 | 0.000244141 |

- We observed $n = 9$. Thus "extreme or more" would be $n = 9, 10, 11$ or $12$, or also $n = 3, 2, 1$ or $0$.
- The total probability of these is 14.6%.
- So we do not reject $H_0$ at the 5% level.

# Coin example: $p$-values

It seems intuitively obvious (and can be Mathematically justified) that when testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ we should reject $H_0$ if the observed number of heads (the value $n$) is very small or very large. In our example, we had $N = 12$, and we can draw up a table of the chance of getting $n$ heads, assuming $H_0$ holds:

| $n$ | probability |
|---|---|
| 0 | 0.000244141 |
| 1 | 0.002929688 |
| 2 | 0.016113281 |
| 3 | 0.053710938 |
| 4 | 0.120849609 |
| 5 | 0.193359375 |
| 6 | 0.225585938 |
| 7 | 0.193359375 |
| 8 | 0.120849609 |
| 9 | 0.053710938 |
| 10 | 0.016113281 |
| 11 | 0.002929688 |
| 12 | 0.000244141 |

- We observed $n = 9$. Thus "extreme or more" would be $n = 9, 10, 11$ or 12, or also $n = 3, 2, 1$ or 0.
- The total probability of these is 14.6%.
- So we do not reject $H_0$ at the 5% level.

# Coin example: $p$-values

It seems intuitively obvious (and can be Mathematically justified) that when testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ we should reject $H_0$ if the observed number of heads (the value $n$) is very small or very large. In our example, we had $N = 12$, and we can draw up a table of the chance of getting $n$ heads, assuming $H_0$ holds:

| $n$ | probability |
|---|---|
| 0 | 0.000244141 |
| 1 | 0.002929688 |
| 2 | 0.016113281 |
| 3 | 0.053710938 |
| 4 | 0.120849609 |
| 5 | 0.193359375 |
| 6 | 0.225585938 |
| 7 | 0.193359375 |
| 8 | 0.120849609 |
| 9 | 0.053710938 |
| 10 | 0.016113281 |
| 11 | 0.002929688 |
| 12 | 0.000244141 |

- We observed $n = 9$. Thus "extreme or more" would be $n = 9, 10, 11$ or 12, or also $n = 3, 2, 1$ or 0.
- The total probability of these is 14.6%.
- So we do not reject $H_0$ at the 5% level.

# Coin example: $p$-values

It seems intuitively obvious (and can be Mathematically justified) that when testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ we should reject $H_0$ if the observed number of heads (the value $n$) is very small or very large. In our example, we had $N = 12$, and we can draw up a table of the chance of getting $n$ heads, assuming $H_0$ holds:

| $n$ | probability |
|---|---|
| 0 | 0.000244141 |
| 1 | 0.002929688 |
| 2 | 0.016113281 |
| 3 | 0.053710938 |
| 4 | 0.120849609 |
| 5 | 0.193359375 |
| 6 | 0.225585938 |
| 7 | 0.193359375 |
| 8 | 0.120849609 |
| 9 | 0.053710938 |
| 10 | 0.016113281 |
| 11 | 0.002929688 |
| 12 | 0.000244141 |

- We observed $n = 9$. Thus "extreme or more" would be $n = 9, 10, 11$ or 12, or also $n = 3, 2, 1$ or 0.
- The total probability of these is 14.6%.
- So we do not reject $H_0$ at the 5% level.

# One-sided test

Maybe we prefer to test

$$H_1 : p > 1/2.$$

This gives a "one-tailed test", so values as or more extreme that $n = 9$ are now only $n = 10, n = 11, n = 12$.

This gives the $p$-value of 7.3%. So we still do not reject $H_0$.

# One-sided test

Maybe we prefer to test

$$H_1 : p > 1/2.$$

This gives a "one-tailed test", so values as or more extreme that $n = 9$ are now only $n = 10, n = 11, n = 12$.

This gives the $p$-value of 7.3%. So we still do not reject $H_0$.

# One-sided test

Maybe we prefer to test

$$H_1 : p > 1/2.$$

This gives a "one-tailed test", so values as or more extreme that $n = 9$ are now only $n = 10, n = 11, n = 12$.
This gives the $p$-value of 7.3%. So we still do not reject $H_0$.

# A different experiment

*I suspect a coin I have is biased. I toss it until I have seen 3 tails: I observe HHTHHHTHHHHT, which is 9 heads.*

The probability model is now different.

- I am interested in the probability of seeing $n$ heads before the 3rd tail is thrown.
- That's the same as tossing the coin $n+2$ times and getting exactly 2 tails, and then throwing a further tail.

$$f(n|p) = \binom{n+2}{n} p^n (1-p)^2 \times (1-p).$$

If $H_0$ is true then

$$f(n) = \binom{n+2}{n} \left(\frac{1}{2}\right)^{n+3}.$$

# A different experiment

*I suspect a coin I have is biased. I toss it until I have seen 3 tails: I observe HHTHHHTHHHHT, which is 9 heads.*

The probability model is now different.

- I am interested in the probability of seeing $n$ heads before the 3rd tail is thrown.
- That's the same as tossing the coin $n+2$ times and getting exactly 2 tails, and then throwing a further tail.

$$f(n|p) = \binom{n+2}{n} p^n (1-p)^2 \times (1-p).$$

If $H_0$ is true then

$$f(n) = \binom{n+2}{n} \left(\frac{1}{2}\right)^{n+3}.$$

# A different experiment

*I suspect a coin I have is biased. I toss it until I have seen 3 tails: I observe HHTHHHTHHHHT, which is 9 heads.*

The probability model is now different.

- I am interested in the probability of seeing $n$ heads before the 3rd tail is thrown.
- That's the same as tossing the coin $n + 2$ times and getting exactly 2 tails, and then throwing a further tail.

$$f(n|p) = \binom{n+2}{n} p^n (1-p)^2 \times (1-p).$$

If $H_0$ is true then

$$f(n) = \binom{n+2}{n} \left(\frac{1}{2}\right)^{n+3}.$$

# A different experiment

> *I suspect a coin I have is biased. I toss it until I have seen 3 tails: I observe HHTHHHTHHHHT, which is 9 heads.*

The probability model is now different.

- I am interested in the probability of seeing $n$ heads before the 3rd tail is thrown.
- That's the same as tossing the coin $n+2$ times and getting exactly 2 tails, and then throwing a further tail.

$$f(n|p) = \binom{n+2}{n} p^n (1-p)^2 \times (1-p).$$

If $H_0$ is true then

$$f(n) = \binom{n+2}{n} \left(\frac{1}{2}\right)^{n+3}.$$

# New $p$-values

For the next experiment, we could get any number of heads before the 3rd tail, so "more extreme" now means obtaining 9 or more heads.

- The sum of all these probabilities is 3.3%.
- So the the 5% level this gives evidence to reject $H_0$.

What's odd is that the *data* was exactly the same as before. So by changing the experimental design, we seem to have changed the statistical significance of the result.

This seems like nonsense to me. (For example, suppose you were spying on me tossing the coin, and your video link happened to break exactly on the 12th throw. Should your interpretation of the data change just because I may or may not have stopped the experiment after you stopped watching?)

# New $p$-values

For the next experiment, we could get any number of heads before the 3rd tail, so "more extreme" now means obtaining 9 or more heads.

- The sum of all these probabilities is 3.3%.
- So the the 5% level this gives evidence to reject $H_0$.

What's odd is that the *data* was exactly the same as before. So by changing the experimental design, we seem to have changed the statistical significance of the result.

This seems like nonsense to me. (For example, suppose you were spying on me tossing the coin, and your video link happened to break exactly on the 12th throw. Should your interpretation of the data change just because I may or may not have stopped the experiment after you stopped watching?)

# New $p$-values

For the next experiment, we could get any number of heads before the 3rd tail, so "more extreme" now means obtaining 9 or more heads.

- The sum of all these probabilities is 3.3%.
- So the the 5% level this gives evidence to reject $H_0$.

What's odd is that the *data* was exactly the same as before. So by changing the experimental design, we seem to have changed the statistical significance of the result.

This seems like nonsense to me. (For example, suppose you were spying on me tossing the coin, and your video link happened to break exactly on the 12th throw. Should your interpretation of the data change just because I may or may not have stopped the experiment after you stopped watching?)

# New $p$-values

For the next experiment, we could get any number of heads before the 3rd tail, so "more extreme" now means obtaining 9 or more heads.

- The sum of all these probabilities is 3.3%.
- So the the 5% level this gives evidence to reject $H_0$.

What's odd is that the *data* was exactly the same as before. So by changing the experimental design, we seem to have changed the statistical significance of the result.

This seems like nonsense to me. (For example, suppose you were spying on me tossing the coin, and your video link happened to break exactly on the 12th throw. Should your interpretation of the data change just because I may or may not have stopped the experiment after you stopped watching?)

# New $p$-values

For the next experiment, we could get any number of heads before the 3rd tail, so "more extreme" now means obtaining 9 or more heads.

- The sum of all these probabilities is 3.3%.
- So the the 5% level this gives evidence to reject $H_0$.

What's odd is that the *data* was exactly the same as before. So by changing the experimental design, we seem to have changed the statistical significance of the result.

This seems like nonsense to me. (For example, suppose you were spying on me tossing the coin, and your video link happened to break exactly on the 12th throw. Should your interpretation of the data change just because I may or may not have stopped the experiment after you stopped watching?)

# Likelihoods and common tests

Almost all the standard textbook statistical tests arise from "(Generalised) Likelihood Ratio Tests" where we compare the likelihood of the data if $H_0$ is true against the likelihood of the data if $H_1$ is true, and reject if this ratio is high.

What I like about this is that again you can use one simple principle (which can even, under special conditions, be proved to be "the best test") to justify a lot of elementary hypothesis testing.
Suddenly statistics does not seem as *ad hoc* as it might.

# Some messages about hypothesis testing

When you perform a test, say from a textbook, keep in mind:

- What are the assumptions about the data? Are they appropriate?
- What are $H_0$ and $H_1$.
- Is rejecting or accepting $H_0$ (against $H_1$) actually what you want to do?

Particularly important is what $p$-values are.

- Suppose we find a $p$-value of 2%.
- This means that, *if* $H_0$ is true, then the chance of seeing data as or more extreme than the data we have, is 2%.
- This is absolutely not "the probability that $H_0$ is true is 2%".

At the 5% level, even if $H_0$ is true, we expect by chance alone to reject $H_0$ about 5% of the time. One in twenty times we'll get "a statistically significant" result just by luck.

# Some messages about hypothesis testing

When you perform a test, say from a textbook, keep in mind:

- What are the assumptions about the data? Are they appropriate?
- What are $H_0$ and $H_1$.
- Is rejecting or accepting $H_0$ (against $H_1$) actually what you want to do?

Particularly important is what $p$-values are.

- Suppose we find a $p$-value of 2%.
- This means that, *if* $H_0$ is true, then the chance of seeing data as or more extreme than the data we have, is 2%.
- This is absolutely not "the probability that $H_0$ is true is 2%".

At the 5% level, even if $H_0$ is true, we expect by chance alone to reject $H_0$ about 5% of the time. One in twenty times we'll get "a statistically significant" result just by luck.

# A better way: Bayesian statistics

Do we have time?

# A better way: Bayesian statistics

Remember the likelihood:

$$\mathrm{lik}(p) = f(n|p).$$

What, intuitively, we want to know is "what is the probability distribution of $p$, given the data we have?"

Bayes Theorem allows us to find this:

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

There are two problems:

- What is $\mathbb{P}(p)$? We need a "prior" belief about what $p$ is. There is a lot of literature on this, and I think it's considered less philosophically suspect than it used to be.
- What is $\mathbb{P}(n)$? This is the "total probability" of seeing our data, averaged over all possible values of $p$. Usually this is impossible to find, except by complicated numerical methods. But in 2020 we now have good software to do this sort of thing.

# A better way: Bayesian statistics

Remember the likelihood:

$$\text{lik}(p) = f(n|p).$$

What, intuitively, we want to know is "what is the probability distribution of $p$, given the data we have?"

Bayes Theorem allows us to find this:

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

There are two problems:

- What is $\mathbb{P}(p)$? We need a "prior" belief about what $p$ is. There is a lot of literature on this, and I think it's considered less philosophically suspect than it used to be.
- What is $\mathbb{P}(n)$? This is the "total probability" of seeing our data, averaged over all possible values of $p$. Usually this is impossible to find, except by complicated numerical methods. But in 2020 we now have good software to do this sort of thing.

# A better way: Bayesian statistics

Remember the likelihood:

$$\text{lik}(p) = f(n|p).$$

What, intuitively, we want to know is "what is the probability distribution of $p$, given the data we have?"
Bayes Theorem allows us to find this:

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

There are two problems:

- What is $\mathbb{P}(p)$? We need a "prior" belief about what $p$ is. There is a lot of literature on this, and I think it's considered less philosophically suspect than it used to be.
- What is $\mathbb{P}(n)$? This is the "total probability" of seeing our data, averaged over all possible values of $p$. Usually this is impossible to find, except by complicated numerical methods. But in 2020 we now have good software to do this sort of thing.

# A better way: Bayesian statistics

Remember the likelihood:

$$\text{lik}(p) = f(n|p).$$

What, intuitively, we want to know is "what is the probability distribution of $p$, given the data we have?"
Bayes Theorem allows us to find this:

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

There are two problems:

- What is $\mathbb{P}(p)$? We need a "prior" belief about what $p$ is. There is a lot of literature on this, and I think it's considered less philosophically suspect than it used to be.
- What is $\mathbb{P}(n)$? This is the "total probability" of seeing our data, averaged over all possible values of $p$. Usually this is impossible to find, except by complicated numerical methods. But in 2020 we now have good software to do this sort of thing.

# A better way: Bayesian statistics

Remember the likelihood:

$$\text{lik}(p) = f(n|p).$$

What, intuitively, we want to know is "what is the probability distribution of $p$, given the data we have?"
Bayes Theorem allows us to find this:

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

There are two problems:

- What is $\mathbb{P}(p)$? We need a "prior" belief about what $p$ is. There is a lot of literature on this, and I think it's considered less philosophically suspect than it used to be.
- What is $\mathbb{P}(n)$? This is the "total probability" of seeing our data, averaged over all possible values of $p$. Usually this is impossible to find, except by complicated numerical methods. But in 2020 we now have good software to do this sort of thing.

# For the coin example

We have

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

We know that $p$ is a value between 0 and 1 and the total probability should be 1, so

$$\int_0^1 \mathbb{P}(p|n)\ dp = 1.$$

This allows us to calculate $\mathbb{P}(n)$ is a roundabout way. We know that

$$f(n|p) \propto p^n(1-p)^{N-n}.$$

Let's impose a "uniform prior", $\mathbb{P}(p) = 1$. This reflects a lack of knowledge about the coin before we did an experiment.
So $\mathbb{P}(p|n) = \lambda p^n(1-p)^{N-p}$ for some constant $\lambda$ chosen to make the integral equal to one. It turns out that

$$\lambda = \frac{n!(N-n)!}{(N+1)!}.$$

# For the coin example

We have

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

We know that $p$ is a value between 0 and 1 and the total probability should be 1, so

$$\int_0^1 \mathbb{P}(p|n)\ dp = 1.$$

This allows us to calculate $\mathbb{P}(n)$ is a roundabout way. We know that

$$f(n|p) \propto p^n(1-p)^{N-n}.$$

Let's impose a "uniform prior", $\mathbb{P}(p) = 1$. This reflects a lack of knowledge about the coin before we did an experiment.

So $\mathbb{P}(p|n) = \lambda p^n(1-p)^{N-p}$ for some constant $\lambda$ chosen to make the integral equal to one. It turns out that

$$\lambda = \frac{n!(N-n)!}{(N+1)!}.$$

# For the coin example

We have

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

We know that $p$ is a value between 0 and 1 and the total probability should be 1, so

$$\int_0^1 \mathbb{P}(p|n)\ dp = 1.$$

This allows us to calculate $\mathbb{P}(n)$ is a roundabout way. We know that

$$f(n|p) \propto p^n(1-p)^{N-n}.$$

Let's impose a "uniform prior", $\mathbb{P}(p) = 1$. This reflects a lack of knowledge about the coin before we did an experiment.

So $\mathbb{P}(p|n) = \lambda p^n(1-p)^{N-p}$ for some constant $\lambda$ chosen to make the integral equal to one. It turns out that

$$\lambda = \frac{n!(N-n)!}{(N+1)!}.$$

# For the coin example

We have

$$\mathbb{P}(p|n) = \frac{f(n|p)\mathbb{P}(p)}{\mathbb{P}(n)}.$$

We know that $p$ is a value between 0 and 1 and the total probability should be 1, so

$$\int_0^1 \mathbb{P}(p|n)\ dp = 1.$$

This allows us to calculate $\mathbb{P}(n)$ is a roundabout way. We know that

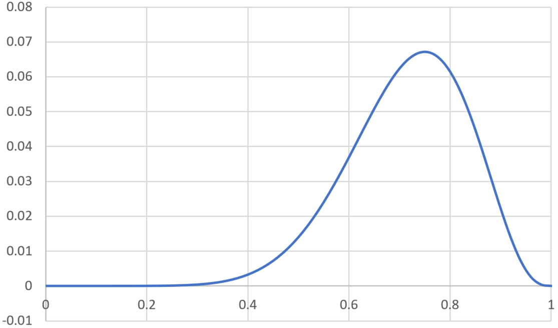
$$f(n|p) \propto p^n(1-p)^{N-n}.$$

Let's impose a "uniform prior", $\mathbb{P}(p) = 1$. This reflects a lack of knowledge about the coin before we did an experiment.
So $\mathbb{P}(p|n) = \lambda p^n(1-p)^{N-p}$ for some constant $\lambda$ chosen to make the integral equal to one. It turns out that

$$\lambda = \frac{n!(N-n)!}{(N+1)!}.$$

# For the coin example (cont.)

We have $N = 12$ and $n = 9$. Then $\mathbb{P}(p|n)$, the "posterior distribution", looks like:




This distribution does suggest that the coin is biased.