A quick (re)tour of statistical hypothesis testing

What Rejecting a Hypothesis Tells Us?

Hypothesis: Electric vehicles (EVs) generate no pollution.

Rejecting this hypothesis tells us that some pollution is being generated.

It does not tell us how much pollution (small or large).

Statistical tests provide evidence against a hypothesis (statement), not its magnitude.

Statistical tests tell us how unlikely zero pollution is - not how large the pollution level is.

A quick (re)tour of statistical hypothesis testing

What Rejecting a Hypothesis Tells Us?

Hypothesis: Electric vehicles (EVs) generate no pollution.

Rejecting this hypothesis tells us that some pollution is being generated.

It does not tell us how much pollution (small or large).

Statistical tests provide evidence against a hypothesis (statement), not its magnitude.

Statistical tests tell us how unlikely zero pollution is - not how large the pollution level is.

NOTE: EVs do not produce direct emissions, but they can generate indirect pollution (e.g., manufacturing, infrastructure).

Statistical Hypothesis Testing: An Intimate Stranger

Widely used and often correctly applied.

Commonly interpreted, but rarely well understood.

The mechanics are familiar; the logic is not.

Statistical hypothesis testing is widely used and easy to apply, but its logic is often misunderstood.

When learned properly, it sharpens numerical intuition and conditional reasoning—not just statistical results.

Statistical Hypothesis Testing: An Intimate Stranger – WHY? Routine in use, mysterious in meaning!

This logic is often non-intuitive to people:

If we cannot assign probabilities to what the true population value is, we can assign probabilities to what it is unlikely to be.

(Frequentist) Statistical hypothesis testing rules out the implausible; it does not quantify belief in the plausible.

Frequentist Statistical hypothesis testing by asking: Which assumptions about the world are incompatible with what we observed?

Frequentist inference cannot assign probabilities to parameters, but it can identify which parameter values are unlikely because they would rarely produce the observed data.

Statistical Hypothesis Testing: An Intimate Stranger – WHY? Routine in use, mysterious in meaning!

Widely used and often correctly applied. Commonly interpreted, but rarely well understood.

Analogy: Signal Detection in a Noisy Instrument; Imagine you have a sensor (e.g., a thermometer). You want a room temperature to be 15.0°C on average. However, the thermometer is known to fluctuate randomly by a small amount due to measurement noise.

You take repeated readings and compute their average – say it was **16.5°C and not the desired 15.0°C.**

Null hypothesis (noise-only): The **true room temperature is 15.0°C**, and any variation in the readings is due solely to **random measurement noise**.

We cannot assign probabilities to temperatures, but we can rule out temperatures that would rarely produce the observed data (i.e., **16.5°C)**.

In other words, we can quantify how unlikely the observed data (**16.5°C)** would be **if the true temperature were 15.0 °C**.

From Research Questions to Statistical Questions

Humans are predominantly right-handed. *Do other animals exhibit handedness as well?* Bisazza et al. (1996) tested this possibility on the common toad.

They sampled (randomly) 18 toads from the wild. They wrapped a balloon around each individual's head and recorded which forelimb each toad used to remove the balloon.

Translating the research question into a statistical question:

Do right-handed and left-handed toads occur with equal frequency in the toad (statistical) population, or is one type more frequent than the other?

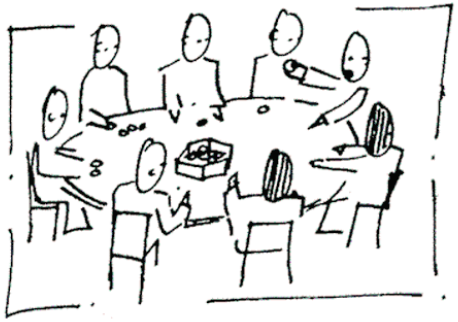
What types of evidence would be most informative for answering this question?

Do right-handed and left-handed toads occur with equal frequency in the toad (statistical) population, or is one type more frequent than the other?

[1] We don't know what the proportion is but is not likely to be 50% right- and 50% left-handed.

[2] We are 95% confident that the proportion of right-handed over left-handed toads varies between 60% and 90%.

[3] We are 95% confident that the proportion of right-handed over left-handed toads varies between 72% and 78%.



What types of evidence would be most informative for answering this question?

Statistical hypothesis testing

We don't know what the proportion is but is not likely to be 50% right- and 50% left-handed = The proportion of right-handed over left-handed toads differs significantly from 0.5 (50%).

"Qualitative" statement

Parameter estimate & uncertainty

We are 95% confident that the proportion of right-handed over left-handed toads varies between 60% and 90%.

We are 95% confident that the proportion of right-handed over left-handed toads varies between 72% and 78%.

"Quantitative" statement

What types of evidence would be most informative for answering this question?

All these three answers provide evidence towards handedness; but humans love yes/no answers (qualitative).

Statistical hypothesis testing

- **We don't know what the proportion is but is not likely to be 50% right- and 50% left-handed = The proportion of right-handed over left-handed toads differs significantly from 0.5 (50%).**

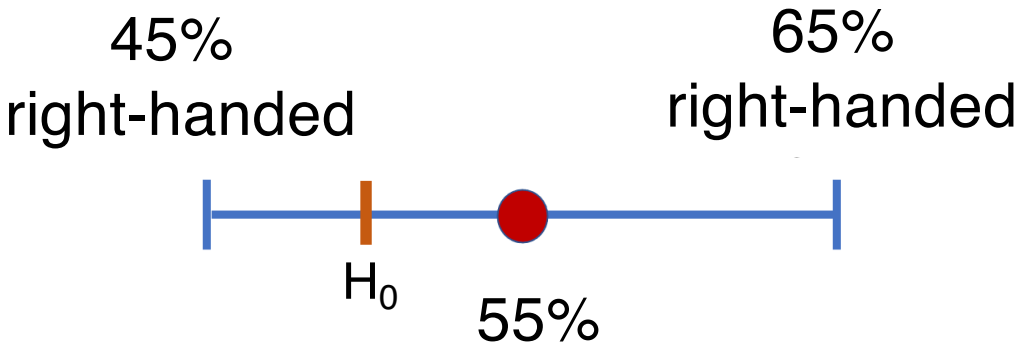
→ **"Qualitative" statement**

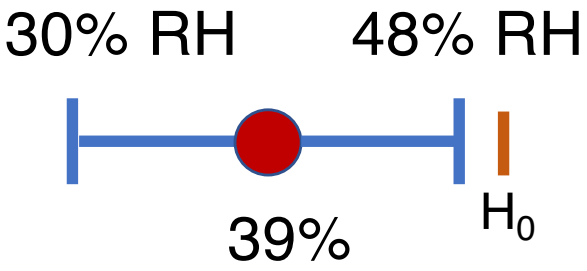
Parameter estimate & uncertainty

- **We are 95% confident that the proportion of right-handed over left-handed toads varies between 60% and 90%.**
- **We are 95% confident that the proportion of right-handed over left-handed toads varies between 72% and 78%.**

→ **"Quantitative" statement**

What Estimation and Hypothesis Testing Agree On - and Where They Differ





"Quantitative" statement

A confidence interval quantifies the uncertainty around an estimate by giving a range of plausible parameter values.

Don't reject H_0; $p > 0.05$

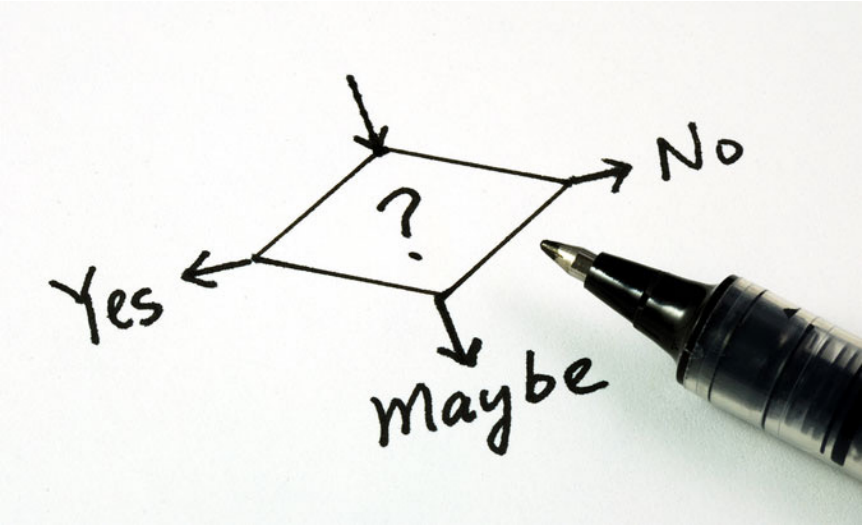
Reject H_0; $p < 0.05$

"Qualitative" Statement (reject / don't reject)

An Intuitive View of Hypothesis Testing

A road map to UNDERSTAND how *evidence-based* decisions / conclusions can be made when knowledge is incomplete.

Remember: Statistics is the science that assists in informing decision making when knowledge is incomplete (e.g., we can't know handedness of each toad on the planet).



Statistical Hypothesis Testing: An Intuitive Demonstration

A demonstration involves using reasoning, evidence, or examples to explain, illustrate, or make a concept clear, often through concrete examples or simple experiments.

Put simply, a demonstration means “to clearly show” (hopefully!).

Statistical Hypothesis Testing: A Very Simple Example

Humans are predominantly right-handed. *Do other animals exhibit handedness as well?* Bisazza et al. (1996) tested this possibility on the common toad.

They sampled (randomly) 18 toads from the wild. They wrapped a balloon around each individual's head and recorded which forelimb each toad used to remove the balloon.

Translating the research question into a statistical question:

Do right-handed and left-handed toads occur with equal frequency in the toad (statistical) population, or is one type more frequent than the other?

RESULTS: 14 toads were right-handed and four were left-handed. Are these results sufficient to generate evidence of handedness in toads?

What is a research (*not statistical*) hypothesis?

A hypothesis is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation; e.g., “animals other than humans also present handedness”.

"*A hypothesis is a proposition made as a basis for reasoning, without any assumption of its truth*” (Oxford dictionary).

Hypothes**e**s [plural form] can be thought as educated guesses that have not been supported by data yet.

Hypotheses ***cannot be proven*** right or wrong based on data alone. Instead, they can be ***supported*** (or not) by the data at hand, while remaining open to being refuted by future evidence.

Hypotheses, Theories and Laws: three different components

Hypotheses cannot be proven right or wrong based on data alone. Instead, they can be supported (or not) by the data at hand, while remaining open to being refuted by future evidence.

Strong research evidence is generated when several studies support (or refute) a particular hypothesis.

"A hypothesis is an idea that is offered or assumed with the intent of being tested. A theory is intended to explain processes already supported or substantiated by data and experimentation" (Marshall Sheperd):

A theory is a well-substantiated explanation for a natural phenomenon. And a law (gravity) is an observation (objects fall towards the ground).

Let’s take a break - 1 minute

Remember the two possible statistical hypotheses:

Null hypothesis (H_0): the proportion of right- and left-handed toads in the population **IS** equal.

Alternative hypothesis (H_A): the proportion of right- and left-handed toads in the population **IS NOT** equal.

The intuition behind the frequentist framework of statistical hypothesis testing

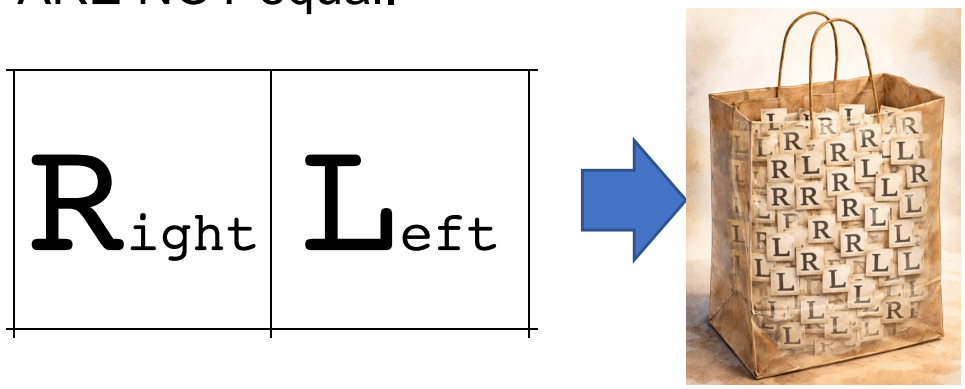
We can generate evidence for or against a biological hypothesis (e.g., handedness) using a simple computational thought experiment involving paper and a bag.

The idea is to assume a specific hypothesis is true (the null hypothesis) and then evaluate whether the observed outcome is consistent with that assumption. If it is not, we reject the null hypothesis in favour of the alternative hypothesis.

A frequentist statistical test is designed to assess how incompatible the data are with the null hypothesis (H_0).

Null hypothesis (H_0): % right- and left-handed toads in the population ARE equal.

Alternative hypothesis (H_A): % of right- and left-handed toads in the population ARE NOT equal.



The intuition behind the frequentist framework of statistical hypothesis testing

⇧

A theoretical statistical population in which 50% of observational units (toads) are left-handed and 50% are right-handed. This population is assumed to be mathematically infinite.

The intuition behind the frequentist framework of statistical hypothesis testing

Randomly draw one observational unit (a piece of paper) from the bag at a time (e.g., close your eyes and draw one).

A theoretical statistical population in which 50% of observational units (toads) are left-handed and 50% are right-handed. This population is assumed to be mathematically infinite.

The intuition behind the frequentist framework of statistical hypothesis testing

Randomly draw one observational unit (a piece of paper) from the bag at a time (e.g., close your eyes and draw one).

Record whether it indicates left or right, then return it to the bag (i.e., sampling with replacement). Repeat this process 18 times, corresponding to the number of toads in the study by Bisazza et al. (1996).

A theoretical statistical population in which 50% of observational units (toads) are left-handed and 50% are right-handed. This population is assumed to be mathematically infinite.

The intuition behind the frequentist framework of statistical hypothesis testing

A theoretical statistical population in which 50% of observational units (toads) are left-handed and 50% are right-handed. This population is assumed to be mathematically infinite.

Randomly draw one observational unit (a piece of paper) from the bag at a time (e.g., close your eyes and draw one).

Record whether it indicates left or right, then return it to the bag (i.e., sampling with replacement). Repeat this process 18 times, corresponding to the number of toads in the study by Bisazza et al. (1996).

1 sample: 14 R & 4 L
2 sample: 8 R & 10 L
.
.
.
Large number of samples (~Infinite)

Sampling distribution of the test statistic under the null (theoretical) population

The intuition behind the frequentist framework of statistical hypothesis testing

A theoretical statistical population in which 50% of observational units (toads) are left-handed and 50% are right-handed. This population is assumed to be mathematically infinite.

Randomly draw one observational unit (a piece of paper) from the bag at a time (e.g., close your eyes and draw one).

Record whether it indicates left or right, then return it to the bag (i.e., sampling with replacement). Repeat this process 18 times, corresponding to the number of toads in the study by Bisazza et al. (1996).

1 sample: 14 R & 4 L
2 sample: 8 R & 10 L
.
.
.
Large number of samples (~Infinite)

Sampling distribution of the test statistic under the null (theoretical) population

*Resampling (sampling with replacement) is important because it ensures that each selection of an observational unit (e.g., a piece of paper) is independent of the others. In other words, drawing one unit (L or R) does not affect the probability of subsequent draws.

This procedure also mimics sampling from a theoretically infinite population, where the composition of the population never changes.

The intuition behind the frequentist framework of statistical hypothesis testing

```
> Sample1 <- sample(c("L", "R"), size = 18, prob = c(0.5, 0.5), replace = TRUE)
> Sample1
 [1] "R" "L" "L" "L" "L" "R" "R" "R" "R" "R" "L" "L" "L" "L" "L" "R" "R" "L"
> sum(Sample1 == "R")
[1] 8
> sum(Sample1 == "L")
[1] 10
```

Sample 1

```
> Sample2 <- sample(c("L", "R"), size = 18, prob = c(0.5, 0.5), replace = TRUE)
> Sample2
 [1] "R" "R" "R" "L" "R" "R" "R" "R" "L" "L" "L" "L" "R" "L" "R" "R" "R" "R"
> sum(Sample2 == "R")
[1] 12
> sum(Sample2 == "L")
[1] 6
```

Sample 2

↓

etc

⇨ Assumed Model (50%/50%) under H_0

Grammar here matters!

Assumed Model
(50%/50%) ***under*** H_0

"under H_0" vs "for H_0"

"Under H_0" means *assuming the null hypothesis is true* and describing what the model, data, or distribution would look like in that hypothetical world. This is the **correct phrasing** in *frequentist* hypothesis testing.

"For H_0" sounds like we are *arguing in favor of* or *supporting* the null hypothesis, which frequentist tests do **not** do.

1 sample: 14 R & 4 L
2 sample: 8 R & 10 L
.
.
.
Large number of samples (~Infinite)

Sampling distribution for the test statistic of interest for the theoretical statistical population

How many samples contain 0 right-handed toads and 18 left-handed toads? 0.000004 or 0.0004%.

If we had drawn 1000000 samples from the population assumed under H_0, only 4 would had been 0 right-handed (the distribution is obviously symmetric).

Number of right-handed toads	Probability of those samples
0	**0.000004**
1	0.00007
2	0.0006
3	0.0031
4	0.0117
5	0.0327
6	0.0708
7	0.1214
8	0.1669
9	0.1855
10	0.1669
11	0.1214
12	0.0708
13	0.0327
14	0.0117
15	0.0031
16	0.0006
17	0.00007
18	**0.000004**
Total	1.0

1 sample: 14 R & 4 L
2 sample: 8 R & 10 L
.
.
.
Large number of samples (~Infinite)

Sampling distribution for the test statistic of interest for the theoretical statistical population

How many samples contain 0 right-handed toads and 18 left-handed toads? 0.000004 or 0.0004%.

How many samples contain 8 right-handed toads and 10 left-handed toads? 0.1669 or 16.69%

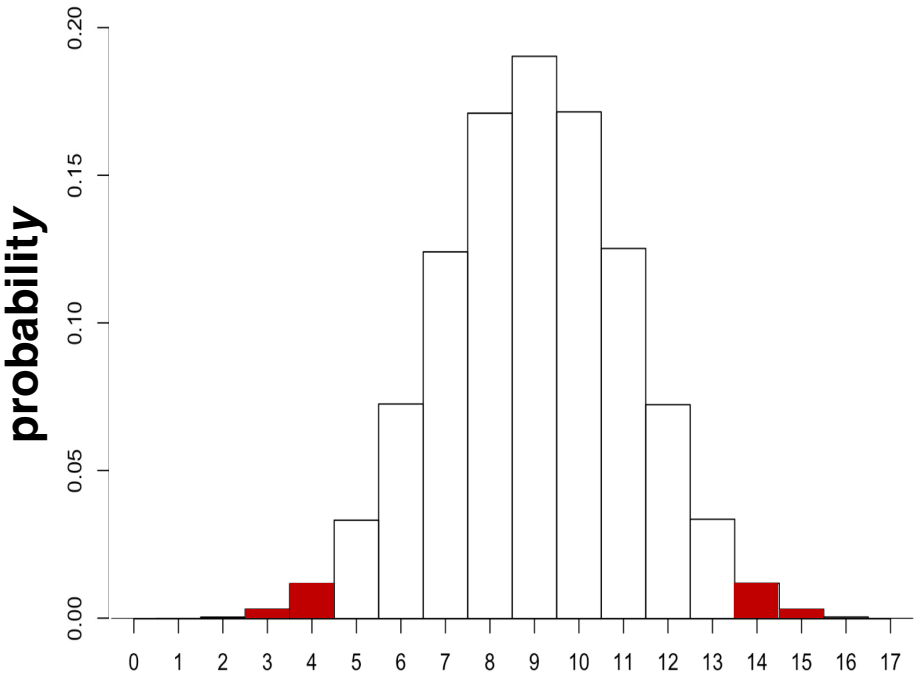
If we had drawn 1000000 samples from the population assumed under H_0, 166900 would had been 8 right-handed and 10 left-handed.

Number of right-handed toads	Probability of those samples
0	0.000004
1	0.00007
2	0.0006
3	0.0031
4	0.0117
5	0.0327
6	0.0708
7	0.1214
8	**0.1669**
9	0.1855
10	**0.1669**
11	0.1214
12	0.0708
13	0.0327
14	0.0117
15	0.0031
16	0.0006
17	0.00007
18	0.000004
Total	1.0

Number of right-handed toads	Probability
0	**0.000004**
1	**0.00007**
2	**0.0006**
3	**0.0031**
4	**0.0117**
5	0.0327
6	0.0708
7	0.1214
8	0.1669
9	0.1855
10	0.1669
11	0.1214
12	0.0708
13	0.0327
14	**0.0117**
15	**0.0031**
16	**0.0006**
17	**0.00007**
18	**0.000004**
Total	1.0

equal or smaller sum [P]=0.0155

equal or greater sum [P]=0.0155

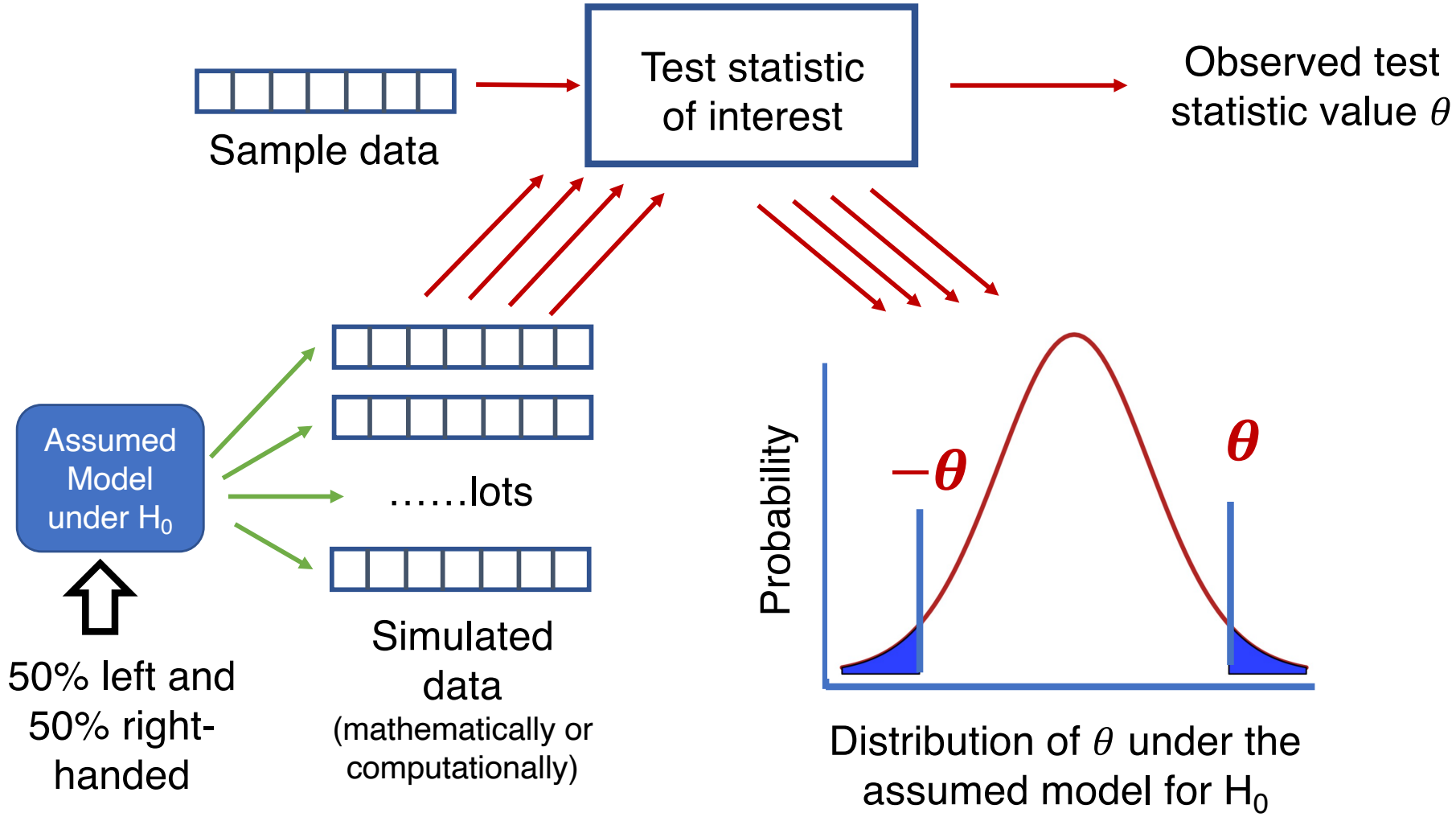


Pr[14 or more right-handed toads] = Pr[14] + P[15] + P[16] + P[17] + P[18] = 0.0155 x 2 (symmetric distribution) = **0.031**

OR: Pr[14 or more right-handed toads] + Pr[4 or less right-handed toads] = 0.031

OR: Pr[14 or more left-handed toads] + Pr[14 or less right-handed toads] = 0.031

The “machinery” behind the framework of the frequentist statistical hypothesis testing

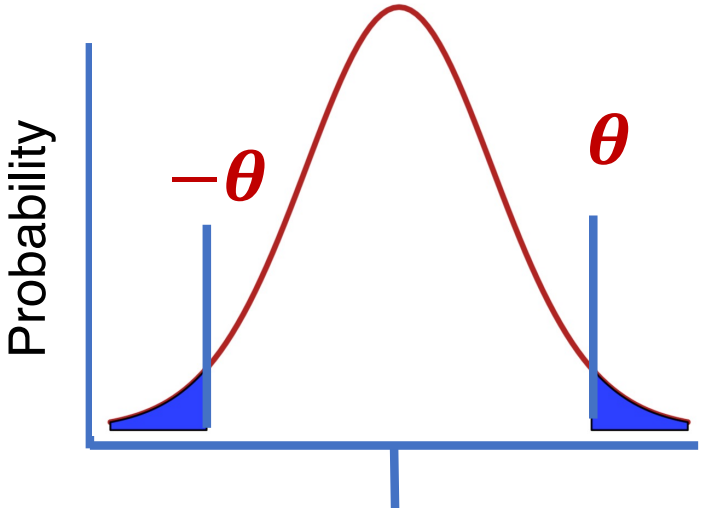


θ = observed number of right-handed toads in the sample

$-\theta$ = observed number of left-handed toads in the sample

Adapted from https://moderndive.com/

The “machinery” behind the framework of the frequentist statistical hypothesis testing



Distribution of θ under the assumed model for H_0

50% left and 50% right-handed

This curve represents the **sampling distribution of number of right-handed toads under H_0.**

The blue shaded areas in the tails represent samples **at least as extreme as the observed θ (and and -θ) under H_0.**

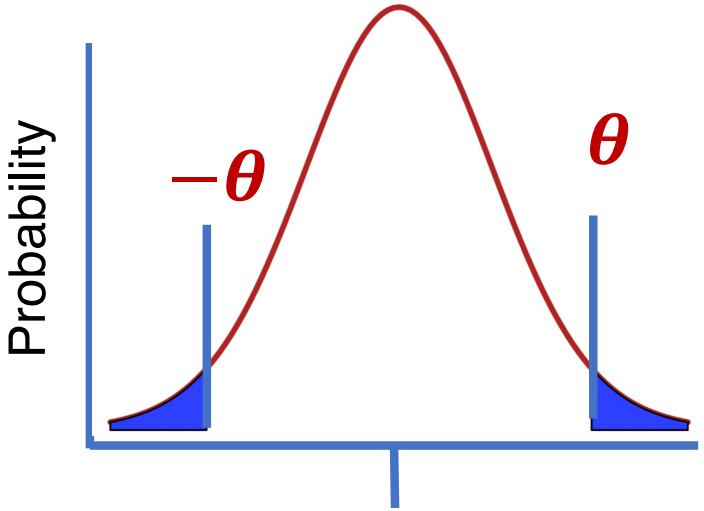
The **p-value is the total probability of those shaded tail regions, which** assumed H_0 as true.

number of right-handed toads equal or larger than the observed

θ = observed number of right-handed toads in the sample

$-\theta$ = observed number of left-handed toads in the sample

The “machinery” behind the framework of the frequentist statistical hypothesis testing



Distribution of θ under the assumed model for H_0

50% left and 50% right-handed

This curve represents the sampling distribution of θ under H_0.

The blue shaded areas in the tails represent outcomes at least as extreme as the observed θ.

The p-value is the total probability of those shaded tail regions, assuming H_0 is true.

SO: The p-value is the probability, calculated under the assumed null hypothesis (H_0), of observing a value of the test statistic (θ) as extreme as, or more extreme than, the one actually observed.

θ = number of right-handed toads equal or larger than the observed

$-\theta$ = number of left-handed toads smaller or larger than the observed

Let’s take a break - 1 minute

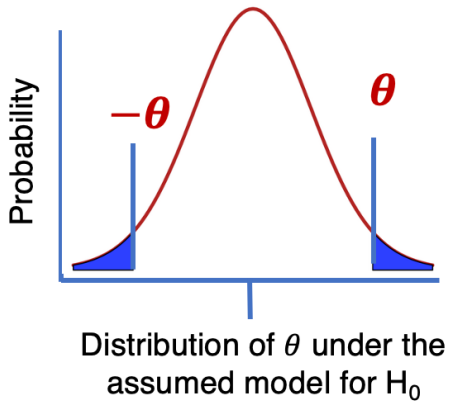
The Frequentist Hypothesis-Testing Framework

Statistical hypothesis testing is a **quantitative inference framework.**

It evaluates how **compatible the data are with an assumed model.**

That model is the **null hypothesis (H_0).**

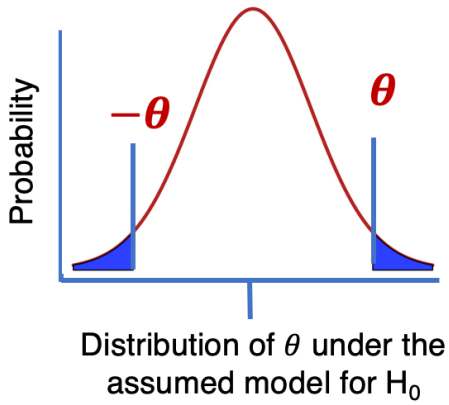
CORE IDEA: we evaluate how surprising the observed data would be if the null hypothesis (H_0) were true.



What's the p-value?

The **p-value** is the probability of observing a result **as extreme or more extreme** than the one observed in the sample.

This probability is calculated **assuming the null hypothesis is true.**



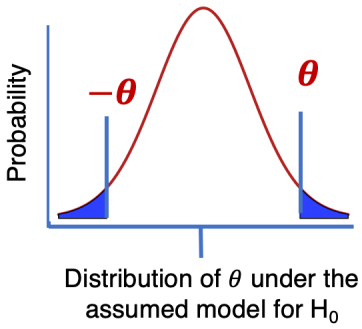
Where the Frequentist Logic Comes From

Imagine many repeated samples drawn from the same theoretical population under H_0 (sampling distribution under H_0.

The sampling distribution describes the long-run behavior of a test statistic.

The observed result is evaluated relative to this distribution.

Frequentist = long-run frequency across repeated samples under H_0.

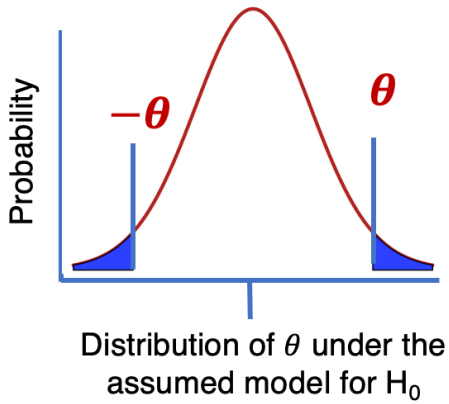


How p-values Generate Evidence

A small p-value means the observed result is rare under the null hypothesis (H_0), and therefore surprising, if H_0 were true.

Rare outcomes (greater surprise, small p-values) indicate incompatibility between data and H_0.

This incompatibility is interpreted as evidence against H_0.



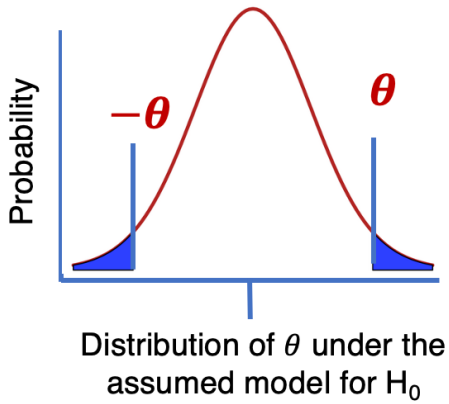
An Important (and Confusing) Point

Statistical tests provide evidence against the null hypothesis.

They do not provide evidence for the null hypothesis.

They also do not provide direct evidence for the statistical alternative.

True but confusing (in the beginning): frequentist tests generates evidence against the implausible (given the data), not for the plausible.



An Important (and Confusing) Point

True but confusing (in the beginning): frequentist tests generates evidence against the *implausible* (given the data), not for the *plausible*.

Implausible (given the data): The **null hypothesis** that *toads are equally likely to be left- or right-handed* (50% / 50%).

Observing **14 right-handed out of 18** would be very unlikely if H_0 were true, so this explanation becomes implausible.

Plausible (but not proven): Any alternative explanation in which right-handedness is more common than left-handedness (e.g., 60%, 70%, 80%, etc.).

These remain plausible because the data are consistent with them—but we cannot say which one is correct.

Why Statistical Tests Give Evidence *Against (implausibility) and* Not *For (plausibility*) the Null

Statistical tests (via their p-values) measure how surprising the observed data are under that assumption (i.e., detect inconsistency).

High surprise (small p-value) → evidence against H_0

Low surprise (large p-value) → no evidence against H_0

Because both a false H_0 and a true H_0 can produce low surprise (hopefully true H_0 less likely to produce low surprise), H_0 can never be confirmed.

Frequentist tests provide evidence against explanations that are implausible given the data, but not for those that remain plausible.

Why the Null Distribution Includes “Extreme” Values

The null distribution shows all outcomes that could occur by chance when H_0 is assumed true.

It is built before seeing the data, not tailored to what is plausible biologically.

Some outcomes are common, others are rare, but all are possible under random sampling (including the ones under the alternative hypothesis, though these will be less frequent).

Extreme outcomes matter because observing them would be very unlikely if H_0 were true.

The null distribution represents the range of outcomes that random sampling can generate under the null hypothesis. Some values (under H_0) are more common than others.

Interpreting the Result (P = 0.031) for the toad study

A p-value of 0.031 indicates high surprise under H_0.

We therefore reject the statistical null hypothesis.

This means the 50/50 assumption is unlikely.

Statistical vs Research Hypotheses

Rejecting H_0 is not the same as proving the alternative, and not rejecting H_0 is not the same as proving the null.

But rejecting H_0 can support the research (biological) hypothesis.

Statistical hypotheses are tools; research hypotheses are the goal.

Statistical tests can rule things out, but they do not prove hypotheses true.

Key Take-Home Messages

Frequentist tests assess compatibility with H_0.

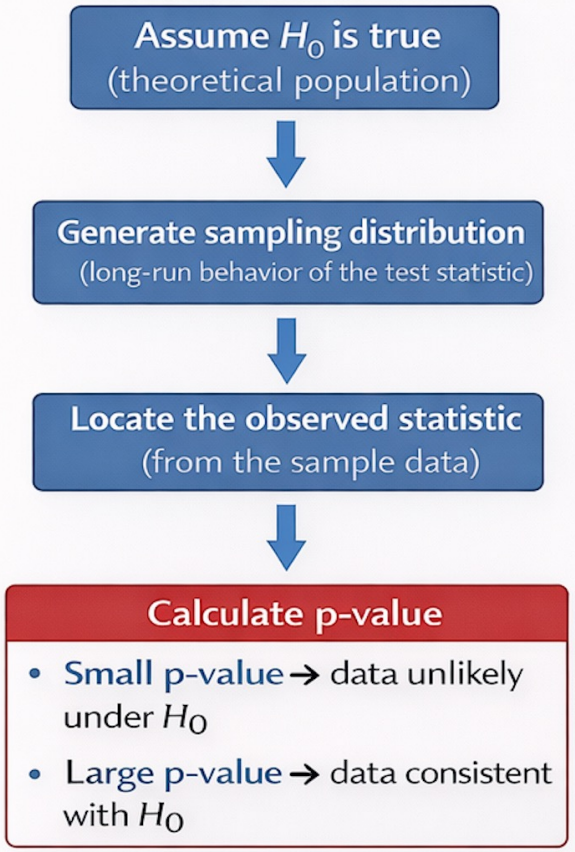
p-values quantify surprise under an assumption.

We rule out what is unlikely, not confirm what is true.

Biological conclusions are indirect, but evidence-based.

Frequency hypothesis testing - summary

Core Idea: We evaluate how surprising the observed data would be if the null hypothesis were true.



- Assume H_0 is true (theoretical population)
- **Generate sampling distribution**
  (long-run behavior of the test tstatci stistic)
- **Locate the observed statistic** (from the sample data)
- **Calculate** p-value = probability of results this extreme or more extreme under H_0

Decision

- Small p-value → data unlikely under H_0 → reject H_0
- Large p-value → data consistent with H_0 → fail to reject H_0

Key take-home messages:

- ✓ Tests provide evidence against the null hypothesis
- ✓ They do not provide evidence for the null or the statistical alternative
- ✓ Evidence against H_0 can support the research (biological) hypothesis

Let’s take a break - 1 minute

Estimation versus Statistical hypothesis testing

Do the conclusions from the two statements below differ?

How?

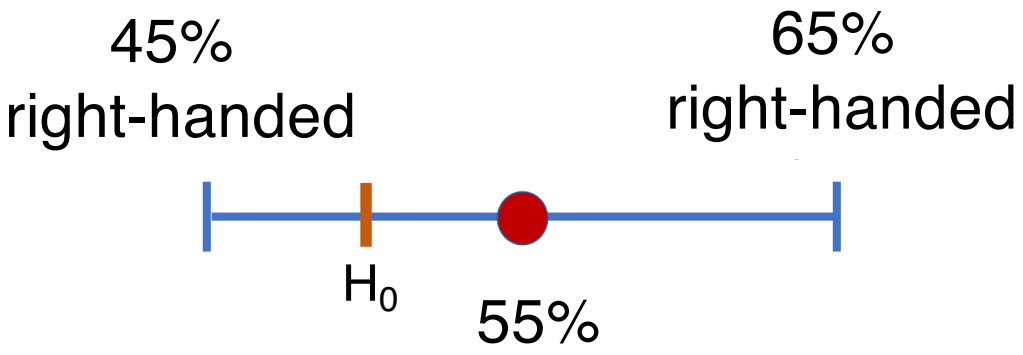
Which one you prefer?

We are confident that the true proportion (right/left) is 0.75 ± 0.03 (i.e., between 0.72 and 0.78)

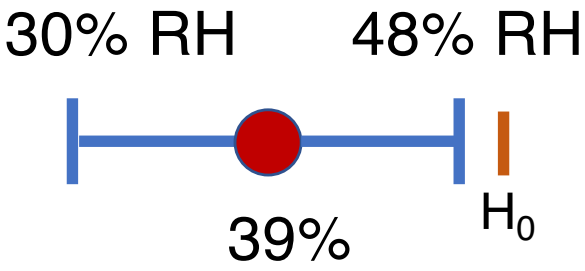
OR

The proportion is different from 0.5

Estimation vs. Hypothesis Testing



Don't reject H_0; $p > 0.05$



Reject H_0; $p < 0.05$

"Quantitative" statement

"Qualitative" Statement (reject / don't reject; yes/no)

A confidence interval quantifies the uncertainty around an estimate by giving a range of plausible parameter values.

Estimation vs. Hypothesis Testing

Both use sample data to make inferences about a population

Estimation:

- Focuses on how large a population parameter might be
- Provides a range of plausible values
- Asks: What is the value of the parameter?
- Example: How many toads are right-handed?

Hypothesis Testing:

- Focuses on whether a specific assumption is unlikely
- Compares the parameter to a null expectation (H_0)
- Asks a yes/no question
- Example: Are right- and left-handed toads equally frequent?

Estimation vs. Hypothesis Testing

Both use sample data to make inferences about a population

Estimation:

- Focuses on how large a population parameter might be
- Provides a range of plausible values
- Asks: What is the value of the parameter?
- Example: How many toads are right-handed?

Hypothesis Testing:

- Focuses on whether a specific assumption is unlikely
- Compares the parameter to a null expectation (H_0)
- Asks a yes/no question
- Example: Are right- and left-handed toads equally frequent?

Estimation asks “how much”; hypothesis testing asks “unlikely or not?”

When Estimation Is Uncertain

With small samples, estimates can be highly uncertain
e.g., wide confidence intervals with only 18 toads.

In the toad study, we cannot confidently estimate a narrow range
e.g., stating the true proportion is between 75%–79% would likely require a much larger sample.

Precise estimation is costly.

What Hypothesis Testing Can Still Do

Instead, we can test whether the data are consistent with a specific assumption.

Assume a 50% / 50% null hypothesis (no handedness); *Ask*: How unusual is observing 14 right-handed toads out of 18 under this assumption?

A surprising result suggests that 50% is unlikely.

This provides evidence for handedness, even if the exact proportion is uncertain (philosophical and probabilistic nuances - more later).

When we cannot estimate precisely, we can still test improbability.

Why We Like Yes/No Answers

Precise estimation often requires large sample sizes.

Hypothesis testing is widely used because it provides simple yes/no answers.

Humans (and decision-makers) tend to prefer clear decisions over ranges of uncertainty.

Testing answers: Is this assumption unlikely or not?