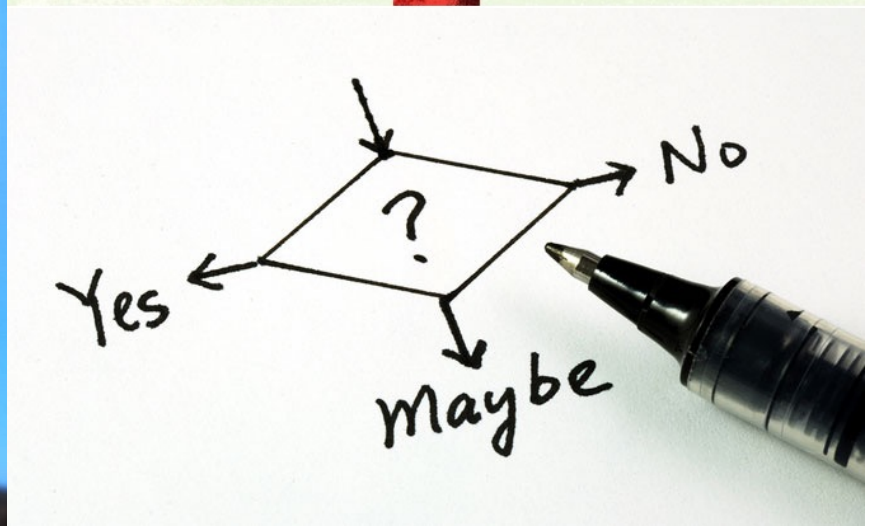
Generating evidence-based conclusions through statistical hypothesis testing, recognizing that biological knowledge is inherently incomplete (i.e., it's based on samples).



What is meant by evidence from the scientific literature

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

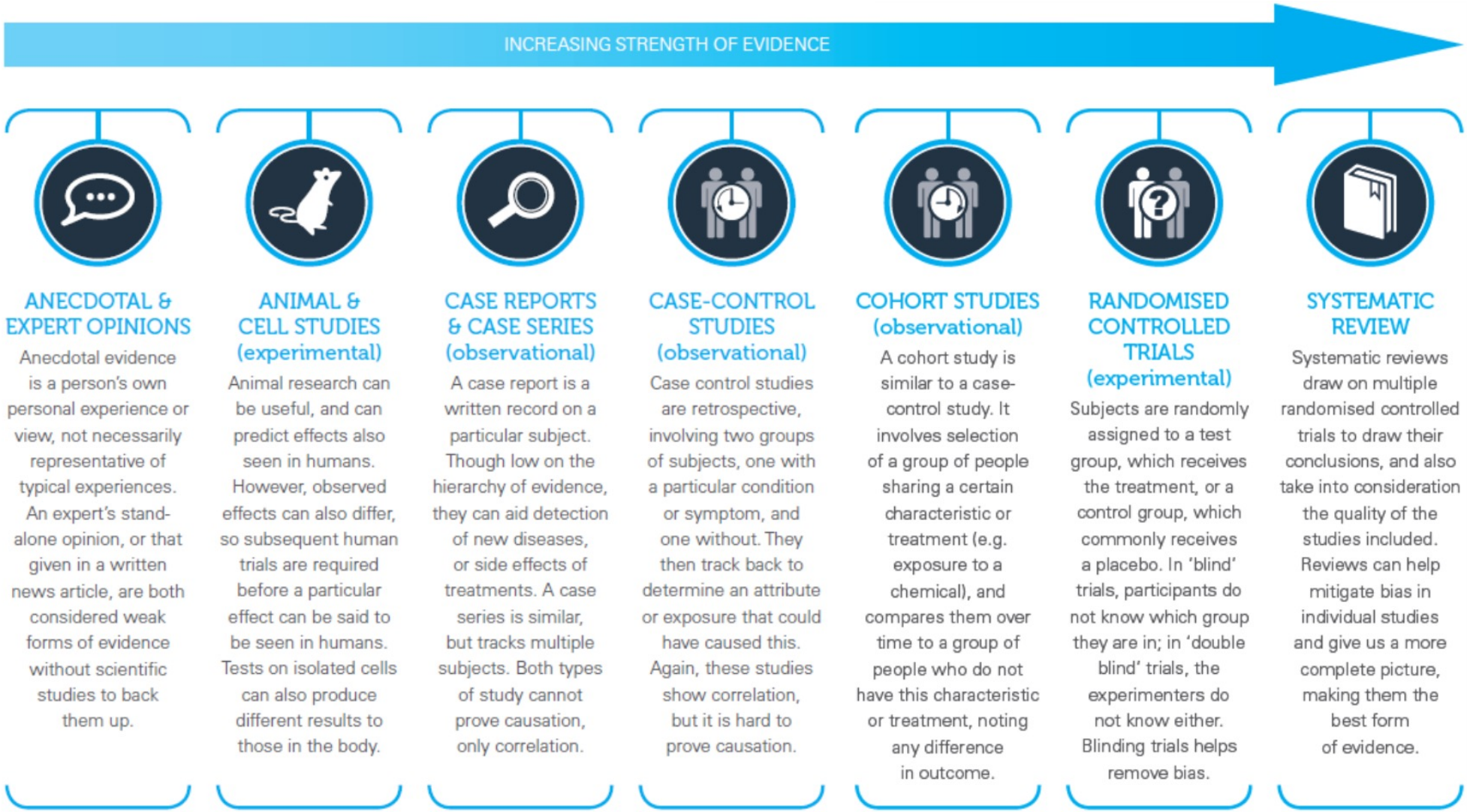
When citing 'evidence from the scientific literature,' people typically mean empirical studies published in peer-reviewed scholarly journals.

However, scientific literature is not the only source of science-based evidence. Government and non-governmental reports also provide relevant information

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

adapted from: https://cebma.org/faq/scientific-evidence/

Types of scientific evidence



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

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

Hypotheses cannot be proven right or wrong from the data. Hypotheses can be said to be either **supported** (or not supported) by the data at hands (and can be potentially **refuted** by future data).

Hypotheses, Theories and Laws: three different components

Research hypotheses cannot be proven right or wrong from the data. Hypotheses can be said to be either supported by the data at hands (and can be potentially refuted by future data).

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):

https://www.forbes.com/sites/marshallshepherd/2019/06/15/theory-hypothesis-and-law-debunking-a-climate-change-contrarian-tactic/#37a3ce047ca7.

Hypotheses, Theories and Laws: three different components

Research hypotheses cannot be proven right or wrong from the data. Hypotheses can be said to be either supported by the data at hands (and can be potentially refuted by future data).

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):

https://www.forbes.com/sites/marshallshepherd/2019/06/15/theory-hypothesis-and-law-debunking-a-climate-change-contrarian-tactic/#37a3ce047ca7.

A scientific theory is a well-substantiated explanation for why something (a natural phenomenon) happens. And a scientific law (gravity) describes what happens (objects fall towards the ground).

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 **research hypothesis**.

The research hypothesis is translated into a statistical question. The statistical question is then stated as two mutually exclusive hypotheses called null hypothesis (H_0) and alternative hypothesis (H_1 or H_A).

The framework most often involves computing a probability value that serves as a quantitative indicator of support for or against the research hypothesis (e.g., generate evidence for or against handedness in toads).

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?

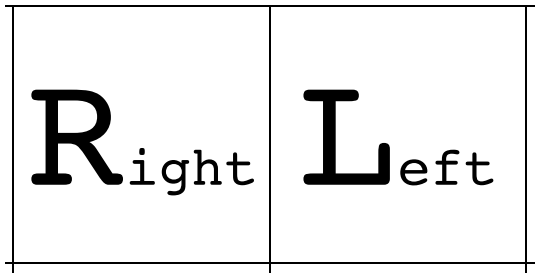
The intuition behind the framework of statistical hypothesis testing

Evidence for or against a hypothesis (such as handedness) can be generated using a simple computational thought experiment with paper and a bag. The process involves assuming a particular hypothesis to be true (the null hypothesis) and then testing whether you can reject it in favor of an alternative hypothesis.

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

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

TODAY: A Roadmap for Drawing Evidence-Based Conclusions in the Absence of Complete Knowledge



Statistical hypothesis testing versus estimation

Both statistical hypothesis testing and estimation (including confidence intervals) use sample data to infer characteristics about the statistical population from which the sample was drawn.

Estimation provides bounds (confidence intervals) around the value of a population parameter, while **statistical hypothesis testing** generates evidence for or against a research hypothesis

Statistical hypothesis testing versus estimation

Both methods (estimation and statistical hypothesis testing) use sample data to make inferences about the population, but they serve different purposes—*estimation focuses on providing a range of values*, while *statistical hypothesis testing focuses on making decisions about the validity of a specific hypothesis*.

Statistical hypothesis testing evaluates whether the observed sample value of a test statistic (i.e., data summary, e.g., mean, variance) significantly differs from the expected value under the null hypothesis, based on the sampling distribution of that statistic for a theoretical population with a specified parameter (null expectation).

Test statistic or Data summary: the proportion of right- and left-handed toads in the population.

Null expectation: the proportion of right- and left-handed toads in the population ARE EQUAL. The null expectation is set in such a way that a sampling distribution for the test statistic can be generated under that expectation.

Statistical hypothesis testing *versus* estimation

Estimation asks - How large is the effect?

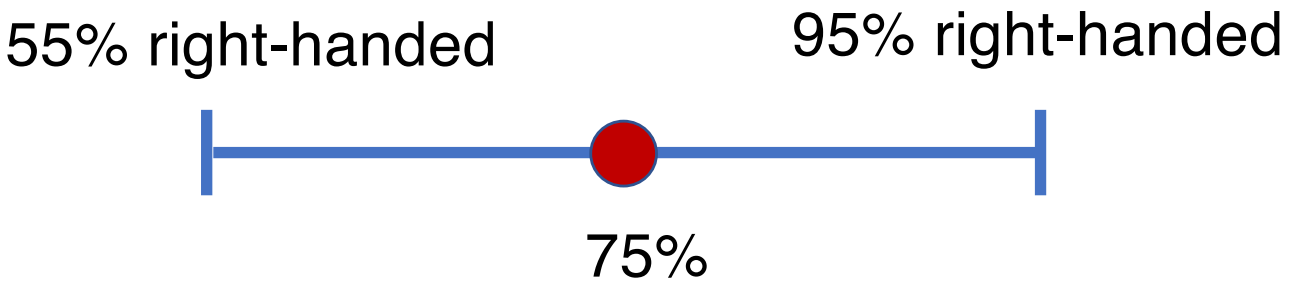
Hypothesis testing asks - Is there any effect at all?

Estimation would ask: What is the proportion of right- and left-handed toads in the population?

Statistical hypothesis testing would ask: Is there a statistically significant difference in the number of toads using their left or right limb to remove the balloon?

Statistical hypothesis testing does not focus on the exact proportion value but on whether there is evidence that the proportion differs from a specified value (commonly 50%/50%, though other values can be tested depending on the hypothesis).

Statistical hypothesis testing *versus* estimation



Estimation thinking: We are 95% confident that the true proportion of right-handed toads is between 55% and 95% of the individuals in the population.

Statistical hypothesis thinking: We are confident that the true proportion of right-handed toads is unlikely to be 50%, indicating that right- and left-handed toads are not equally distributed.

In estimation, we express what the value is likely to be, while in hypothesis testing, we indicate what value is likely not true.

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

Statistical hypothesis thinking: We are confident that the true proportion of right-handed toads is unlikely to be 50%, indicating that right- and left-handed toads are not equally distributed.

In estimation, we express what the value is likely to be, while in hypothesis testing, we indicate what value is likely not true.

In statistical hypothesis testing, one quantifies how unusual the observed sample data (e.g., 4/18 left-handed or 14/18 right-handed) are compared to the assumption of a 50%/50% split.

This is done by contrasting the observed number of right-handed individuals against a sampling distribution generated from a theoretical statistical population where the true proportion is 50%.

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

Is the sample proportion of right-handed (14/18 = 0.78) and left-handed (4/18 = 0.22) toads significantly different from what we would expect in a population where the proportion is 0.5?

Remember that samples can vary due to sampling variation. Because of chance, we don't necessarily expect to see exactly nine right-handed and nine left-handed toads when sampling from a population where the true proportion is 50%/50%.

So, how can we determine if the observed ratio of 14 right-handed to 4 left-handed toads is statistically different from 0.5?

Let’s take a break – 1 minute

The intuition behind the framework of statistical hypothesis testing

You can generate evidence for or against a hypothesis (such as handedness) using something as simple as paper and a bag. The process involves assuming a thypothesis is true (the null hypothesis) and then testing whether you can reject it in favor of the alternative hypothesis.

A statistical theoretical population assumes that 50% of observational units (toads) are left-handed and 50% are right-handed. This theoretical population is considered mathematically infinite.

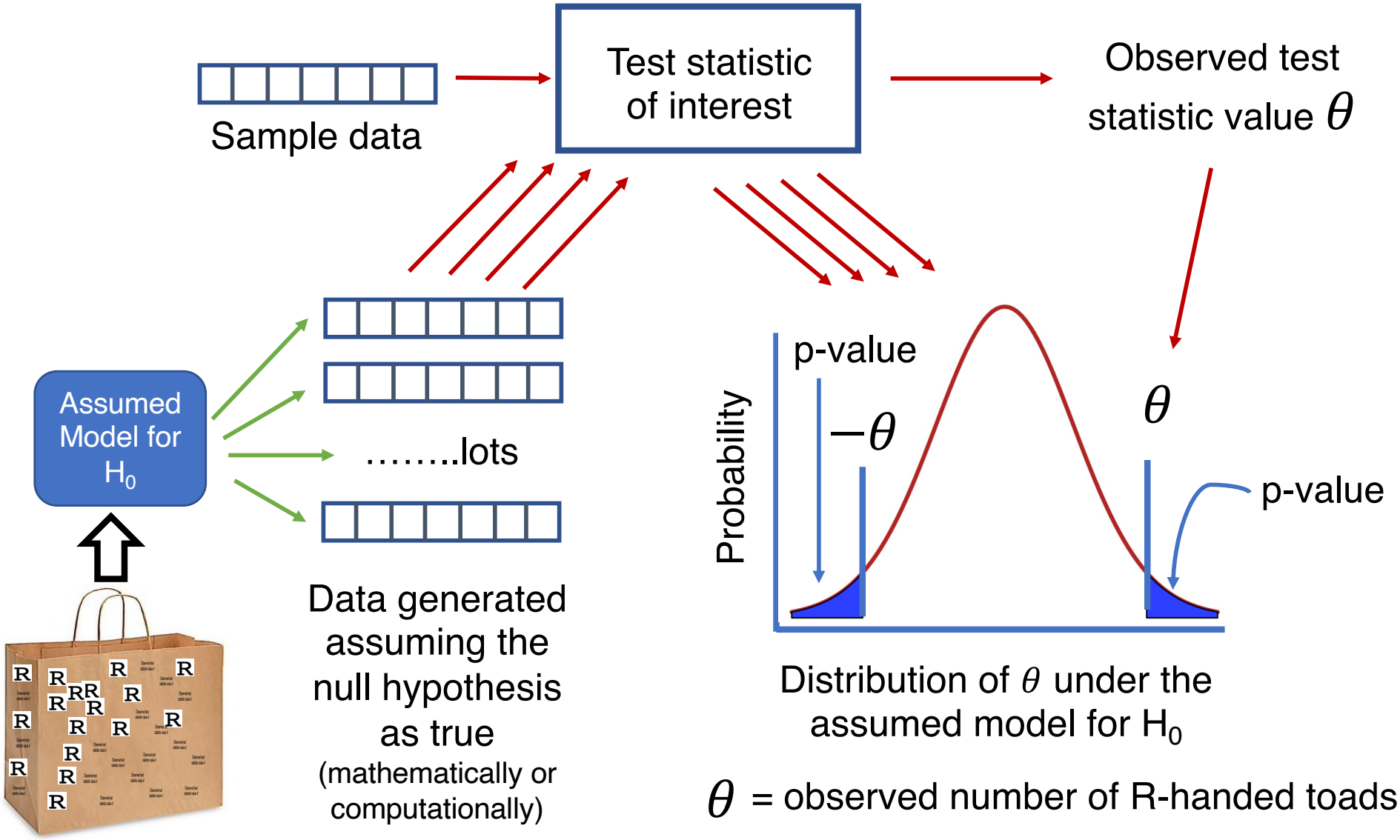
Take one observational unit (a piece of paper) randomly from the bag by closing your eyes and selecting a paper. Record whether it represents 'left' or 'right,' and then return it to the bag (i.e., sampling with replacement). Repeat this process 18 times, matching the number of toads used in the study 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 of interest (number of L and R) for the theoretical population.

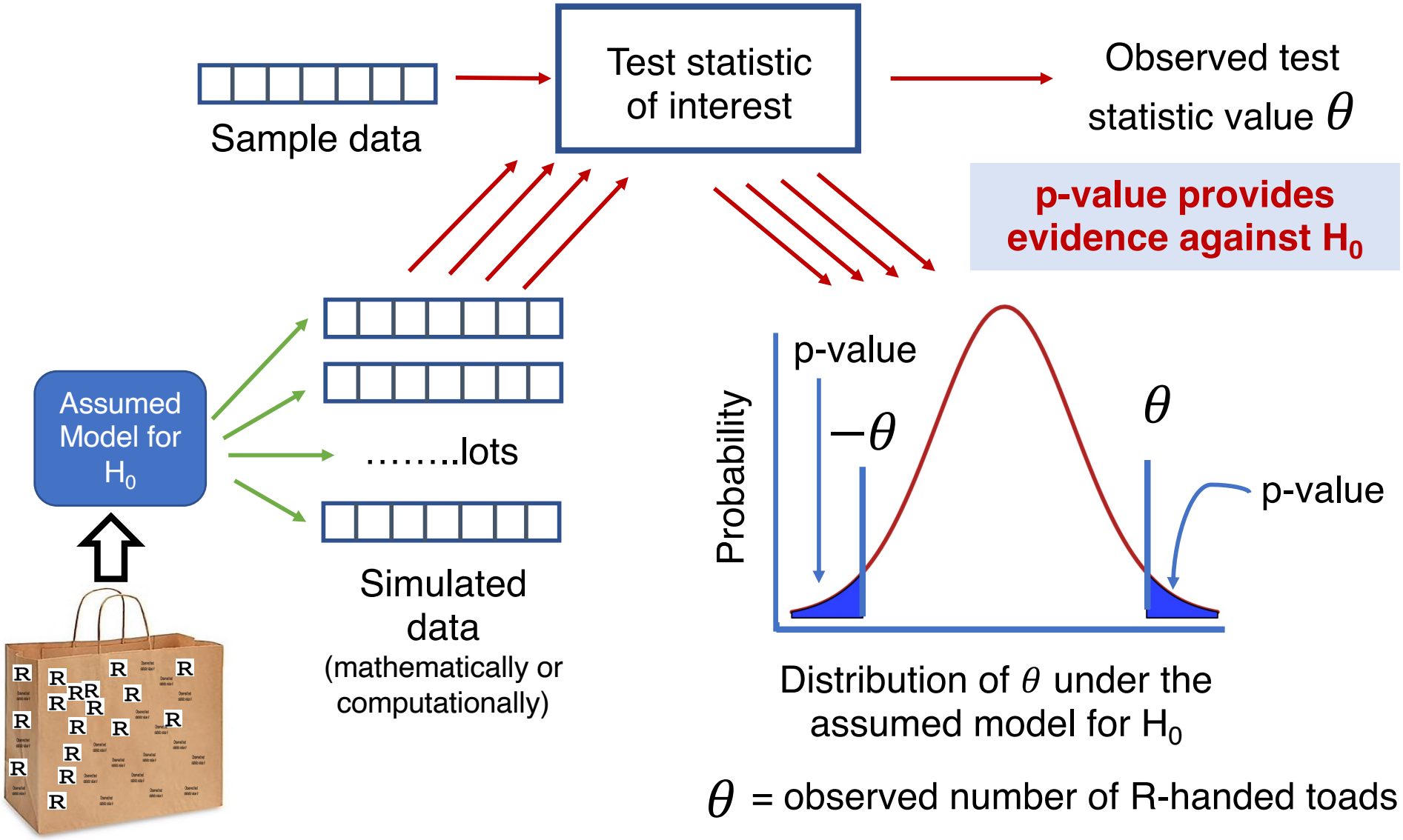
*Resampling is important to ensure that the selection of observational units from the population (e.g., individual pieces of paper) is independent. That is, the selection of any unit (e.g., L or R) must not influence the selection of any other unit.

The “machinery” behind the framework of statistical hypothesis testing



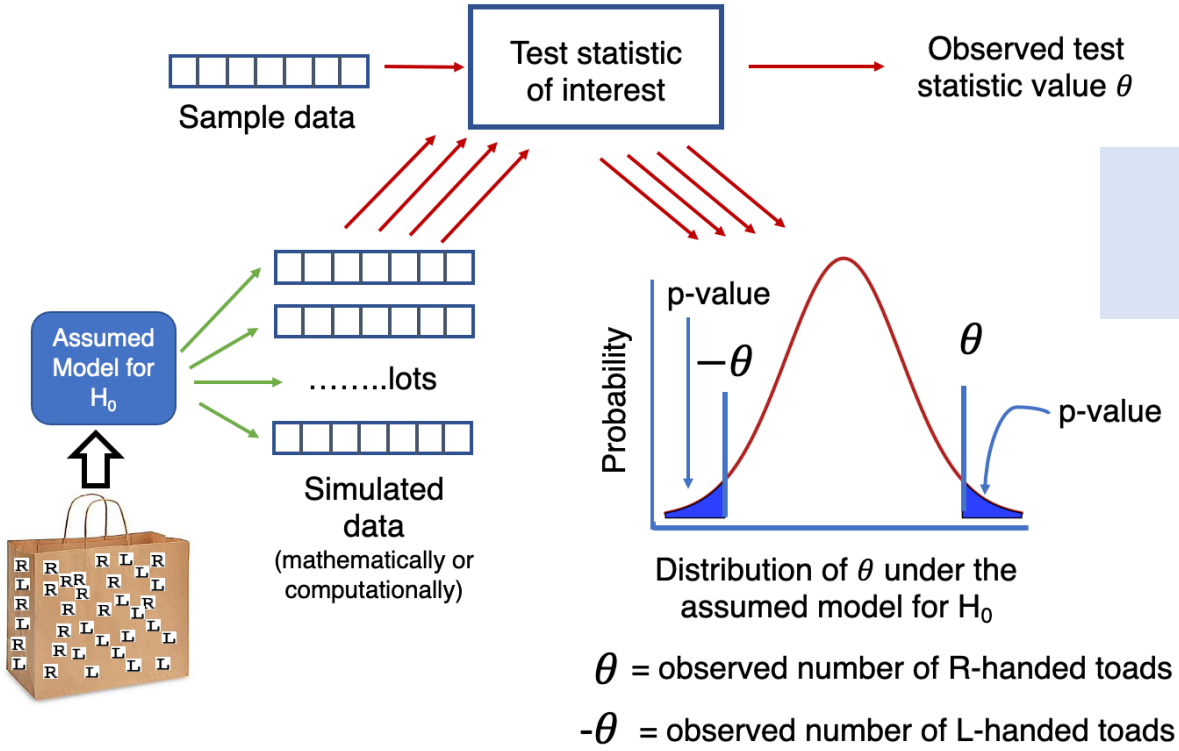
Adapted from https://moderndive.com/

The "machinery" behind the framework of statistical hypothesis testing



Adapted from https://moderndive.com/

The “machinery” behind the framework of statistical hypothesis testing

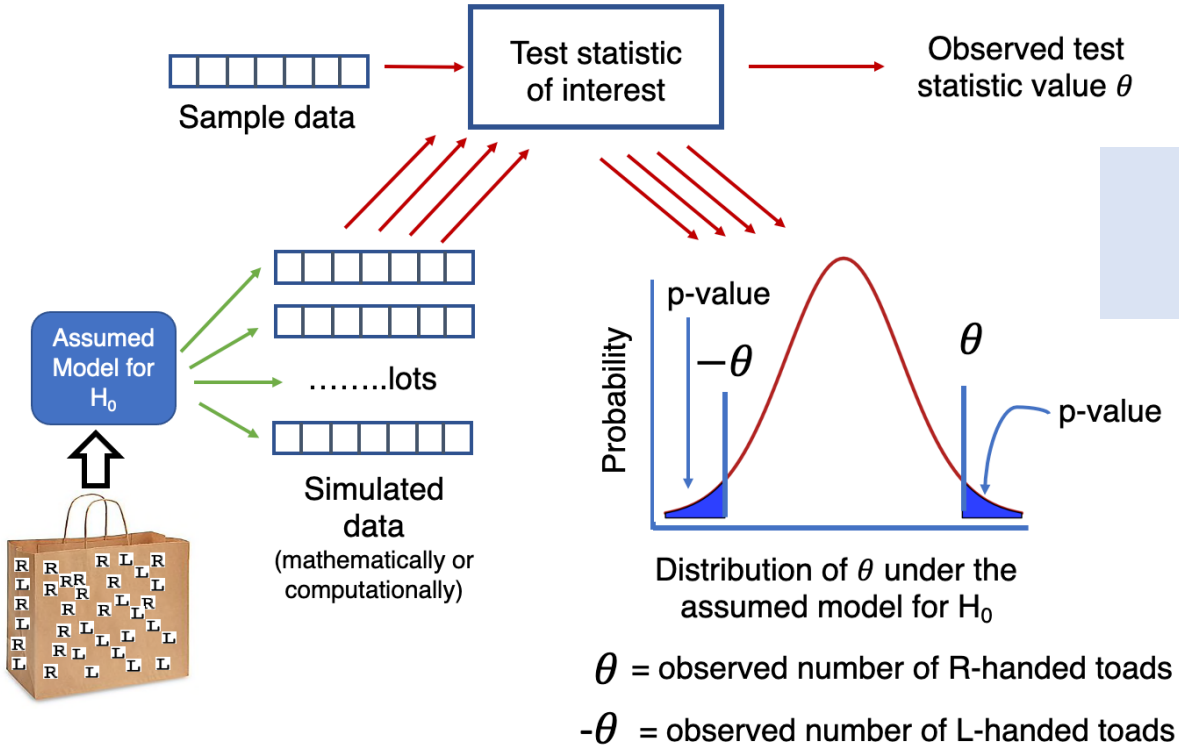


p-value provides evidence against H_0

P-values = (i.e., assuming a theoretical population (model) where 50% of individuals are left-handed and 50% right-handed).

The proportion of samples in the probability distribution that are either equal to or greater than the observed value, or equal to or smaller than the observed value.

The "machinery" behind the framework of statistical hypothesis testing



p-value provides evidence against H_0

In other words, the p-value is the probability of obtaining results at least as extreme as those observed, assuming that the null hypothesis (H_0) is true.

```
> 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
```

```
> 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
```

Assumed theoretical population under H_0

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

Sampling distribution of the test statistic of interest for the theoretical 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 the null hypothesis (H_0), only 4 would have been 0 right-handed, and only 4 would have been 18 right-handed (since the distribution is symmetric). This gives a probability of P = 0.000004.

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 of the test statistic of interest for the theoretical population.

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

How many samples in the null distribution (i.e., assuming the null hypothesis as true) are made of 8 right-handed toads and 10 left-handed toads? 0.1669 or 16.69%

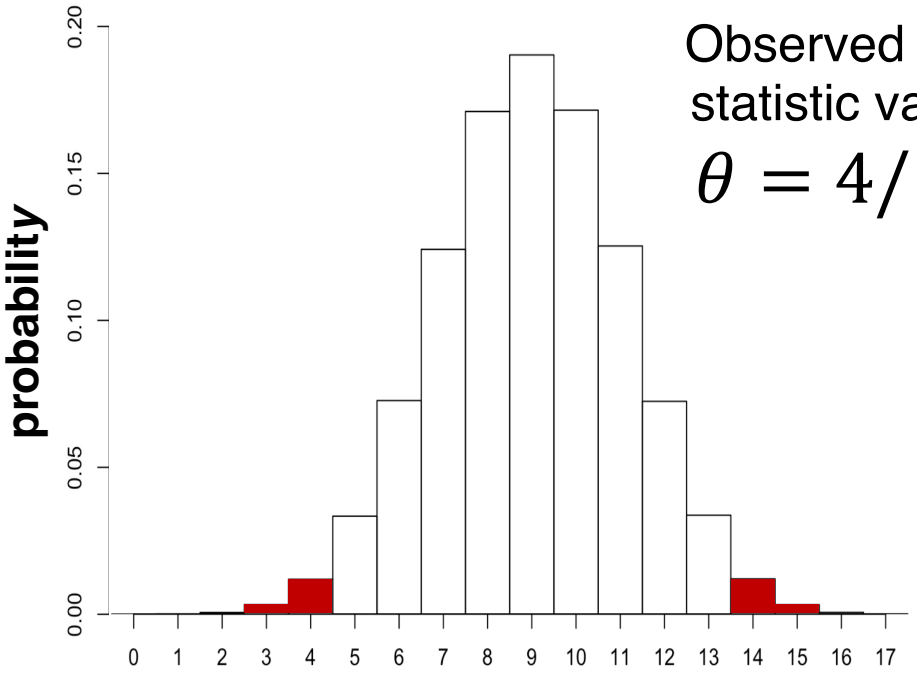
If we had drawn 1,000,000 samples from the population assumed under H_0, 166,900 of them would have resulted in 8 right-handed and 10 left-handed individuals.

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



Observed test statistic value

$\theta = 4/14$

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 statistical hypothesis testing

The **p-value** is a measure used in statistical hypothesis testing to assess the strength of the evidence against the null hypothesis (**H_0**). A **small p-value** (typically less than a significance level, such as 0.05) suggests that the observed data are unlikely under the null hypothesis, providing evidence against it.

However, it’s important to note that the p-value does not *prove* the null hypothesis (H_0) is false, but rather indicates whether the observed data are inconsistent with H_0 based on the chosen significance level.

Decision in statistical hypothesis testing – what do P-values represent?

The statistical hypothesis testing framework typically involves computing a p-value, which serves as a quantitative indicator of support for or against the research hypothesis (e.g., generating evidence for or against handedness in toads).

P-values are used as quantitative evidence against a hypothesis that is usually of no interest (i.e., the null hypothesis), which assumes that the parameter for the theoretical population is true. In this case, it assumes that the proportion of right- and left-handed toads is equal.

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

Decision in statistical hypothesis testing – what do P-values represent?

$P = 0.031$

The p-value is the probability of obtaining results at least as extreme as the one observed, assuming the null hypothesis (H_0) is true.

A p-value quantifies how unusual (i.e., smaller or greater) the observed sample is, based on a theoretical population where the number of right- & left-handed toads is equal. The sampling distribution of this theoretical population represents the null distribution (under the null hypothesis).

Decision in statistical hypothesis testing – what do P-values represent?

P = 0.031

The p-value is the probability of obtaining results at least as extreme as those actually observed, assuming the null hypothesis (H_0) is true.

A p-value quantifies how unusual (i.e., smaller or greater) the observed sample is, based on a theoretical population where the number of right- & left-handed toads is equal. The sampling distribution of this theoretical population represents the null distribution (under the null hypothesis).

Another way to state this is by using the definition of the p-value adopted by the *American Statistical Association*: "The probability under *a specified statistical model* that a statistical summary of the data would be equal to or more extreme than its observed value."

Under this definition:

A specified statistical model is the sampling distribution of the test statistic based on a theoretical population, assuming a particular parameter of interest (e.g., an equal number of right- and left-handed toads).

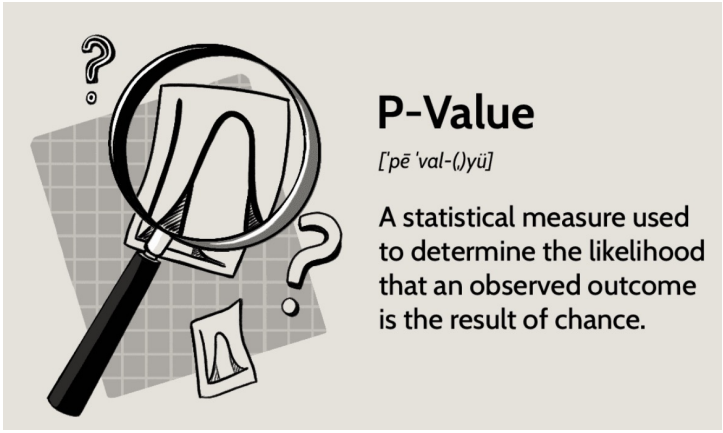
Decision in statistical hypothesis testing – what do P-values represent?

P = 0.031

A p-value quantifies how unusual (i.e., smaller or greater) the observed sample is compared to a theoretical population where the number of right- and left-handed toads is the same.

The smaller the p-value, the stronger the evidence against the initial assumption about the parameter of the theoretical population (i.e., the null hypothesis).

IMPORTANT: This evidence does not confirm that handedness is true, but rather that we have strong evidence against saying the opposite (i.e., claiming that handedness is not true).



Decision in statistical hypothesis testing – what do P-values represent?

P = 0.031

VERY IMPORTANT and potentially "confusing":

P-values provide evidence against the statistical null hypothesis (e.g., that toads do not have handedness and are distributed 50%/50%). However, p-values do not provide evidence for the alternative hypothesis (e.g., that toads have handedness).

Therefore, we can say we have evidence to reject the null statistical hypothesis, but we cannot claim evidence to accept the alternative statistical hypothesis.

However, by rejecting the null statistical hypothesis, we build support for the research hypothesis (remember not to confuse statistical hypotheses with research hypotheses).

Decision in statistical hypothesis testing – what do P-values represent?

P = 0.031

The smaller the P-value, the stronger the evidence against the initial assumption based on the parameter assumed for the theoretical population (i.e., the null hypothesis).

RESULT: Given that the P-value for the toad handedness study was relatively small (P = 0.031), we have sufficient evidence to reject our initial assumption of no handedness. Therefore, the sample data support the hypothesis that toads exhibit handedness.

Remember: Hypotheses cannot be definitively proven or disproven based on sample data.

Instead, hypotheses can only be supported or refuted by the evidence gathered from the data.

Let’s take a break - 1 minute

Statistical hypothesis testing (the handedness of toads)

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

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

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either *likely* or *unlikely* that we would observe a data summary (such as the number of right-handed toads) among the possible values that can occur due to sampling variation (chance alone) from a statistical population assumed to be true under the null hypothesis (50% right- and left-handed).

The p-value is a quantitative measure of the likelihood (chance) to collect the evidence we did given the initial assumption (i.e., based on the theoretical population with equal number of individuals with right- and left-handed).

The decision of 'likely' or 'unlikely' is based on a criterion (the confidence level, as discussed earlier).

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either ***likely*** or ***unlikely*** that we would observe the evidence we collected (i.e., the sample proportion) given the initial assumption of a theoretical population with equal numbers of right- and left-handed individuals.

If the observed results are considered ***likely***, we **'do not reject'** our initial assumption (the null hypothesis, based on the parameter assumed for the theoretical population). In this case, there is not enough evidence to suggest otherwise. In other words, any observed difference between the sample (14 right-handed and 4 left-handed) and the theoretical population value (50% right- and 50% left-handed) can be attributed to chance or sampling variation alone.

Adapted from https://onlinecourses.science.psu.edu/stat504/?q=book/export/html/20

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either ***likely*** or ***unlikely*** that we would observe the evidence we collected (i.e., the sample proportion) given the initial assumption of a theoretical population with equal numbers of right- and left-handed individuals.

If the observed results are considered ***likely***, we **'do not reject'** our initial assumption (the null hypothesis, based on the parameter assumed for the theoretical population). In this case, there is not enough evidence to suggest otherwise. In other words, any observed difference between the sample (14 right-handed and 4 left-handed) and the theoretical population value (50% right- and 50% left-handed) can be attributed to chance or sampling variation alone.

if the observed (sample) data is ***unlikely***, then one of two things must be true:

Our initial assumption (that the proportions are equal) is incorrect, and we should 'reject' the initial assumption. In this case, we could say, 'we have strong evidence against the initial assumption.', i.e., H_0.

Alternatively, our initial assumption could still be correct, but we have experienced a rare event, meaning we drew a very unusual sample purely by chance. In this case, we may have made an error in rejecting the initial assumption based on this unlikely outcome.

Adapted from https://onlinecourses.science.psu.edu/stat504/?q=book/export/html/20

The process of statistical hypothesis testing

Population of interest (toads)

Sample

Test statistic: number of right-handed toads

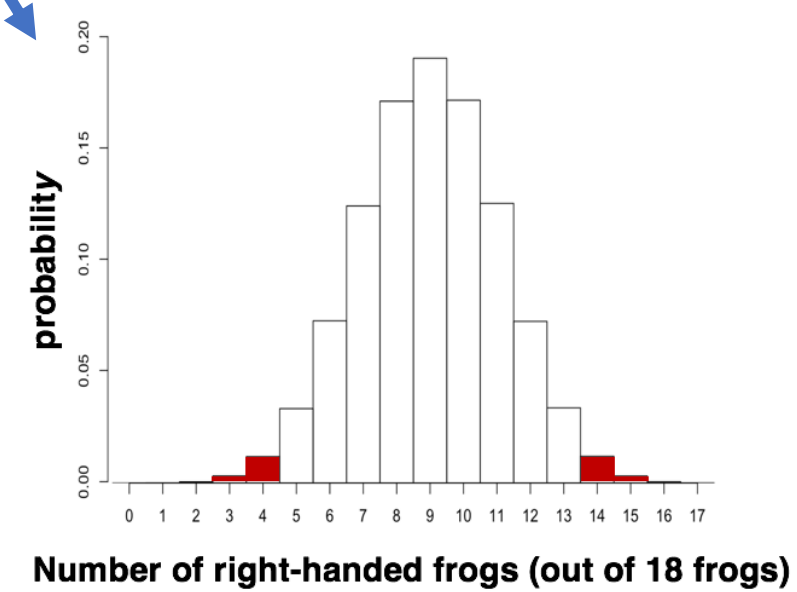
Compare the observed test statistic from the sample data (14 right-handed and 4 left-handed) to the expected distribution under the null hypothesis (null distribution).

Decisions:

Do not reject the initial assumption, i.e., parameter for the theoretical value (equal right and left) for the test of statistic of interest (large p-value). OR **Reject** the initial assumption (small p-value).

Theoretical population - parameter assumed = 50% R & 50% L

Generate the sampling distribution for the test statistic of interest (e.g., the number of right-handed and left-handed toads) based on the theoretical population, also known as the null distribution.



The two hypotheses in statistical hypothesis testing

Our initial assumption (parameter) to build the sampling distribution for the theoretical population is called:

H_0 (null hypothesis): any observed difference between the sample and the theoretical population value is due to chance alone; i.e., the observed sample data is a common sample within the theoretical population (initial assumption; in the toad case - equal proportion of right- and left-handed).

The null hypothesis is a specific claim about a theoretical population parameter, made to serve as a basis for argument and to generate evidence for or against it. Typically, it represents a hypothesis of no particular interest.

H_0: the proportion of right- and left-handed toads in the population are equal.

The two hypotheses in statistical hypothesis testing

Our initial assumption (parameter) to build the sampling distribution for the theoretical population is called:

The alternative hypothesis (H_A) encompasses all other possible parameter values aside from the one specified in the null hypothesis.

In other words, if our initial assumption (theoretical population value) is incorrect, the observed sample data likely come from a population that does not have an equal number of right- and left-handed individuals. Unlike the null hypothesis, H_A is not specific and represents a broader range of possible outcomes.

H_A: the proportion of right- and left-handed toads in the population differ.

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

Do other animals exhibit handedness as well?

H_0 (null hypothesis): any observed difference between the sample proportion and the theoretical population proportion assumed for the sake of argument is due to chance alone.

H_0 (null hypothesis): the number of right- & left-handed toads are equal; i.e., the true population value for the ratio between right- and left-handed toads is 1.

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

Do other animals exhibit handedness as well?

H_0 (null hypothesis): any observed difference between the sample proportion and the theoretical population proportion assumed for the sake of argument is due to chance alone.

H_0 (null hypothesis): the number of right- & left-handed toads are equal; i.e., the true population value for the ratio between right- and left-handed toads is 1.

H_A (alternative hypothesis): includes all other possible parameter values, i.e., all possible toad populations except the one stated in the null hypothesis.

H_A (alternative hypothesis): the number of right- and left-handed toads differ in the population; i.e., the true population value (ratio) for the ratio between right- and left-handed toads does not equal 1.

Drawing a conclusion using the P-value as evidence for or against a research hypothesis

Do other animals exhibit handedness as well?

P = 0.031

The decision threshold is called *significance level* and its symbol is α (alpha). In biology, the mostly used α = 0.05 (and often α = 0.01). If P is smaller or equal than α, we assume to have enough evidence to reject the null hypothesis (H_0) in favour of the alternative (H_A).

CONCLUSION:

Assuming a significance level of 0.05 (the threshold for deciding whether to reject H0), the results from the balloon experiment suggest strong evidence that toads exhibit handedness.

The smaller the p-value, the stronger the evidence against the statistical null hypothesis (H0), and consequently, the stronger the support for the scientific hypothesis of handedness.

Drawing a conclusion using the P-value as evidence for or against a research hypothesis

Do other animals exhibit handedness as well?

Note that these two statistical hypotheses are about populations and not samples:

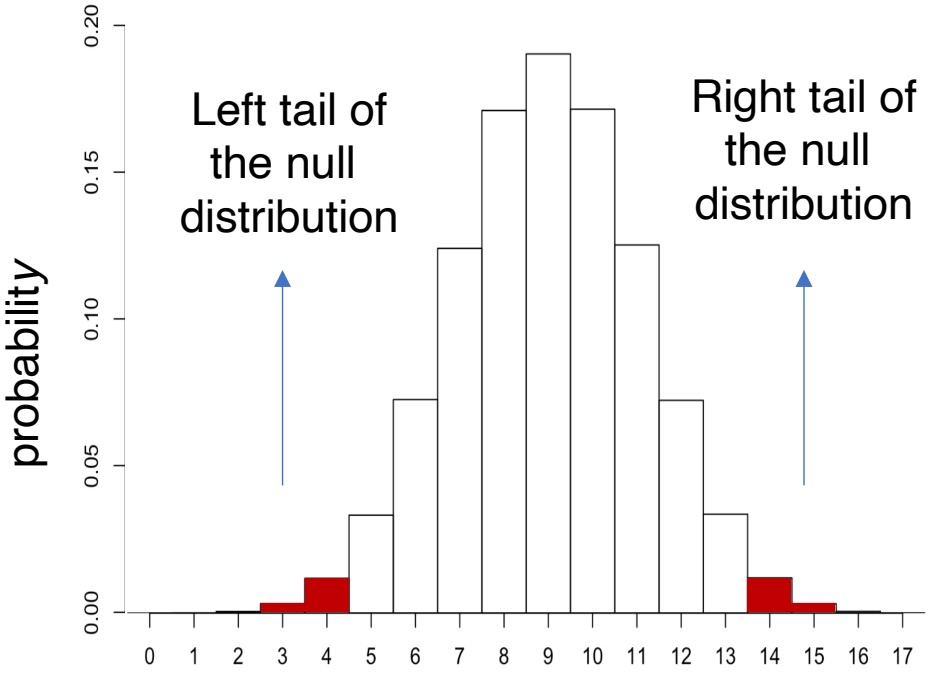
H_0 (null hypothesis): the true population value for the ratio between right- and left-handed toads is 1.

H_A (alternative hypothesis): the true population value for the ratio between right- and left-handed toads does not equal 1.

Drawing a conclusion using the P-value as evidence for or against a research hypothesis

Do other animals exhibit handedness as well?
Directionality in hypothesis testing:

This is called a two-sided (or two-tailed) test because we would reject the null hypothesis if the sample had either a much higher or much lower ratio of left- to right-handed toads compared to 1. Both extremes, whether more left- or more right-handed, provide evidence against the null hypothesis of equal proportions.



In this case, we were interested in determining whether there was any preference for limb use, but not in whether the right limb or left limb was specifically preferred over the other. Investigating a specific preference for one limb would require a one-tailed test, which will be covered in a later lecture.

The process of statistical hypothesis testing: critical details

Statistical hypothesis testing examines how unusual the observed sample data is when compared to the distribution assumed under the null hypothesis.

Hypotheses are statements about populations but are tested using sample data.

Hypothesis testing (usually) assumes that sampling is random.

The null hypothesis is typically the simplest or default statement, while the alternative hypothesis reflects the hypothesis of greater interest.

The null hypothesis is often a specific statement about a population parameter, whereas the alternative hypothesis tends to be less specific.

Decision in statistical hypothesis testing: In light of the 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 involve:

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

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

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.

Computer 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).

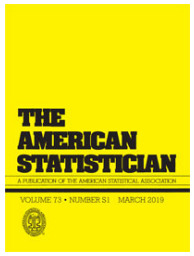
Draw a conclusion by contrasting the p-value against the significance level (α). If the p-value is greater than α, then do not reject H_0; if P-value is smaller or equal than α, then reject H_0.

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)

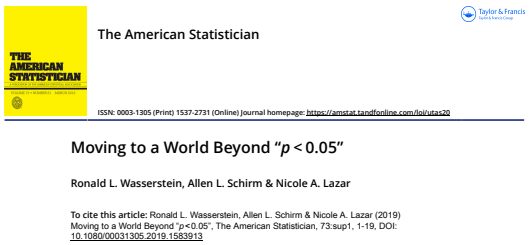
1. P-values can indicate how incompatible the observed data are with a specified statistical model (e.g., the one assumed under H_0).

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

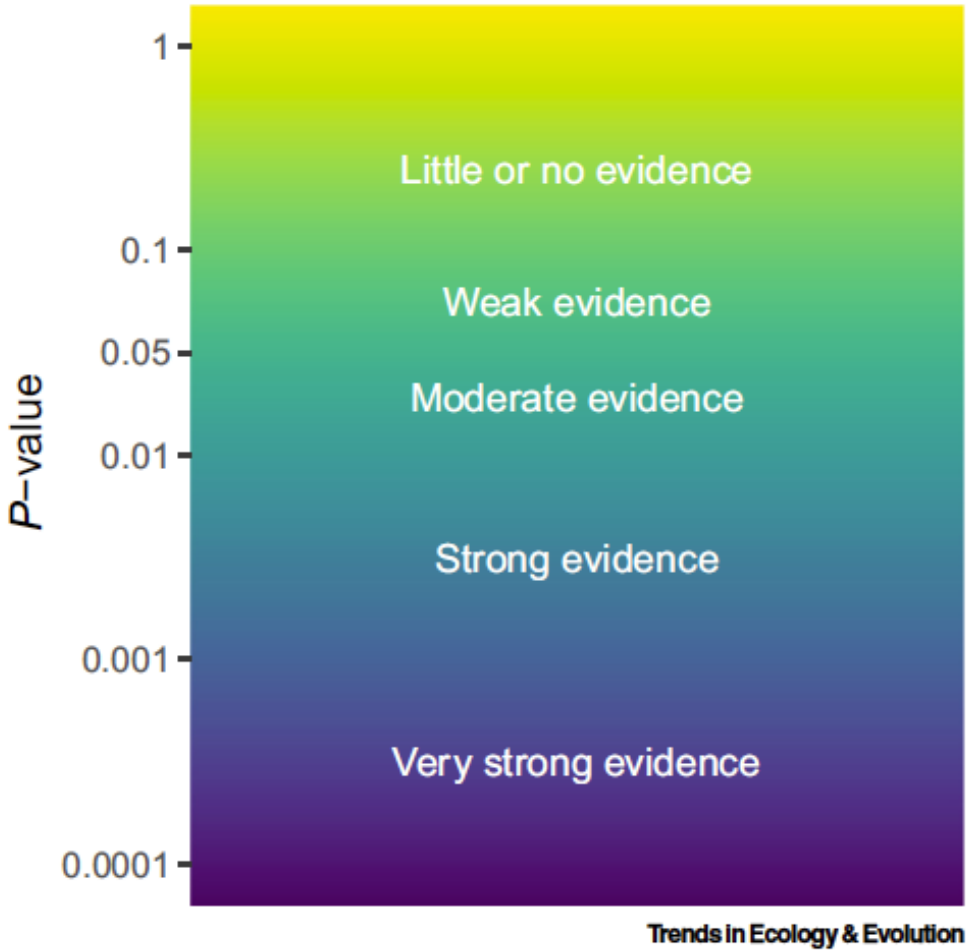


Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

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