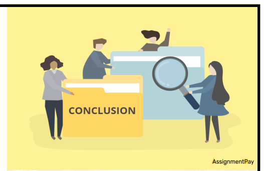
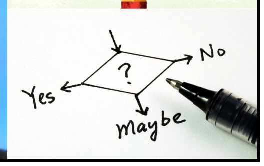
Generating evidence-based conclusion without complete biological knowledge





1

What is meant by evidence from the scientific literature

Evidence in general means information, facts or data supporting (or contradicting) a claim, prediction assumption or hypothesis.

When referring to “evidence from the scientific literature”, people generally mean the empirical studies published in peer-reviewed scholarly journals.

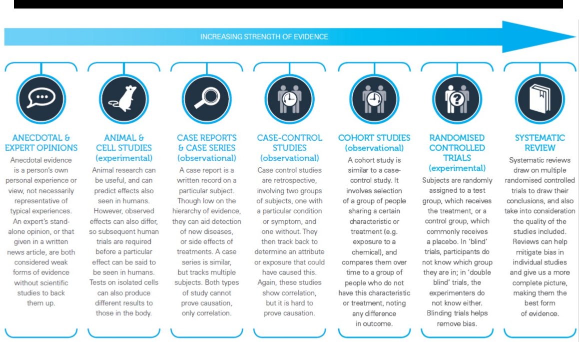
But not only the scientific literature is relevant to science-based evidence. Many government and non-governmental reports are also relevant

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

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

2

Types of scientific evidence



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

3

Statistical hypothesis testing as a quantitative framework to generate evidence for 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 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.

4

What is a research hypothesis?!

A hypothesis is a supposition or proposed explanation made on the basis of 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).

5

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.

6

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).

7

Tackling research hypotheses using 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 estimating 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).

8

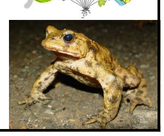
Back to statistically testing the hypothesis of handedness

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

They 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?

9

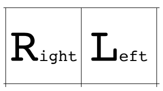
The intuition behind the framework of statistical hypothesis testing

You can generate evidence for or against a hypothesis (handedness) using a computational thought experiment based on paper and a bag. All you need is to assume a particular hypothesis as true (null hypothesis) and then reject it (or not) is support of an alternative hypothesis!

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.

TODAY: A road map for understanding *evidence-based conclusions* without complete knowledge



10

Statistical hypothesis testing versus estimation

Both statistical hypothesis testing and estimation use sample data to make inferences about the statistical population from which the sample was taken.

While estimations puts bounds (confidence intervals) on the value of a population parameter.

And statistical hypothesis testing generates evidence for or against a research hypothesis.

11

Statistical hypothesis testing versus estimation

Both statistical hypothesis testing and estimation use sample data to make inferences about the statistical population from which the sample was taken.

While estimations puts bounds (confidence intervals) on the value of a population parameter.

And statistical hypothesis testing generates evidence for or against a research hypothesis.

Statistical hypothesis testing asks whether the observed sample value for a given test statistic (i.e., data summary) differs from a specific "null" expectation (null hypothesis) based on the sampling distribution of the same test statistic for a theoretical statistical population assuming a particular theoretical parameter.

12

Statistical hypothesis testing versus estimation

Both statistical hypothesis testing and estimation use sample data to make inferences about the statistical population from which the sample was taken.

Statistical hypothesis testing asks whether the observed sample value for a given test statistic (i.e., data summary) differs from a specific "null" expectation (null hypothesis) based on the sampling distribution of the same test statistic for a theoretical statistical population assuming a particular theoretical parameter.

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.

13

Statistical hypothesis testing *versus* estimation

Estimation asks - How large is the effect?

Hypothesis testing asks - Is there any effect at all?

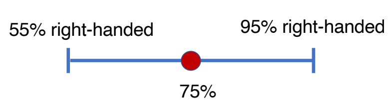
Estimation would ask: What is the proportion of right- and left-handed frogs in the population?

Statistical hypothesis would ask: Is there a statistical difference (effect) in the number of toads that used their left or right limb to remove the balloon?

Statistical hypothesis testing is not about the exact proportion value but whether we can generate evidence that it differs or not from a value of (usually) **NO** interest (here 50%/50% but other values can be used).

14

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 not likely to be in equal proportion (50% right- and 50% left-handed).

Instead of stating what the value is likely (estimation), we state what value is likely not (hypothesis testing)!

15

Statistical hypothesis testing: generating evidence-based conclusion without complete biological knowledge

Statistical hypothesis thinking: We are confident that the true proportion of right-handed toads is not likely to be in equal proportion (50% right- and 50% left-handed).

Instead of stating what the value is likely (estimation), we state what the value is likely not (statistical hypothesis testing)!

In statistical hypothesis testing, one quantifies how unusual the observed sample data (4/18 left or 14/18 right) is in contrast to the assumption that they are 50%/50%

This is done by contrasting the observed number of right-handed individuals against a sampling distribution of number of right-handed toads for a theoretical statistical population where the proportion is truly 50%).

16

Statistical hypothesis testing: generating evidence-based conclusion without complete biological knowledge

Is the sample proportion of right-handed (14/18 = 0.78) and right-handed (4/18 = 0.22) toads really different from what would expected from a statistical population of toads that would have a proportion equal to 0.5?

Remember that samples vary due to sampling variation. Because of the effects of chance during sampling, we don't really expect to see exactly nine right-handed and nine left-handed toads when we sample from a statistical population in which 50%/50% are truly left/right handed!

So, how can we generate evidence that 14 right-handed frogs against 4 left-handed frogs is statistically different from 0.5?

17

Let's take a break - 2 minutes

18

The intuition behind the framework of statistical hypothesis testing

You can generate evidence for or against a hypothesis (handedness) even using paper and a bag. All you need is to assume a particular hypothesis as true (null hypothesis) and then reject it (or not) in support of the alternative hypothesis!

Statistical theoretical population where 50% of observational units (toads) are left-handed and 50% right-handed. This theoretical population is mathematically infinite.

Take one observational unit (piece of paper) randomly at the time (close eyes and take a paper) out of the bag, write it down whether a left or right and return to the bag (i.e., sampling with replacement*. Repeat this 18 times (i.e., number of toads used by the toad study (Bisazza et al. 1996).

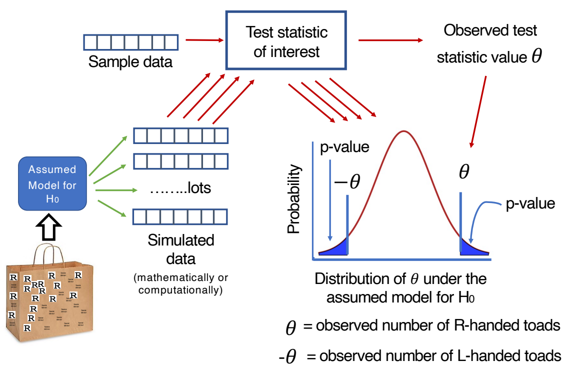
1 sample: 14 R & 4 L
2 sample: 8 R & 10 L
.
.
.
Large number of samples (~Infinite)

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

*Resampling is important to assure that the selection of observational units in the population (e.g., individual piece of paper here) must be independent, i.e., the selection of any unit (e.g., L or R) of the population must not influence the selection of any other unit.

19

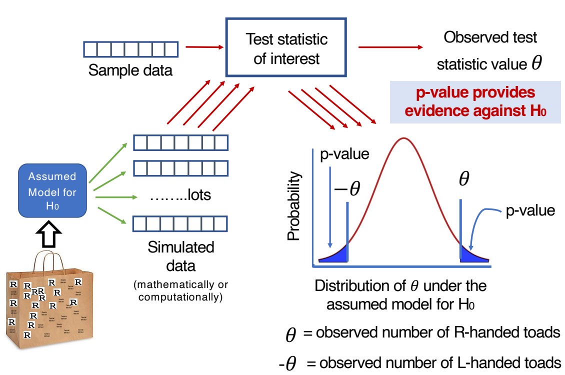
The "machinery" behind the framework of statistical hypothesis testing



Adapted from https://moderndive.com/

20

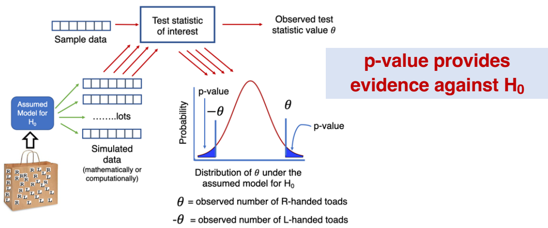
The "machinery" behind the framework of statistical hypothesis testing



Adapted from https://moderndive.com/

21

The "machinery" behind the framework of statistical hypothesis testing

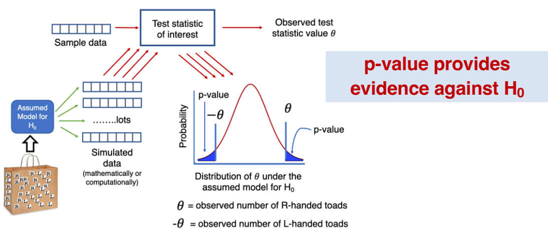


P-values = (i.e., assuming a theoretical population (model) where 50% of individuals are left-handed and 50% right-handed).

The proportion of samples in the frequency distribution (probability distribution) that were equal or greater than the observed AND equal or smaller than the observed.

22

The "machinery" behind the framework of statistical hypothesis testing



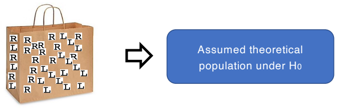
P-values = (i.e., assuming a theoretical population (model) where 50% of individuals are left-handed and 50% right-handed). The proportion of samples in the frequency distribution (probability distribution) that were equal or greater than the observed AND equal or smaller than the observed.

In other words, the P-value is the probability of obtaining results at least as extreme as the results actually observed assuming the H_0 as true.

23

```
> 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
```



24

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 and only 4 would had been 18 right-handed (the distribution is obviously symmetric), i.e., 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

25

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.

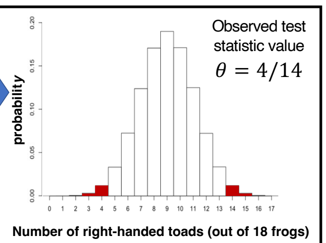
Number of right-handed toads	Probability of those samples
0	0.000004
1	0.00007
2	0.0006
3	0.0031
4	0.0117
5	0.0327
6	0.0708
7	0.1214
8	**0.1669**
9	0.1855
10	**0.1669**
11	0.1214
12	0.0708
13	0.0327
14	0.0117
15	0.0031
16	0.0006
17	0.00007
18	0.000004
Total	1.0

26

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

27

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

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

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

28

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 results actually observed assuming the H_0 as true.

A *P*-value then estimates how unusual* (i.e., smaller or greater) the observed sample is according to a theoretical population where the number of right- and left-handed toads are the same. The sampling distribution of the theoretical population in the null distribution (under the null hypothesis.

29

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 results actually observed assuming the H_0 as true.

A *P*-value then estimates how unusual* (i.e., smaller or greater) the observed sample is according to a theoretical population where the number of right- and left-handed toads are the same. The sampling distribution of the theoretical population in the null distribution (under the null hypothesis.

Another way of stating the above is by using the definition of 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:

Specified statistical model = Sampling distribution of the same test statistic but based on a theoretical population assuming a particular parameter of interest (e.g., number of right- & left-handed toads are equal).

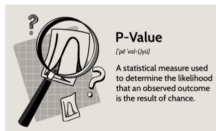
30

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

P = 0.031

A *P*-value then estimates how unusual* (i.e., smaller or greater) the observed sample is according to a theoretical population where the number of right- and left-handed toads are the same.

The smallest the P-value, the stronger the evidence against the initial assumption based on the parameter assumed for the theoretical population (i.e., null hypothesis). IMPORTANT: That's not to say handedness is true but rather that we have strong evidence not to say the contrary (i.e., to say that handedness is not true).



31

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

P = 0.031

VERY IMPORTANT and "confusing":

The way p-values are estimated, they provide evidence against the statistical null hypothesis (i.e., that toads do not have handedness, 50%/50%) but p-values do not provide evidence for the alternative hypothesis (i.e., handedness).

So we can say that we have evidence to reject the null statistical hypothesis BUT we cannot say that we have evidence to accept the alternative statistical hypothesis.

BUT, by rejecting the statistical null hypothesis, we build evidence towards the research hypothesis (do not confuse statistical with research hypotheses).

32

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

P = 0.031

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

RESULT: Given that the p-value for the toad handedness study was small (P=0.031), there is evidence to reject (refute) our initial assumption of no handedness. Therefore the sample data support the hypothesis of handedness

Remember: Hypotheses cannot be proven right or wrong from sample data. Hypotheses can only be said to be supported by the data.

33

Let's take a break - 2 minutes

34

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.

35

Statistical hypothesis testing (the handedness of toads)

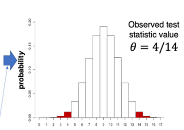
Null hypothesis (H_0): the proportion of right- and left-handed toads in the population ARE equal.

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

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

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 frequency distribution (probability distribution) that were equal or greater than the observed AND equal or smaller than the observed.

In other words, the P-value is the probability of obtaining results at least as extreme as the results actually observed assuming the H_0 as true.

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

36

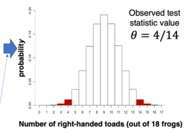
The P-value represents the chance of observing a data summary (e.g., number of right-handed toads) as extreme as or more extreme than what can be observed within the frequency distribution assumed under H0 (null distribution).

As such, the p-value is a quantitative measure of the agreement or disagreement (fit) with the value assumed under H0. Low p-values, low agreement, therefore perhaps consider H0 to be false (reject it).

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

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 frequency distribution (probability distribution) that were equal or greater than the observed AND equal or smaller than the observed.

In other words, the P-value is the probability of obtaining results at least as extreme as the results actually observed assuming the H_0 as true.

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

37

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either ***likely*** or ***unlikely*** that we would observe a data summary (from the data we have; number of right-handed toads) among the possible values that can be obtained under sampling variation (chance alone) from a population (statistical) assumed to be true (H0) for the sake of argument (50%/50%).

The p-value is a quantitative measure of the likelihood (change) 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 "Likely" or "unlikely" is based on a criterium (more later)

38

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either ***likely*** or ***unlikely*** that we would collect the evidence we did (i.e., the proportion we found in the sample data) given the initial assumption (theoretical population with equal number of individuals with right- and left-handed).

If we consider as **likely**, then we "**do not reject**" our initial assumption (the null distribution generated under the parameter assumed for the theoretical population). There is not enough evidence to do otherwise. In other words, any observed difference between the sample (14 right-handed and 4 right-handed) and the theoretical population value (50%/50%) is due to chance alone (due to chance under sampling variation alone).

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

39

Decision in statistical hypothesis testing – using p-values

P = 0.031

It is either ***likely*** or ***unlikely*** that we would collect the evidence we did (i.e., the proportion we found in the sample data) given the initial assumption (theoretical population with equal number of individuals with right- and left-handed).

If it is **likely**, then we "**do not reject**" our initial assumption (the null distribution generated under the parameter assumed for the theoretical population). There is not enough evidence to do otherwise. In other words, any observed difference between the sample (14 right-handed and 4 right-handed) and the theoretical population value (50%/50%) is due to chance alone.

If it is **unlikely**, then either:

- Our initial assumption (proportion is equal) is truly incorrect and we should **"reject"** the initial assumption. We could say "we have strong evidence against the initial assumption".
- OR our initial assumption is correct. We then experienced a truly unusual sample data (i.e., we made a mistake in rejecting the initial assumption); here, just by chance we sampled a very unusual sample.

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

40

The process of statistical hypothesis testing

Test statistic: number of right-handed toads

Population of interest (toads)

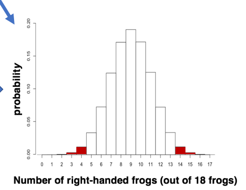
Sample

contrast the test statistic for the sample data (14 right handed and 4 left-handed) against the null distribution

Decisions:
Do not reject the initial assumption, i.e., parameter for the theoretical value (9 right- handed toads) 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 (here number of right-handed and left-handed toads) for the theoretical population (i.e., null distribution).



41

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).

The null hypothesis is a specific statement about a theoretical population parameter made for the purposes of argument and generating evidence for or against it (usually the hypothesis of no interest).

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

42

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).

The null hypothesis is a specific statement about a theoretical population parameter made for the purposes of argument and generating evidence for or against it.

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

In other words, our initial assumption (theoretical value for the population) is incorrect. As such, it is more likely that the observed sample data come from a population that does not have an equal number of individuals that are right- and left-handed. H_A (unlike H_0) is not specific.

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

43

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 and the theoretical population value is due to chance alone.

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

44

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 and the theoretical population value is due to chance alone.

H_0 (null hypothesis): the number of right and 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 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.

45

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 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 (decision threshold about whether reject or not the H_0), the data generated by the balloon experiment generated evidence that toads exhibit handedness.

The smaller the p-value, the stronger the evidence is that the statistical null hypothesis should be rejected; and the stronger the evidence towards the scientific hypothesis of handedness.

46

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.

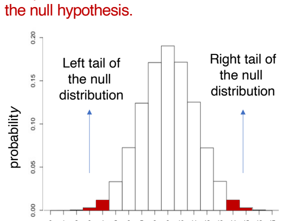
H_A (alternative hypothesis): the true population value for the ratio between right- and left-handed toads does not equal 1.

47

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:

We call the test used here **two-sided (or two-tailed test)** because either a sample with a much higher ratio of left- to right-handed toads than 1 **OR** a sample with a much smaller ratio than 1 would have led to the rejection of the null hypothesis.



Number of right-handed frogs (out of 18 frogs)

Here we were interested whether there was a preference but not whether the right limb or left limb were preferred over the other (that would had been a one tail test; more on that in a later lecture).

48

The process of statistical hypothesis testing: critical details

Statistical hypothesis testing asks how unusual it is to get the observed value for the sample data within the distribution built assuming the null hypothesis as true.

Statistical hypothesis are about populations but are tested with data from samples.

Statistical hypothesis (usually) assumes that sampling is random.

The null hypothesis is usually the simplest statement, whereas the alternative hypothesis is usually the statement of greatest interest.

A null hypothesis is often specific (specific parameter for the theoretical population); an alternate hypothesis often is not.

49

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: p-value is not the probability of a null hypothesis being true; it is simply a quantitative metric that allows us to state strong or small evidence against H_0.

50

Statistical hypothesis testing involve:

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. Estimate 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 estimated 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.

51

What does the significance level (α level) represent?

There is a lot of disagreement among statisticians and users about whether to accept or reject statistical hypotheses based on p-values.

i.e., whether to use α as a threshold for making a decision to state whether a p-value is non-significant (do not reject H_0) or a p-value is significant (reject H_0 in favour of H_A).

Although I agree with these arguments it is unlikely that radical changes will arrive in research behaviour any time soon!


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

52

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.

53

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 significance is easily said than done. The majority of researchers do report results as significant or non-significant. We will try to guide you in a more nuanced ways in BIOL322 but hard to get away from this common culture in the statistical applications biology and most other fields.

54

Use p-values using "the language of evidence" against H_0

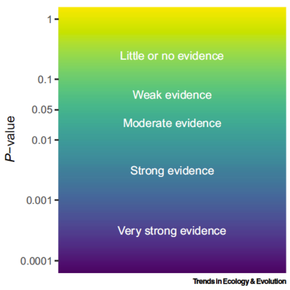


Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

Note: because the p-value is based on the H_0, the evidence is against H_0 and not in favour of H_A. So, we have evidence to reject H0 (one fixed assumed parameter) but not accept HA (many potential parameters can fit HA (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.

55