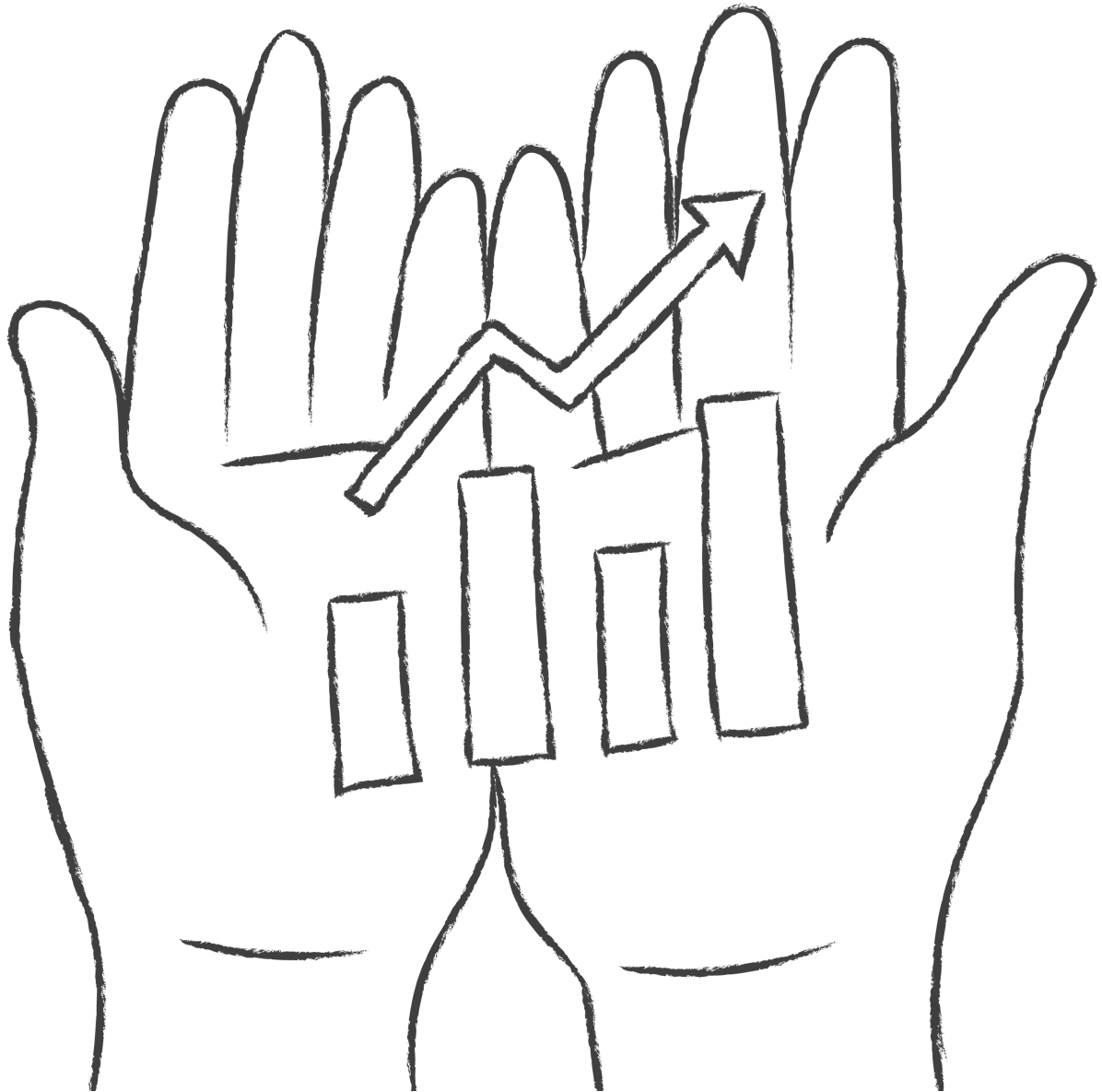
Hypothesis testing

Rachel Warnock

03.04.2023



Today

schedule	start.time
Course introduction	10:00
Group exercise: mini reading group	
Break (15 mins)	
Introduction to hypothesis testing	11:00
Lunch (45 mins)	12.00
Group exercise: defining hypotheses	13:00
(Re)Introduction to the normal distribution	13:30

Times are approximate.

About this course

The course focuses on statistical **hypothesis testing** and **reproducibility** in science.

You will learn:

- how develop and test your own hypotheses
- perform basic statistical tests
- apply this knowledge to reproduce (and potentially improve) published results.

Each participant is assigned to a working and each group is assigned a published scientific paper.

Group 1

Isotopic and anatomical evidence of an herbivorous diet in the Early Tertiary giant bird Gastornis

- Lars (online?)
- Nida
- Toshiro

Group 2

High coral diversity is coupled with reef-building capacity during the Late Oligocene warming

- Bastian
- Taeya

Course evaluation

The goal is to reproduce the results of a published scientific article.

On **Friday afternoon** each group will present their findings.

You can use this *Google Slides Template* to prepare your presentation.

Within your working groups, try to do everything within time allocated for the course.

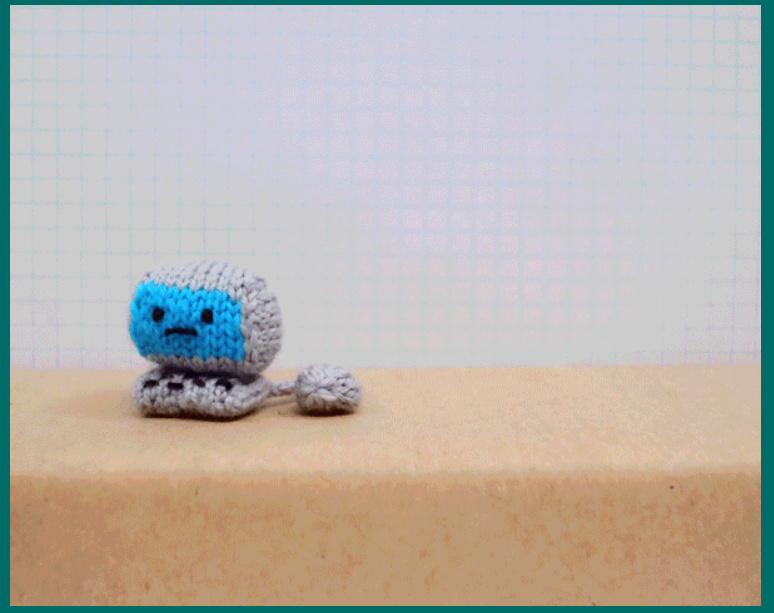
Will I have to use R?

Not necessarily.

The focus of the is on the **concepts** behind hypothesis testing and reproducibility, not programming, although both things are easier if you use R.

It's up to each group how you divide the tasks.

I'm here all week to help!




FULLY EXPECTING ~~JOY~~ HATE THIS CLASS!
STATS
CODE
@allison_horst

Why study statistics?

- Humans are **biased**.
- Science is **complicated** and data is **messy**.
- It is deeply intertwined with **research design**.
- It makes the **literature** more accessible, since many papers you encounter will include statistics.
- Where there's **data** there's statistics!
- Having some statistical knowledge gives you a **superpower** 🦸

Why does palaeobiology feature so much statistics?

What do you think? 🤔

Mini reading group

First group exercise

In your working groups:

Introduce yourselves and discuss your paper.

- What is the paper about?
- What was the general aim?
- Did you like / not like it?

To report back: prepare a three sentence summary.

15:00

Time for a break

Introduction to hypothesis testing

Day 1 - Part 2

Objectives

- Statistical hypotheses
- Null vs. alternative hypotheses
- Type I and type II erros
- Significance and *p*-values

Further reading *Learning Statistics with R*

Hypothesis testing

At its core, hypothesis testing is a simple idea – you, the researcher, have some theory about how the world works, and want to determine whether or not available data support your theory.

A **research hypothesis** involves make a substantive*, testable scientific claim.

*having a firm basis in reality

Some examples of research hypotheses

🎵 **Listening to music reduces your ability to concentrate on other things** – this is a claim about the causal relationship between two psychologically meaningful concepts (listening to music and paying attention to things). This is a reasonable scientific hypothesis.

🧠 **Intelligence is related to personality** – this is another relation claim, this time about two psychological constructs (intelligence and personality). The claim is weaker because it proposes correlation not causation.

🧠 **Intelligence is the speed of information processing** – this is not a relation claim, it's an ontological claim about the definition of intelligence. This is a research questions, but it's usually more straightforward to ask *"does X affect y?"* than *"what is X?"*.

Most everyday research questions tend to be relational.

Examples *Learning Statistics with R*

What are examples of research hypotheses in palaeobiology?

Examples of statements that are not research hypotheses

Love is a battlefield – too vague to be testable. A research hypothesis can be vague to some extent, but you have to be able to break into down into testable theories. This statement can't be converted into a concrete research design.

The first rule of tautology* club is the first rule of tautology club. – this is by definition true. Your hypotheses must have the possibility of being wrong!

More people in my experiment will say "yes" than say "no" – this is a claim about the data, not about a scientific theory. This hypothesis actually sounds more like a statistical hypothesis.

Examples *Learning Statistics with R*

tautology = saying the same thing twice in different words

Research vs. statistical hypotheses

So to recap, a **research hypothesis** involves make a substantive, testable scientific claim, i.e., has a firm basis in reality.

Sometimes a research hypothesis can be a bit fuzzy, but ultimately research hypotheses are scientific claims.

In contrast, a **statistical hypothesis** must be mathematically precise and must correspond to specific claims about the data generating mechanism (i.e., the underlying population).

Statistical hypothesis example

Say we have a species of bird and the birds can be either **blue** or **red**. We want to know, does being blue confer some advantage?

- Birds have the same chance of being blue as being red.
- More birds are blue.
- More birds are red.

Statistical hypothesis example

Evidence for this might come from the numbers for each color.

θ = the probability of being blue.

- Birds have the same chance of being blue as being red. If this is true, then $\theta = 0.5$.
- More birds are blue. If this is true, $\theta > 0.5$.
- More birds are red. If this is true, $\theta < 0.5$.
- If I know a different number of birds are blue than are red, but I didn't keep my field notes well organised, I might not know which way round the numbers go. Then, $\theta \neq 0.5$.

Statistical hypothesis example

These examples are **statistical hypotheses** because they are statements about a 'population' parameter and are meaningfully relevant to the research hypothesis.

Research hypothesis: being blue is better. Statistical hypothesis: $\theta > 0.5$.

A statistical (hypothesis) test is a *test* of the statistical hypothesis, not the research hypothesis.

Null vs. alternative hypotheses

The **null hypothesis** H_0 corresponds to the exact opposite of the thing I want to believe.

The thing I'm actually interested in is the **alternative hypothesis** H_1.

In our example, the null hypothesis is $\theta = 0.5$, since that's what we'd except if there was no advantage of being blue. The alternative is $\theta > 0.5$.

The goal is not to show that the alternative hypothesis is (probably) true, the goal is to show that the null hypothesis is (probably) false.

Hypotheses

Second group exercise

Discuss the hypotheses in your paper.

Try to identify the following:

- The research hypothesis
- The alternative hypothesis
- The null hypothesis

Make a note to add your presentation. Don't worry about getting it perfect, you can go back and refine it later.

If your paper includes more than one, just pick one.

10:00

Two types of errors

If we flip a coin 10 times and all 10 times we get heads, this is pretty strong evidence that our coin is biased. But there's a 1 in 1024 chance that this would happen even if the coin was fair.

We **always** have to accept there's a chance that the results of any experiment are wrong.

The goal behind statistical hypothesis testing is not to *eliminate* but to *minimize* errors.

Two types of errors

After we run the test, one of four things might have happened:

- H_0 is true $-$ correct decision
- H_0 is true $-$ incorrect decision (type I error)
- H_0 is false $-$ correct decision
- H_0 is false $-$ incorrect decision (type II error)

Two types of errors

If we reject a null hypothesis that is actually true, then we have a **type I error**.

If we retain the null hypothesis that is actually false, then we have a **type II error**.

One of the most important design principles of hypothesis testing is to control the probability α of a **type I error**. α is called the significance level. By convention we often use α of 0.05, 0.01, 0.001.

A hypothesis test that is said to have a significance level α has a type I error rate is no larger than α.

Two types of errors

What about type II error rate? We care about this probability β too. We refer to the **power** of the test, which is the probability that we reject the null hypothesis when it really is false, which is $1 - \beta$. A powerful test has a small value of β. Note we don't have a corresponding level for β.

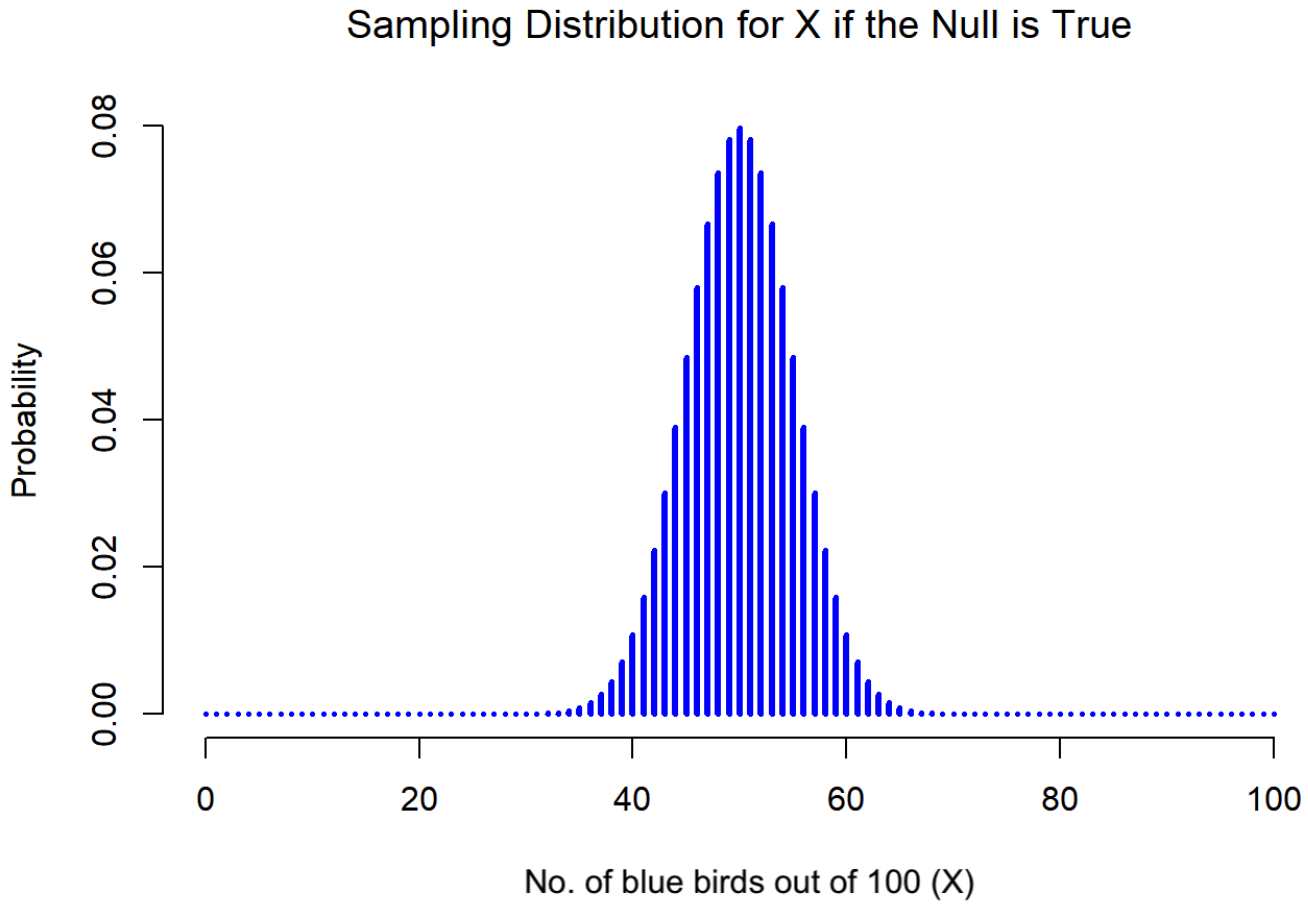
Statistical tests are designed to minimise α, not β.

Sampling distributions

If the probability of being a blue bird is $\theta = 0.5$, what would we expect the data to look like?

We can also say $X/N = 0.5$ (approximately). This is our **test statistic**.

We need to determine what the **sampling distribution of the test statistic** would be if the null hypothesis was true. This distribution tells us what values we can expect if H_0 is true. We use this a tool for assessing how closely the null agrees with our data.



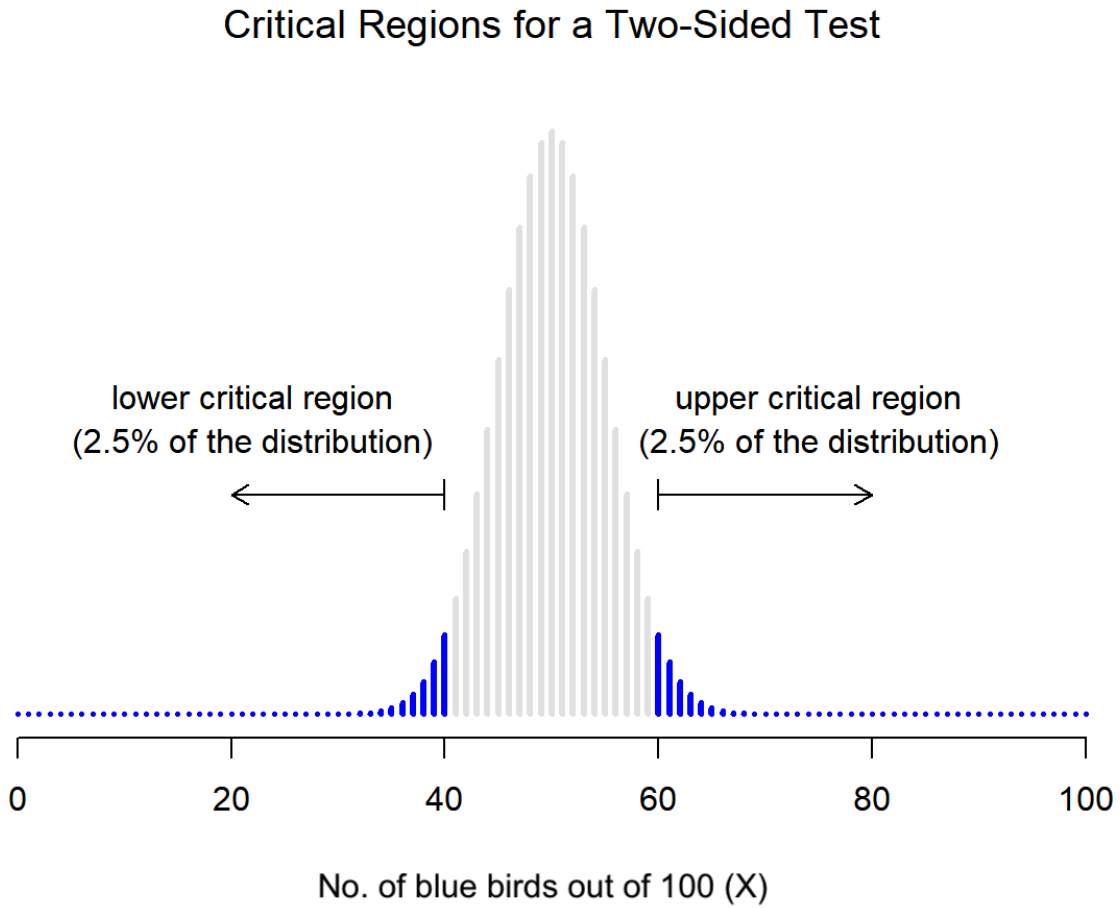
The null hypothesis predicts that X is *binomially distributed*. It says $X = 50/100$ is the most likely outcome, so we'd expect to see somewhere between 40 and 60 blue birds.

Critical regions and critical values

X should be very big or very small in order to reject the null hypothesis.

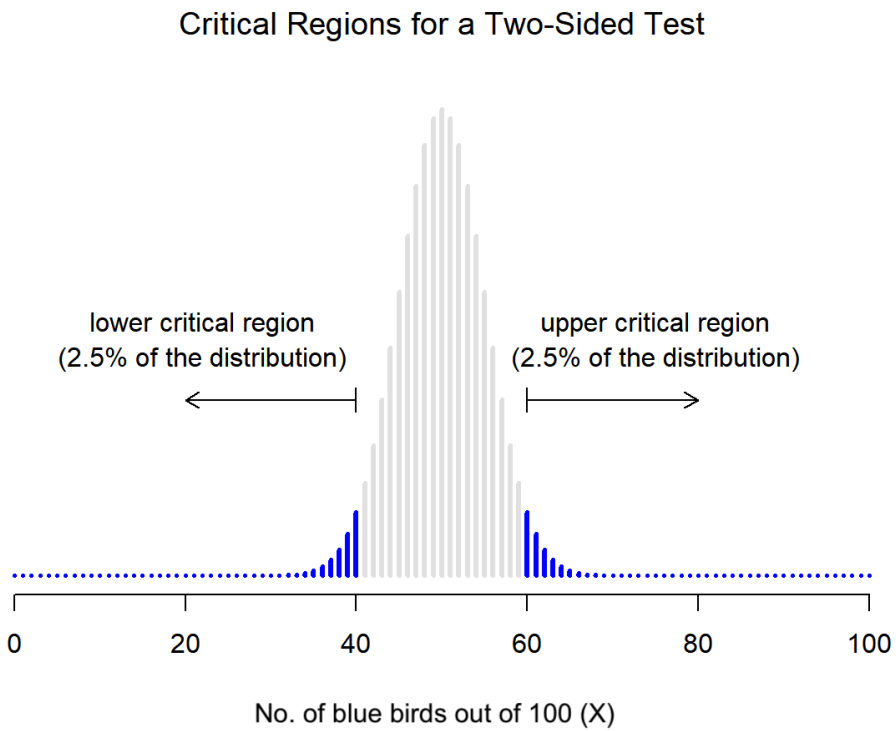
If the null hypothesis is true, the sampling distribution of X is Binomial $(0.5, N)$.

If $\alpha = 0.05$, the critical region must cover 5% of this sampling distribution.



Our critical region consists of the most extreme values, known as the **tails** of the distribution.

The **critical region** corresponds to the values of X for which we would reject the null hypothesis, while the **sampling distribution** describes the probability that we would obtain a particular value of X if the null hypothesis were actually true.

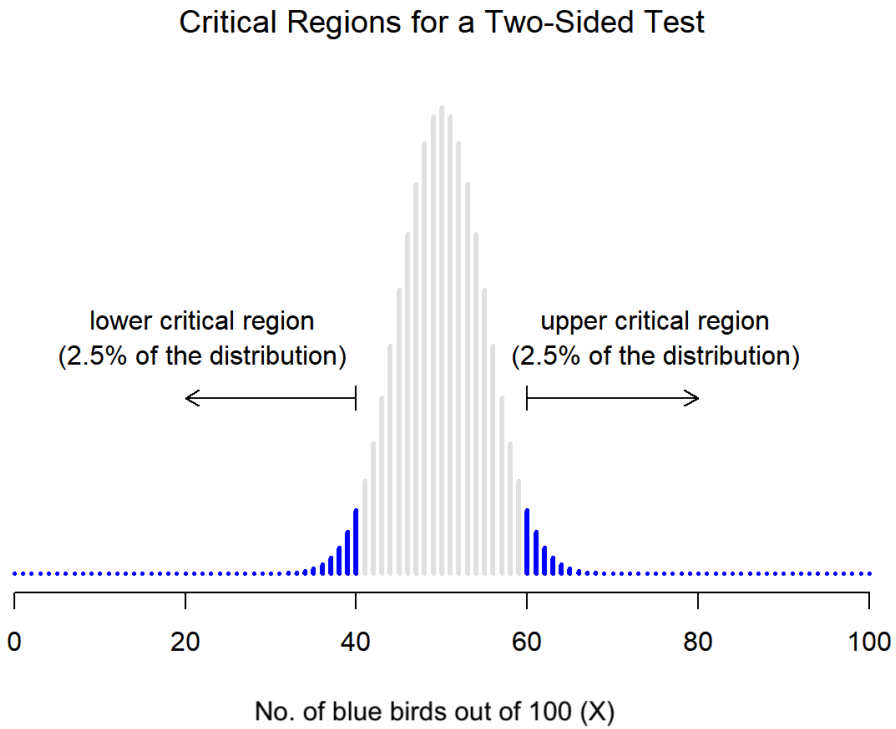


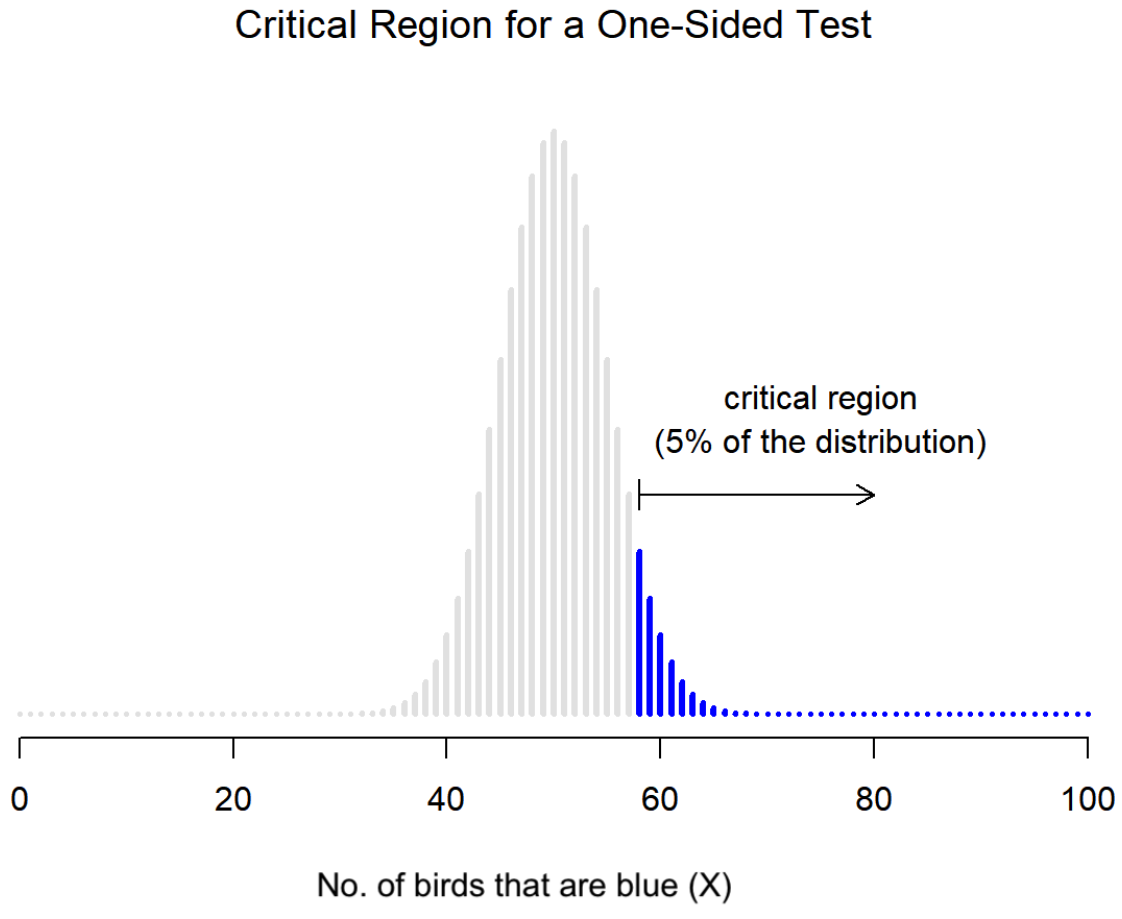
For $\alpha = .05$, our critical regions correspond to $X \leq 40$ and $X \geq 60$.

The numbers 40 and 60 are our **critical values**.

If the number of blue birds is between 41 and 59, then we should retain the null hypothesis (birds have the same chance of being blue as being red).

If the number of blue birds is between 0 to 40 **or** between 60 to 100, then we should reject the null hypothesis — this is a **two tailed test**.





The critical region for a **one sided test**. We would use this to test for $\theta > 0.5$ (i.e, more birds are blue).

Statisical significance

If the data allow us to reject the null hypothesis, we say that "the result is statistically significant", which is often shortened to "the result is significant".

This terminology reflects a time when "significant" meant something like "indicated", rather than its more recent meaning, which is closer to "important".

p-values

p can be defined to be the smallest type I error rate α that you are willing to tolerate if you want to reject the null hypothesis.

In the bird example, $X = 62$ blue birds gives us *p* = 0.021. The results can be interpreted as shown in the table, given *p* = 0.021.

$X = 97$ blue birds would give us *p* = 1.36 X 10^{-25}, which is a tiny, tiny type I error rate.

Value of α	Reject the null?
0.05	Yes
0.04	Yes
0.03	Yes
0.02	No
0.01	No

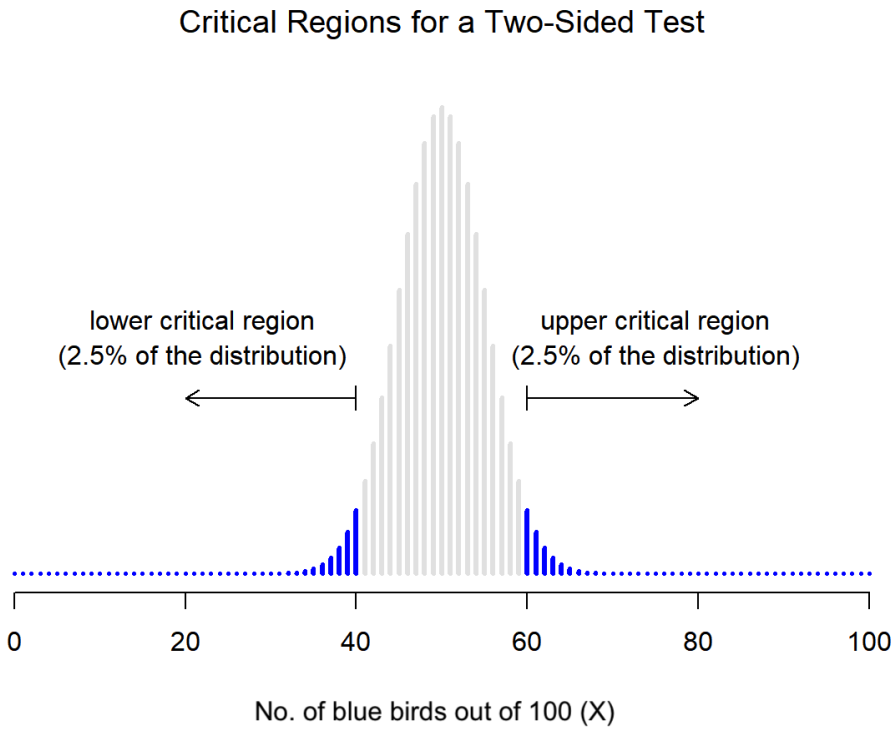
p can be calculated in R or looked up in a statistical table.

p-values

Recall that the critical region corresponds to the tails (extremes) of the distribution.

p can therefore also be defined as the probability of observing a test statistic that is (at least) as extreme as the one we actually get.

If the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.



p can be calculated in R or looked up in a statistical table.

Reporting *p-values*

Option 1 — you can state only that $p < \alpha$ for a significance level that you chose in advance, e.g., p<.05.

But this means we're being forced to treat p=.051 in a fundamentally different way to p=.049.

Option 2 — just report the actual p value and let the reader make up their own minds about what an acceptable type I error rate is.

But if you get p=.062, then it means that you have to be willing to tolerate a type I error rate of 6.2% to justify rejecting the null. If you personally find 6.2% intolerable, then you retain the null.

Intepreting *p*-values

Third group exercise

How have the authors in your paper interpreted the *p* values?

Spend 15 minutes putting together the 'Scientific background' and 'Hypothesis' slides.

15:00

Time for a break