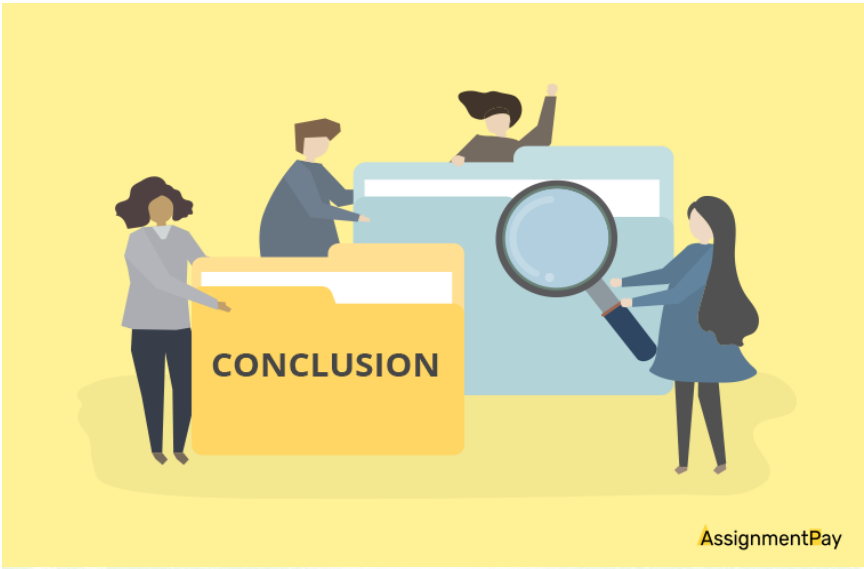
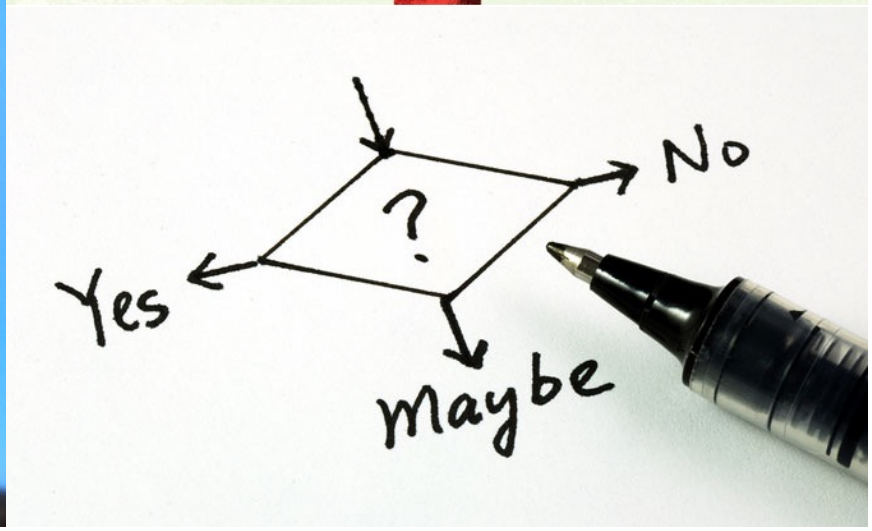
Generating evidence-based conclusions through statistical hypothesis testing, while recognizing that biological knowledge is inherently uncertain because it is inferred from samples rather than entire populations.





What is meant by evidence from the scientific literature

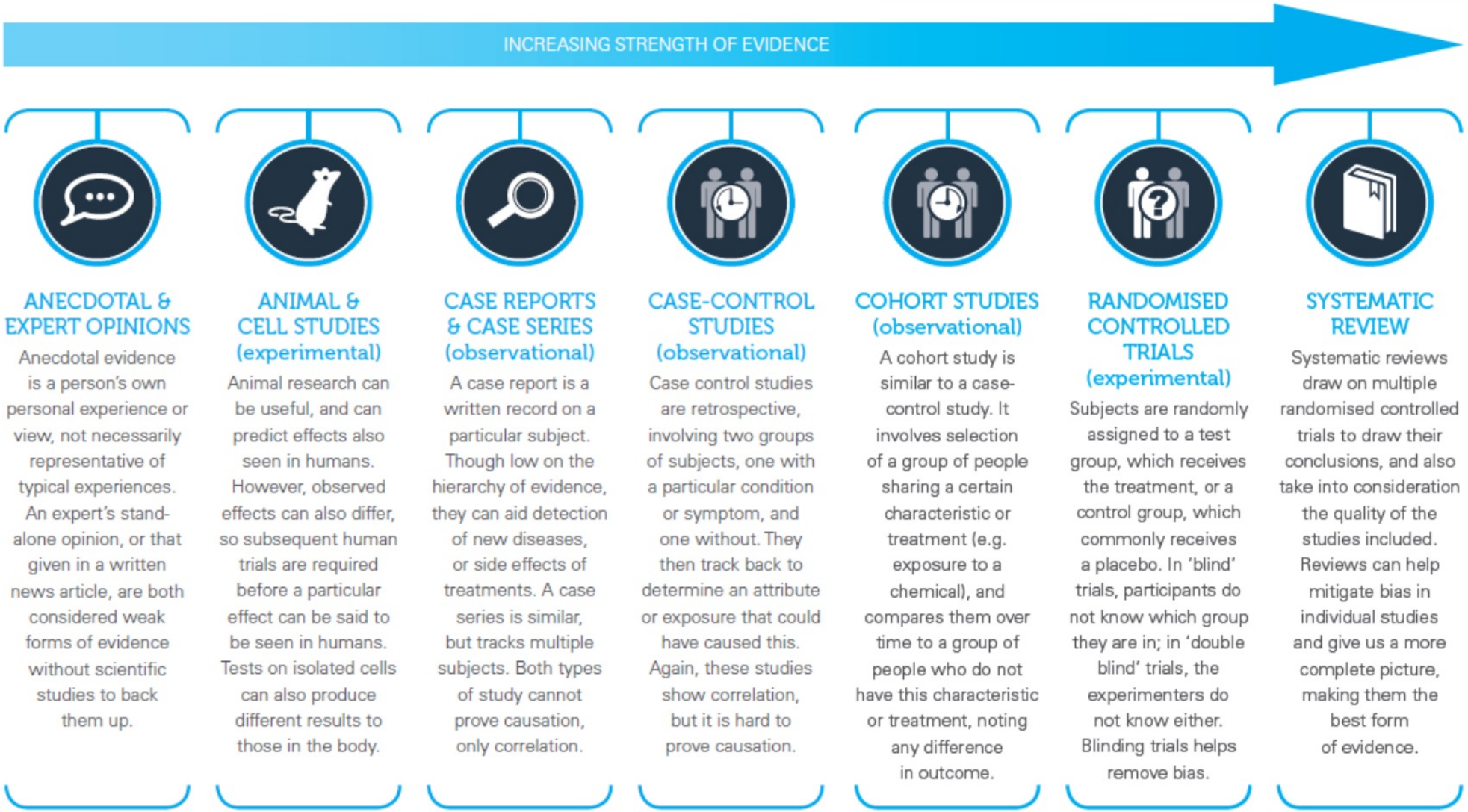
Evidence refers to information, facts, or data that support or challenge a claim, prediction, assumption, or hypothesis.

When people refer to “evidence from the scientific literature,” they typically mean empirical findings reported in peer-reviewed scholarly journals.

Government reports, technical assessments, and publications from non-governmental organizations can also provide relevant and rigorously collected information, particularly when they are grounded in systematic data collection and transparent methodologies.

(e.g., the Intergovernmental Panel on Climate Change [IPCC] from the United Nations).

Types of scientific evidence



Source: https://www.sciencemediacentre.co.nz/coveringscience/types-of-scientific-evidence/

The role of probability in evidence

Quantifying Statistical Evidence from Data (i.e., from samples)

Suppose you flip a coin 20 times and observe 18 heads (and 2 tails).

Would you find that surprising?

Is this outcome unusual enough to make you question whether the coin is truly fair?

Are these data consistent with the assumption of a fair coin?

If the coin were truly fair, would these data (18/20 heads) be considered a typical outcome?

Quantifying Statistical Evidence from Data (i.e., from samples)

Suppose you flip a coin 20 times and observe 18 heads.

The probability of observing 18 heads out of 20 flips purely by chance (if the coin were fair) is **0.0004025.**

This probability means that if you repeated the experiment of flipping a fair coin 20 times over and over again (say 1,000,000 times or infinite times) you would expect to see 18 heads (or more extreme outcomes) only **0.0004025 × 1,000,000 ≈ 403 times out of one million repetitions.**

So, although it is possible to obtain 18 heads with a fair coin, it would happen very rarely in repeated sampling from a fair coin.

0.0004025 represents the ***statistical evidence*** against the assumption of fairness. It tells us how surprising the observed data (18 heads in 20 flips) would be if the coin truly had a 50% chance of heads.

Would you consider that outcome reasonably compatible with a fair coin, or so unlikely that it would lead you to question the assumption of fairness?

Quantifying Statistical Evidence from Data (i.e., from samples)

The probability of observing 18 heads out of 20 flips purely by chance (if the coin were fair) is **0.0004025.**

In statistics, we refer to this probability as the value predicted by the model that assumes the coin is fair.

Quantifying Statistical Evidence from Data (i.e., from samples)

The probability of observing 18 heads out of 20 flips purely by chance (if the coin were fair) is **0.0004025.**

In statistics, we refer to this probability as the value predicted by the model that assumes the coin is fair.

This model is called the null hypothesis (H_0) because it makes explicit that a specific mechanism must be assumed (e.g., a fair coin) in order to calculate probabilities.

That assumption defines a probability distribution for the data, and we then evaluate whether the observed outcome appears *typical* or *atypical* under that distribution.

Quantifying Statistical Evidence from Data (i.e., from samples)

Which one is more surprising than the other?

If the coin were truly fair, which of these data be considered a typical outcome?

Number of flips	Number of heads	Probability under the assumption of a fair coin
20	12	0.5034
20	18	0.00040
20	15	0.04139
200	110	0.179
200	120	0.00569
200	150	< 0.0000000001

If the coin were truly fair, which of these data be considered a typical outcome? In statistics we refer to the probability predicted by a model assuming a fair coin.

Quantifying Statistical Evidence from Data (i.e., from samples)

What do we see overall? As the sample size increases:

The same proportional difference becomes far more unlikely. The ***evidence against the fair-coin assumption*** becomes stronger for large samples.

In short: with larger samples, even modest departures from 50% can become highly inconsistent with the fair-coin model.

This illustrates a key principle of statistical inference: **sample size strongly influences how unusual an observed result appears under a given assumption (model).**

Number of flips	Number of heads	Probability under the assumption of a fair coin
20	12	0.5034
20	18	0.00040
20	15	0.04139
200	110	0.179
200	120	0.00569
200	150	< 0.0000000001

Quantifying Statistical Evidence from Data (i.e., from samples)

What do we see overall? As the sample size increases:

The same proportional difference becomes far more unlikely. The ***evidence against the fair-coin assumption*** becomes stronger for large samples.

In short: with larger samples, even modest departures from 50% can become highly inconsistent with the fair-coin model.

This illustrates a key principle of statistical inference: **sample size strongly influences how unusual an observed result appears under a given assumption (model).**

Number of flips	Number of heads	Probability under the assumption of a fair coin
20	12	0.5034
20	18	0.00040
20	15	0.04139
200	110	0.179
200	120	0.00569
200	150	< 0.0000000001

Quantifying Statistical Evidence from Data (i.e., from samples)

The probability of observing 18 heads out of 20 flips purely by chance (if the coin were fair) is **0.0004025.**

AGAIN: In statistics, we refer to this probability as the value predicted by the model that assumes the coin is fair.

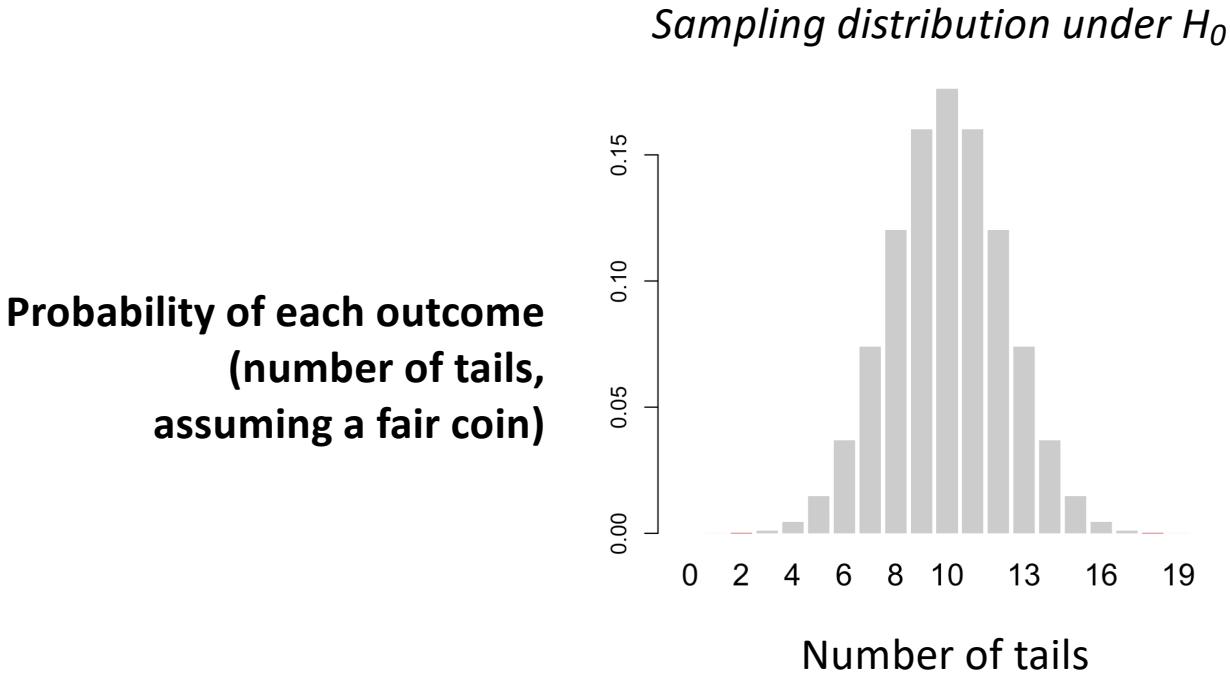
This model is called the null hypothesis (H_0) because it makes explicit that a specific mechanism must be assumed (e.g., a fair coin) in order to calculate probabilities.

Quantifying Statistical Evidence from Data (i.e., from samples)

We consider results that are equally or more extreme because, once we determine that 18 heads is unusual under a fair-coin model, any outcome that departs even further from what fairness predicts provides at least as much statistical evidence against that model.

In other words, if 18 heads is already strong evidence that the fairness assumption may not hold, then 19 or 20 heads would represent even stronger evidence against the fair-coin model, and 2, 1 or 0 heads do the same in the opposite direction.

When assessing how surprising a result is, we therefore include all outcomes that contradict the model at least as much as the one we observed.1



Quantifying Statistical Evidence from Data (i.e., from samples)

STRONG statistical evidence is generated when the observed data would be very unlikely under the assumption being tested (model /H_0), indicating that the assumption is inconsistent with what was observed.

WEAK statistical evidence is generated when the observed data are somewhat unlikely under the assumption being tested (model / H_0), suggesting limited inconsistency with what was expected under that assumption.

NO statistical evidence against the assumption is generated when the observed data are reasonably likely under the assumption being tested (model /H_0), indicating that the data are consistent with what was expected.

Quantifying Statistical Evidence from Data (i.e., from samples)

STRONG statistical evidence is generated when the observed data would be very unlikely under the assumption being tested, indicating that the assumption is inconsistent with what was observed.

WEAK statistical evidence is generated when the observed data are somewhat unlikely under the assumption being tested, suggesting limited inconsistency with what was expected under that assumption.

NO statistical evidence against the assumption is generated when the observed data are reasonably likely under the assumption being tested, indicating that the data are consistent with what was expected.

Statistical evidence	Number of flips	Number of heads	Probability under the assumption of a fair coin
No	20	12	0.5034
Strong	20	18	0.00040
Moderate	20	15	0.04139
Weak to none	200	110	0.179
Strong	200	120	0.00569
Very strong	200	150	< 0.0000000001

Statistical hypothesis testing provides a quantitative framework for generating evidence in support of or against a biological phenomenon.

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

They randomly sampled 18 wild toads, placed a balloon over each one's head, and recorded which forelimb the toads used to remove it to determine their preferred limb.

What is a research hypothesis?!

A hypothesis is a supposition or proposed explanation made based on limited evidence as a starting point for further investigation (Oxford dictionary); e.g.,

"animals, other than humans, also have a preferred limb (handedness)".

Hypotheses cannot be definitively proven true or false based on a single dataset. They can only be described as **supported or not supported** by the data at hand, and they always remain open to revision or refutation in light of future evidence.

What is a research hypothesis?!

A hypothesis is a supposition or proposed explanation made based on limited evidence as a starting point for further investigation (Oxford dictionary); e.g.,

“animals, other than humans, also have a preferred limb (handedness)”.

Hypotheses cannot be definitively proven true or false based on a single dataset. They can only be described as **supported or not supported** by the data at hand, and they always remain open to revision or refutation in light of future evidence.

Inference is based on limited sample data rather than the entire population. As such, conclusions are always conditional on the evidence observed and therefore cannot establish absolute truth.

Strong statistical evidence is generated when strong statistical evidence is generated when the observed data would be very unlikely under the assumption being tested, indicating that the assumption is inconsistent with what was observed.

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

Back to statistically testing the hypothesis of handedness

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

They randomly sampled 18 wild toads, placed a balloon over each one's head, and recorded which forelimb the toads used to remove it to determine their preferred limb.

Translating the research question into a statistical question:

Do right-handed and left-handed toads occur with equal frequency in the (population, or is one more common than the other?

RESULTS: 14 toads were right-handed and four were left-handed. Do these results provide sufficient evidence to demonstrate handedness in toads?

Addressing research hypotheses within the framework of statistical hypothesis testing.

The **statistical hypothesis framework** (most often involving statistical testing) is a quantitative method of statistical inference that allows to generate evidence for or against a **hypothesis**.

In the frequentist framework of inference, we typically compute a probability value (p-value) that quantifies how compatible the observed data are with a specified null hypothesis (H_0), thereby providing a measure of evidence against that hypothesis (e.g., testing for handedness in toads).

Two possible statistical hypotheses:

The research hypothesis is translated into a statistical question. In the frequentist framework, the statistical question is then stated as two mutually exclusive hypotheses called null hypothesis (H_0) and alternative hypothesis (H_A).

A null hypothesis, which represents a specific assumption about the population parameter (often reflecting no effect or no difference), and

An alternative hypothesis, which represents a competing claim suggesting that the parameter differs from that assumption.

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 null hypothesis is a model of “nothing systematic going on”

It is called “null” because it represents a baseline or default assumption; typically, that there is no effect, no difference, or no deviation from some reference value.

It assumes that any observed differences or patterns arise purely from random variation rather than from a real underlying effect.

In many classical tests, the null hypothesis states that nothing systematic is happening beyond random variation. For example, in the coin case, the null says the coin is fair; in a mean comparison, the null says the difference is zero.

Historically, the term “null” emphasizes that the hypothesis often specifies the absence of a phenomenon.

The intuition behind the frequentist framework of statistical hypothesis testing

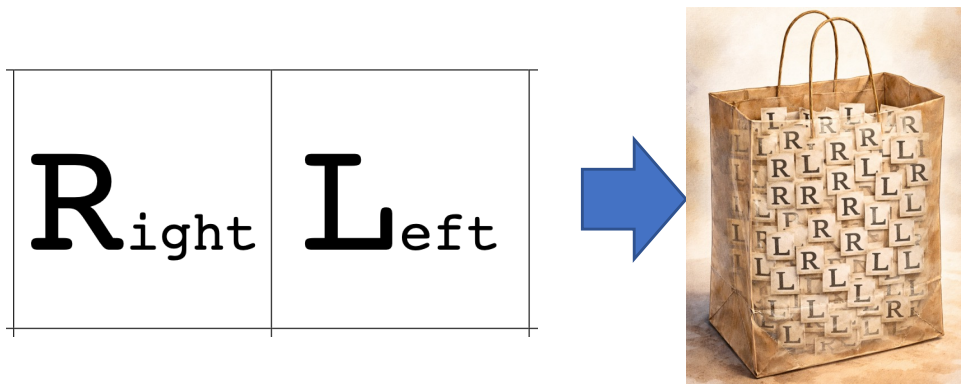
Let's use a simple computational thought experiment involving paper and a bag.

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): 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

⇧

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] "L" "L" "R" "R" "R" "R" "R" "R" "L" "L" "L" "R" "R" "R" "L" "R" "L" "R"

> sum(Sample1 == "R")
[1] 11
> sum(Sample1 == "L")
[1] 7

> Sample2 <- sample(c("L", "R"), size = 18, prob = c(0.5, 0.5), replace = TRUE)

> Sample2
 [1] "L" "L" "L" "L" "L" "L" "L" "R" "R" "L" "L" "R" "L" "L" "R" "L" "L" "L"

> sum(Sample2 == "R")
[1] 4
> sum(Sample2 == "L")
[1] 14
```

Sample 1

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 distribution of data (samples) would look like in that hypothetical world. "*Under*" is the **correct phrasing** in *frequentist* hypothesis testing.

"For H_0" sounds like we are *arguing in favour 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 1,000,000 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 8 right-handed toads and 10 left-handed toads? 0.1669 or 16.69%

If we had drawn 1,000,000 samples from the population assumed under H_0, 166,900 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

Quantifying Statistical Evidence from Data (i.e., from samples)

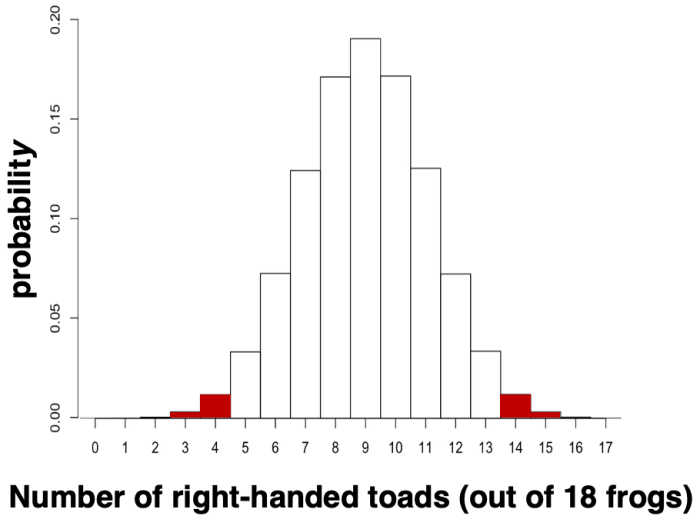
We consider results that are equally or more extreme because, under a model assuming 50% right-handed and 50% left-handed, outcomes that deviate further from this expected balance are at least as incompatible with the model as the observed result.

Therefore, when evaluating surprise under H_0, we include all outcomes that depart from the model at least as much as the one observed. In this case, values such as 0, 1, 2, 3, 4 & 15, 16, 17, or 18 right-handed depart even more strongly from the 50%/50% ratio.

When evaluating how surprising the observed result is, we therefore include all outcomes that contradict the model at least as much as the one we observed.

Probability of each outcome (assuming a population with 50% right- & 50% left-handed)

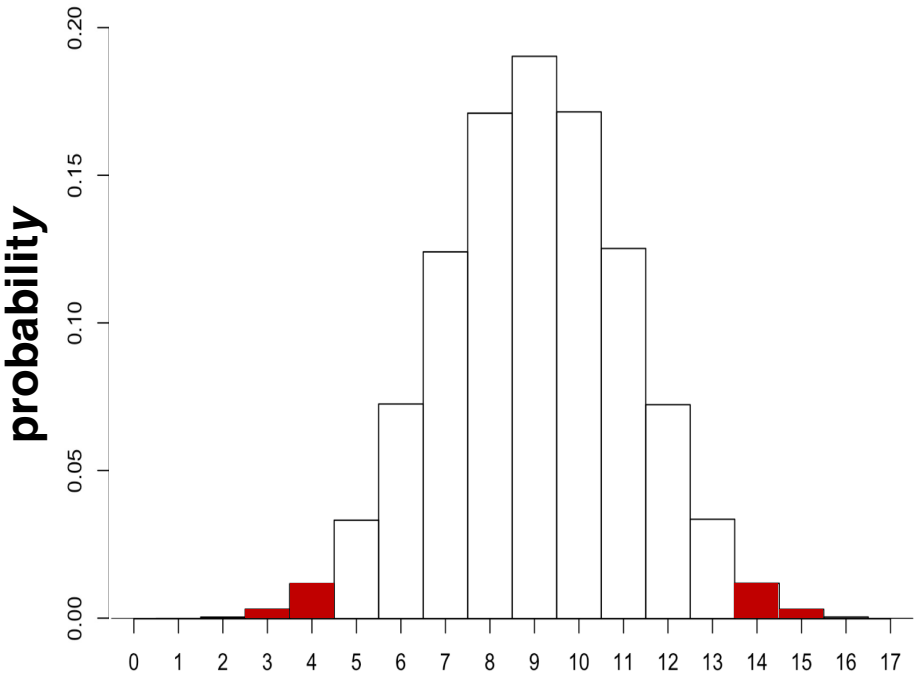
Sampling distribution under H_0



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



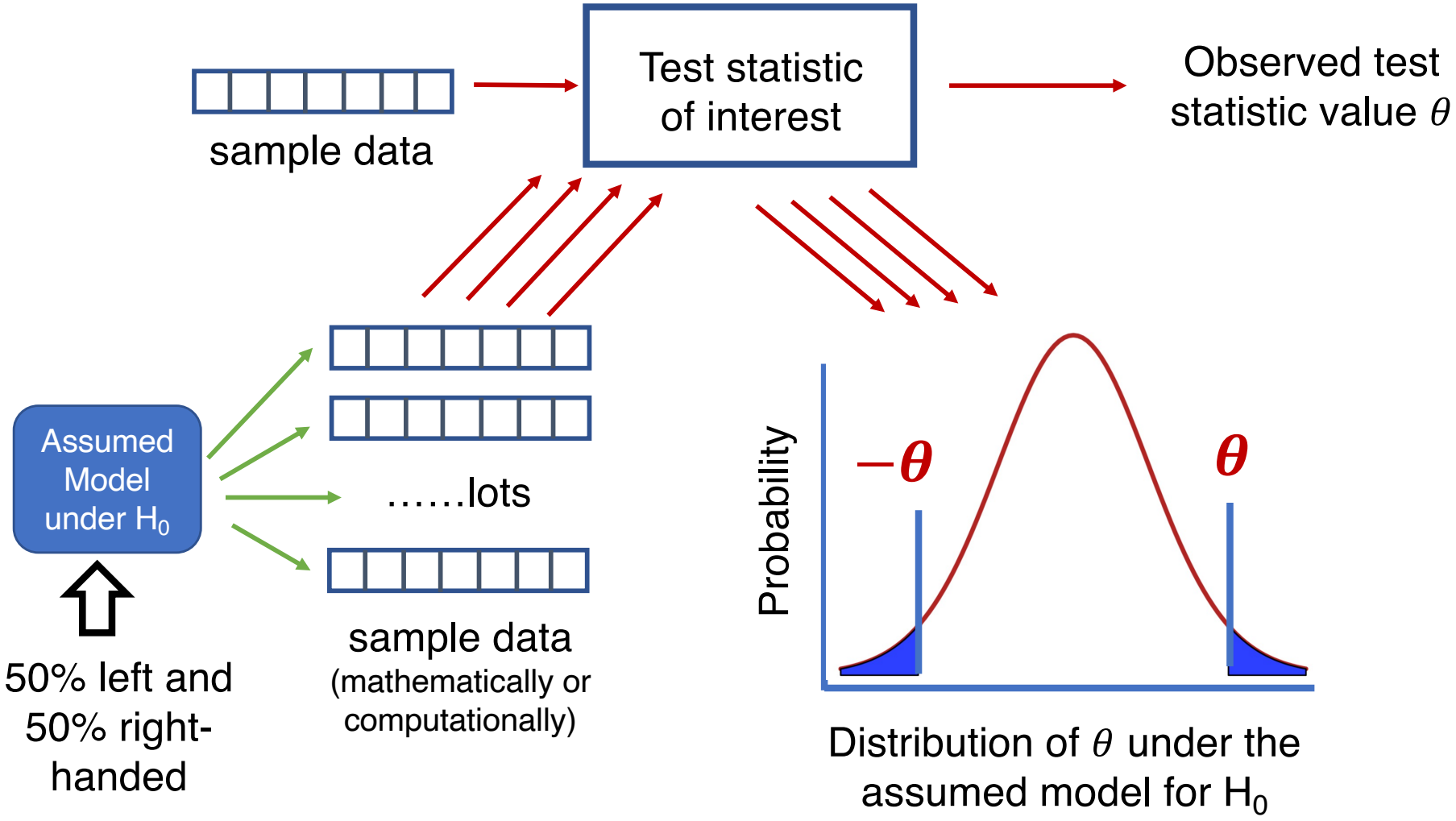
P-value =

P[0] + P[1] + P[2] + P[3] + P[4] +

P[14] + P[15] + P[16] + P[17] + P[18] =

0.0155 + 0.0155 = **0.031.**

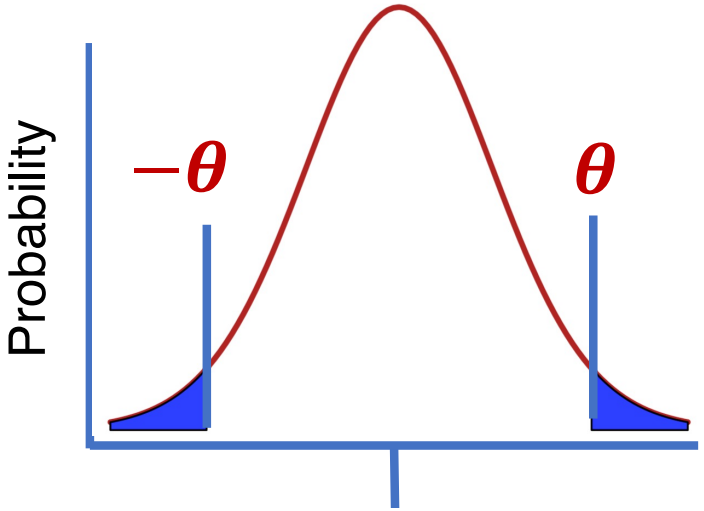
The “machinery” behind the framework of the frequentist statistical hypothesis testing



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

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

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.**

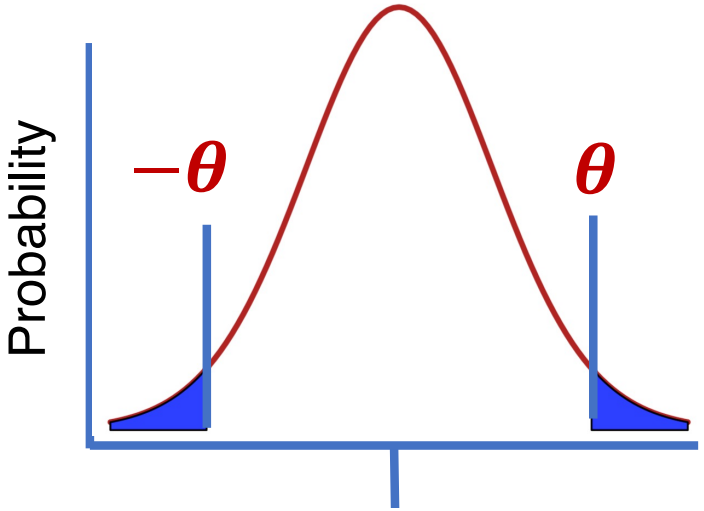
The blue shaded areas in the tails represent samples **at least as extreme as the observed θ (and and -θ) under H_0.**

The **p-value is the total probability of those shaded tail regions calculated by assuming** H_0 as true.

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

The **p-value is the total probability of those shaded tail regions calculated by assuming** H_0 as 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

Quantifying Statistical Evidence from Data (i.e., from samples)

REMEMBER: We consider results that are equally or more extreme because, once we determine that 14 right-handed is unusual under a fair-coin model, any outcome that departs even further from what fairness predicts provides at least as much statistical evidence against that model.

In other words, if 14 right-handed is already strong evidence that the handedness (i.e., 50%/50% - akin to the fair coin) assumption may not hold, then 15, 16, 17 or 18 right-handed would represent even stronger evidence against handedness, and 4, 3, 2, 1 or 0 right-handed do the same in the opposite direction.

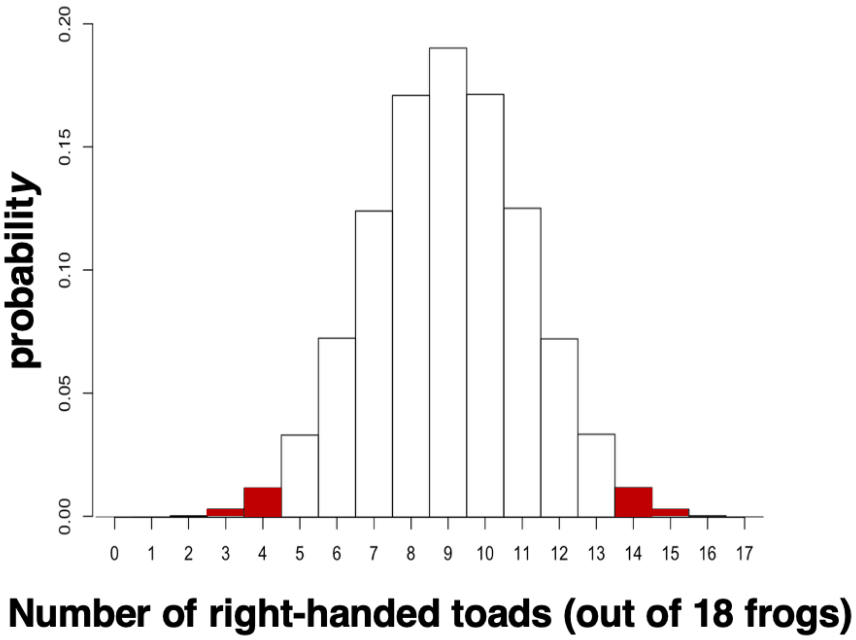
When assessing how surprising a result is, we therefore include all outcomes that contradict the model at least as much as the one we observed.

P-value =

P[0] + P[1] + P[2] + P[3] + P[4] +

P[14] + P[15] + P[16] + P[17] + P[18] =

0.0155 + 0.0155 = **0.031.**



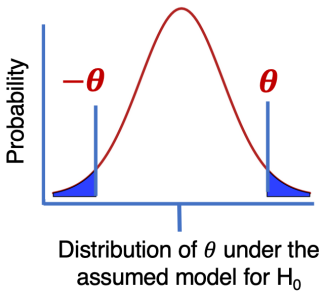
The Frequentist Hypothesis-Testing Framework

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.



An Important (and Confusing) Point

In frequentist inference, we do not prove that a model is correct; we evaluate how incompatible the observed data are with the model (H_0: null hypothesis).

Strong evidence arises when the data would be very unlikely if the model were true.

We test the null hypothesis (or null model) directly. If the observed data are highly incompatible with it, we reject the null and regard the alternative hypothesis as more plausible in light of the evidence.

An Important (and Confusing) Point

In frequentist inference, we do not prove that a model is correct; we evaluate how incompatible the observed data are with the model (H_0: null hypothesis).

Strong evidence arises when the data would be very unlikely if the model were true.

We test the null hypothesis (or null model) directly. If the observed data are highly incompatible with it, we reject the null and regard the alternative hypothesis as more plausible in light of the evidence.

However, this does not mean the alternative hypothesis has been proven true; it simply means the null model does not adequately explain the data.

The p-value quantifies that consistency: small values indicate greater inconsistency with H_0, whereas large values indicate that the observed data are reasonably consistent with H_0 (though they do not prove it is true).

P-values Measure How Incompatible the Data Are with the Null Hypothesis (Null Model)

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

If the probability of observing 18 or more heads in 20 flips is 0.0004025, then in one million repetitions of that experiment we would expect to see such extreme outcomes only about 403 times.

So, observing 18 heads in a single trial is witnessing an outcome that occurs only very rarely under the fair-coin mode; on the order of a few hundred times in a million similar experiments.

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

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

It indicates that the observed data would occur about 3.1% of the time if the null model were true.

That is relatively uncommon, and under conventional thresholds (like 0.05), it is considered sufficiently incompatible with the null to justify rejecting it.

While this provides evidence against H_0, the p-value is not extremely small, so the evidence can be considered moderate rather than overwhelming.

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.

What the p-value Does NOT Tell Us

The p-value tells us how surprising the data are if H_0 were true.

It does not tell us the probability that H_0 is true.

It does not tell us the probability that we are making a mistake.

It does not tell us the probability that we should reject H_0.

Deciding on when to reject the null hypothesis

The significance level, denoted by α (alpha), is the threshold we set before analyzing the data to decide how much incompatibility with the null model we are willing to tolerate before rejecting it.

It represents the probability of rejecting the null hypothesis when it is actually true (more on this later).

α is not a law of nature or probabilities. It is a chosen threshold that reflects how cautious we want to be about claiming evidence against the null model.

Biology usually uses 0.05 or 0.01 – tradition and consistency.

Key Take-Home Messages about statistical hypothesis testing

- Frequentist tests assess compatibility with H_0.

- p-values quantify surprise under an assumption (i.e., H_0).

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

- Biological conclusions are indirect, but evidence-based.

Decision in statistical hypothesis testing: In light of the statistical evidence (P-value), should we favour H_0 or H_A?

Mark Chang (2017) well stated: "A smaller p-value indicates a discrepancy between the hypothesis and the observed data. In this sense, p-value measures the strength of evidence against the null hypothesis".

CRITICAL: the p-value does not represent the probability that the null hypothesis (H_0) is true. Instead, it is a quantitative measure indicating the strength of evidence against H_0. A smaller p-value suggests stronger evidence against the null hypothesis.

Statistical hypothesis testing involves:

1) How the research hypothesis should be transformed into a statistical question.

2) State the null (parameter for the theoretical population) and alternative hypotheses.

3) Compute the observed value for a particular metric of interest (i.e., based on the sample data, i.e., observed summary statistic). This is called *test statistic*. In our toad example it was simply the number of right-handed individuals.

4) Compute the P-value by contrasting the sample (observed) value against a sampling distribution that assumes the null hypothesis to be true (around the parameter of interest for a theoretical population).

5) Draw a conclusion by contrasting the p-value against the significance level (α). If the p-value is greater than α, then do not reject H_0; if P-value is smaller or equal than α, then reject H_0.

What does the significance level (α level) represent?

There is a lot of disagreement among statisticians and users about whether to 'do not reject' or 'reject' statistical hypotheses based on p-values (i.e., decision based on a threshold).

i.e., Decide whether to use α as a threshold for determining if a p-value is non-significant (fail to reject H_0) or significant (reject H_0 in favor of H_A).

While I agree with these arguments, it seems unlikely that we will see radical changes in research behaviour any time soon.

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

The don’ts about P values and statistical hypothesis testing (Wasserstein et al. 2019)

1. P-values can indicate how incompatible the observed data are with a specified statistical model (e.g., the one assumed under H_0).

2. P-values do not measure the probability that the studied research hypothesis is true.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold (alpha).

4. A p-value, or statistical significance, does not measure the biological importance of a result.

- There are many other important don’ts that we will continue to cover in the course.

The don'ts about P values and hypothesis testing (Wasserstein et al. 2019)

Despite the limitations of p-values, we are not recommending that the calculation and use of p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., p = 0.08) and not yes/no reject the null hypothesis [even though in BIOL322 we will use this tradition because it is the most used and unlikely to change anytime soon].

The biggest push today is to abandon the idea of statistical significance. In other words, to abandon the almost universal and routine practice to state that if the probability is smaller than or equal to alpha, than we should state that the results are significant.

Abandoning the concept of significance is easier said than done. The majority of researchers still report results as either significant or non-significant. In BIOL322, we will guide you towards more nuanced interpretations, but it is challenging to break away from the common practice in statistical applications across biology and most other fields.

Use p-values using “the language of evidence” against H_0

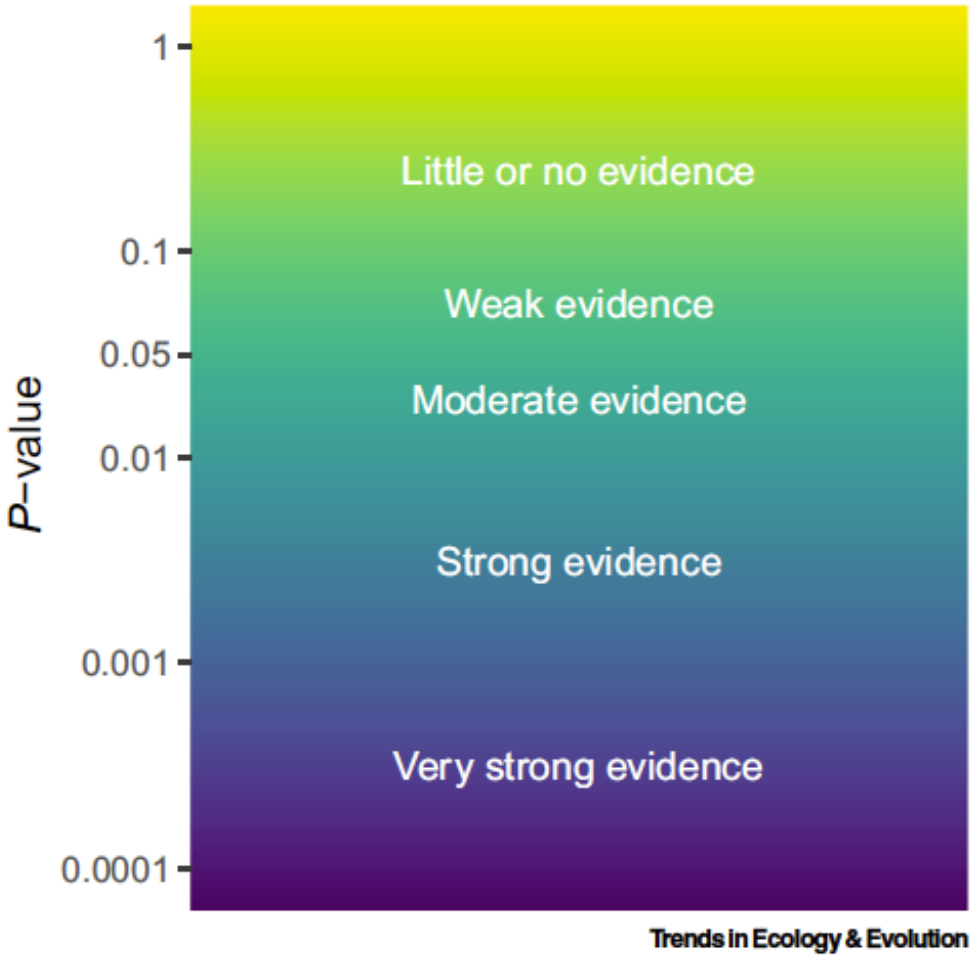


Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

Note: Since the p-value is derived under the assumption of H_0, it provides evidence against H_0, but not in favour of H_A.

This means we can gather evidence to reject H_0 (which assumes one specific parameter), but we cannot confirm H_A, as many potential parameter values could fit H_A (e.g., 55%/45%, 80%/20% right-handed, etc.)

Stefanie Muff et al. 2022. Rewriting results sections in the language of evidence. Trends in Ecology and Evolution 3:203-210.