A quick (re)tour of statistical hypothesis testing

Think of this statement:

Hypothesis: Electric vehicles (EVs) do not generate pollution (i.e., generate zero pollution).

Rejecting this null does not tell us how much pollution it generates…could be very small or could be very large. By rejecting that hypothesis, all we can say that it is likely that generate some.

So statistical hypothesis testing can generate evidence against a statement but (in most cases) it does not tell us by how much.

Note: EVs do not generate direct emissions but emissions are not the only source of pollution and EVs do generate indirect emissions (e.g., construction of the vehicle, technical support, etc).

statistical hypothesis testing is an intimate stranger!!

Most users know how to implement and interpret it, but they don't really understand its philosophy and how it really works.

Research question and what approach would you prefer?

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?

Which specific types of evidence-based data would be most valuable in addressing this question?

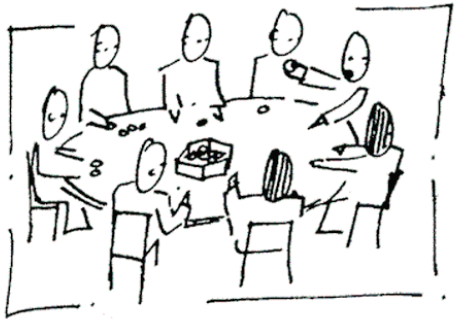
Translating the research question into a statistical question:

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?

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

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

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



Which specific types of evidence-based data would be most valuable in addressing this question?

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?

Statistical hypothesis testing

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

"Qualitative" statement

Parameter estimate & uncertainty

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

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

"Quantitative" statement

Which specific types of evidence-based data would be most valuable in addressing this question?

Humans love yes/no questions (qualitative)

However, all these three answers provide evidence towards handedness.

Statistical hypothesis testing

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

"Qualitative" statement

Parameter estimate & uncertainty

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

We are 95% confident that the proportion of right-handed over left-handed toads varies between 72% and 75%.

"Quantitative" statement

Estimation [& associated confidence intervals] and statistical hypothesis testing agree but have different interpretations

45% right-handed

65% right-handed

H_0

55%

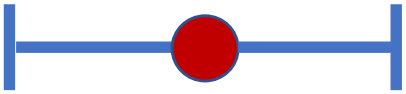
"Quantitative" statement

Don't reject H_0; $p > 0.05$

"Qualitative" statement

30% RH

48% RH



39%

"Quantitative" statement

Reject H_0; $p < 0.05$

"Qualitative" statement

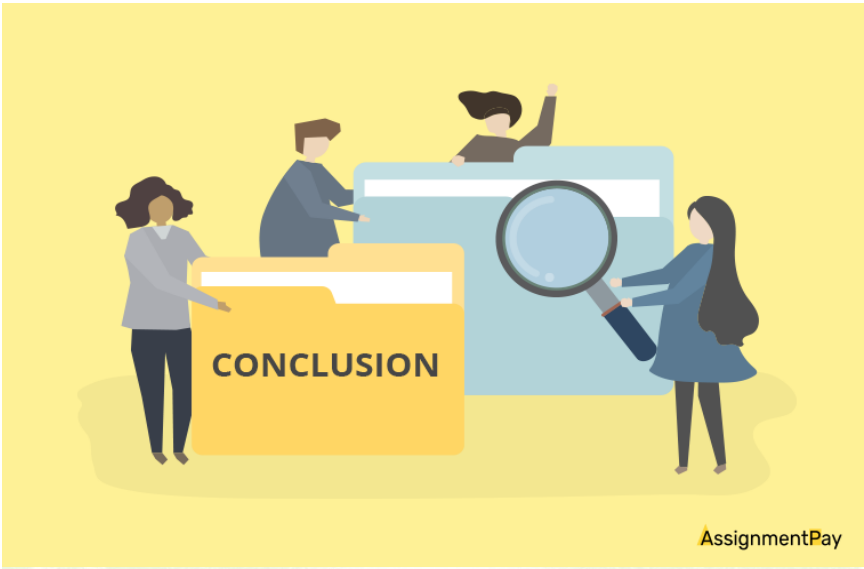
Generating evidence-based conclusions without complete biological knowledge!

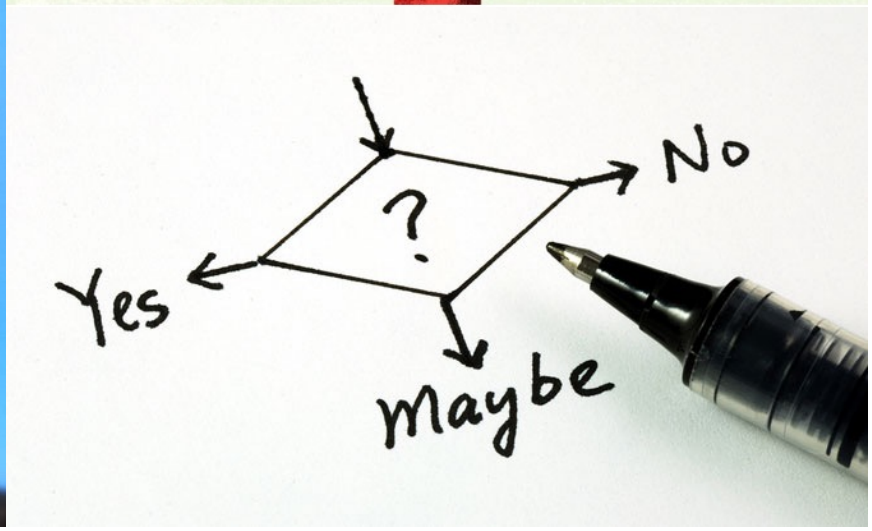
The case of Statistical Hypothesis Testing:

A statistical framework that suggests how an uninteresting value (value assumed under the null hypothesis) is likely (large p-values) or unlikely (small p-values).

But if the value of no interest is unlikely, IT does NOT indicates which other values are likely (any other value differing from the one assumed under the null hypothesis).

Generating evidence-based conclusions without complete biological knowledge

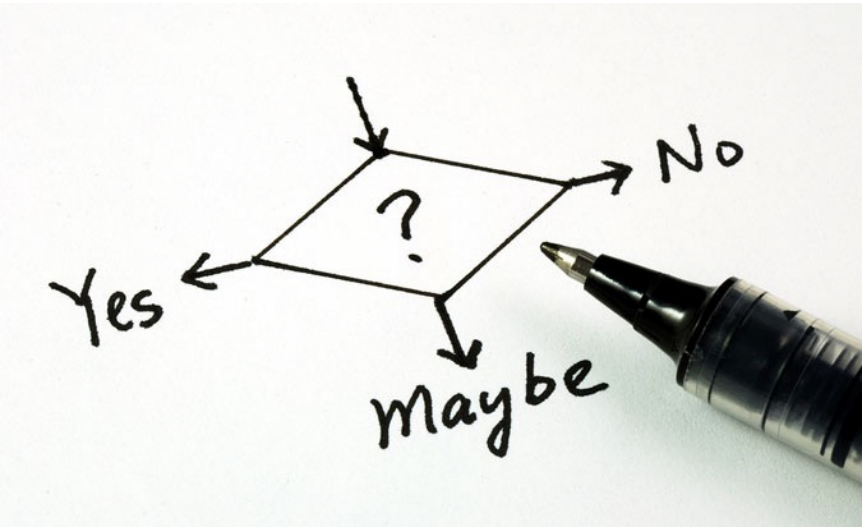




Intuitive demonstration underlying statistical hypothesis testing

A road map to UNDERSTAND how *evidence-based* decisions / conclusions can be made without complete knowledge

Statistics - the science that assists in informing decision making without complete knowledge!!



“intuitive” Demonstration: statistical hypothesis testing

Statistical Hypothesis Testing is an intimate stranger!!

Demonstration involves showing by reason or proof, explaining or making clear by use of examples or experiments.

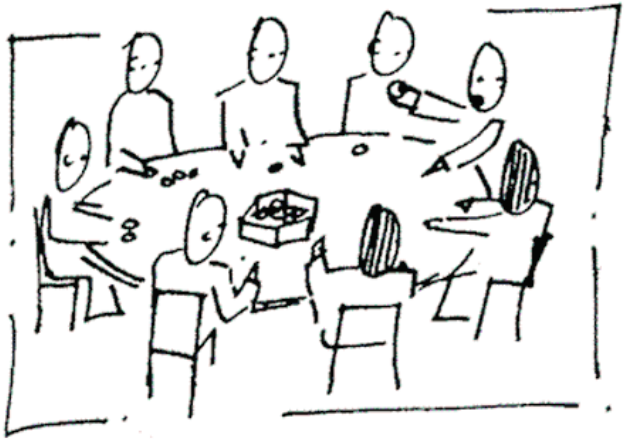
Put more simply, demonstration means 'to clearly show’! (hopefully)

Class-participation demonstration: Statistical Hypothesis Testing

I. INTRODUCTION

Classroom demonstrations, a standard component of science courses in schools and universities, are commonly believed to help students learn science and to stimulate student interest. There is little doubt that well-performed demonstrations achieve the latter objective; one study found that demonstrations are among students' favorite elements of introductory undergraduate physics courses.[1] However, research on student learning from demonstrations suggests that traditional demonstrations may not effectively help students grasp the underlying scientific concepts or recognize and correct scientific misconceptions they may have.[2–4]

Science education research shows that most students learn more from instruction that actively engages them rather than from traditional methods in which they are passive spectators.[5] A number of approaches to instruction that are designed to engage students more actively have therefore been developed. Many of the most successful approaches consist of a set of carefully refined student activities designed to address research-identified student difficulties with the material. These approaches specify both the instructional methods and the content to be covered.[6] For example, Sokoloff and Thornton's Interactive Lecture Demonstrations (ILD)[7] replace 1 h of lecture per week with a sequence of five to seven highly interactive, demonstration-based activities.[7]



Crouch et al. 2004. American Journal of Physics **72**, 835; https://doi.org/10.1119/1.1707018

A very simple statistical hypothesis testing example

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

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

Translating the research question into a statistical question:

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

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

What is a research 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 present handedness”.

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

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 by the data at at hands (and can be potentially refuted by other 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 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 theory is a well-substantiated explanation for a natural phenomenon. And a law (gravity) is an observation (objects fall towards the ground).

Tackling research hypotheses using the framework of statistical hypothesis testing

The **statistical hypothesis framework** (most often involving statistical tests) is a quantitative method of statistical inference that allows to generate evidence for or against a research hypothesis (often based on a question of interest).

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

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).

Let’s take a break - 2 minutes

Remember the two possible statistical hypotheses:

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

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

Tackling research hypotheses using the framework of statistical hypothesis testing

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

CONFUSING: IT ONLY GENERATES SUPPORT *AGAINST* THE STATISTICAL NULL HYPOTHESIS (NOT FOR). It also doesn't generate support for (or against) the alternative hypothesis.

But by building support AGAINST a statistical null hypothesis, one builds support FOR the research (alternative) hypothesis (i.e., other animals do present handedness).

Remember: a small p-value makes us reject the null hypothesis of equal proportion of limb usage and therefore provides support to the research hypothesis of handedness.

A very simple statistical hypothesis testing example

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

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

Translating the research question into a statistical question:

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

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

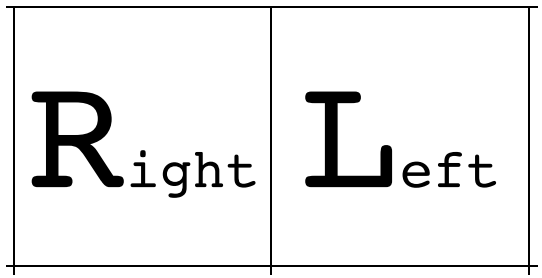
The intuition behind the framework of statistical hypothesis testing

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



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 and support the alternative hypothesis or not!

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

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

*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.

```
> 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 Model for H_0

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

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

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

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

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

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

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

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

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

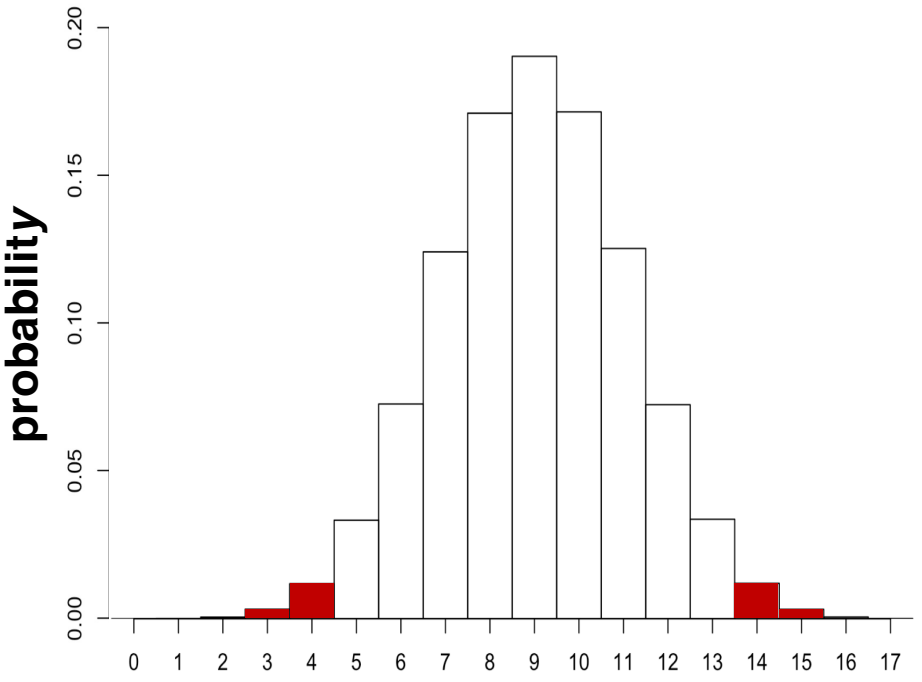
If we had drawn 1000000 samples from the population assumed under H_0, 166900 would had been 8 right-handed and 10 left-handed.

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

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

equal or smaller sum [P]=0.0155

equal or greater sum [P]=0.0155

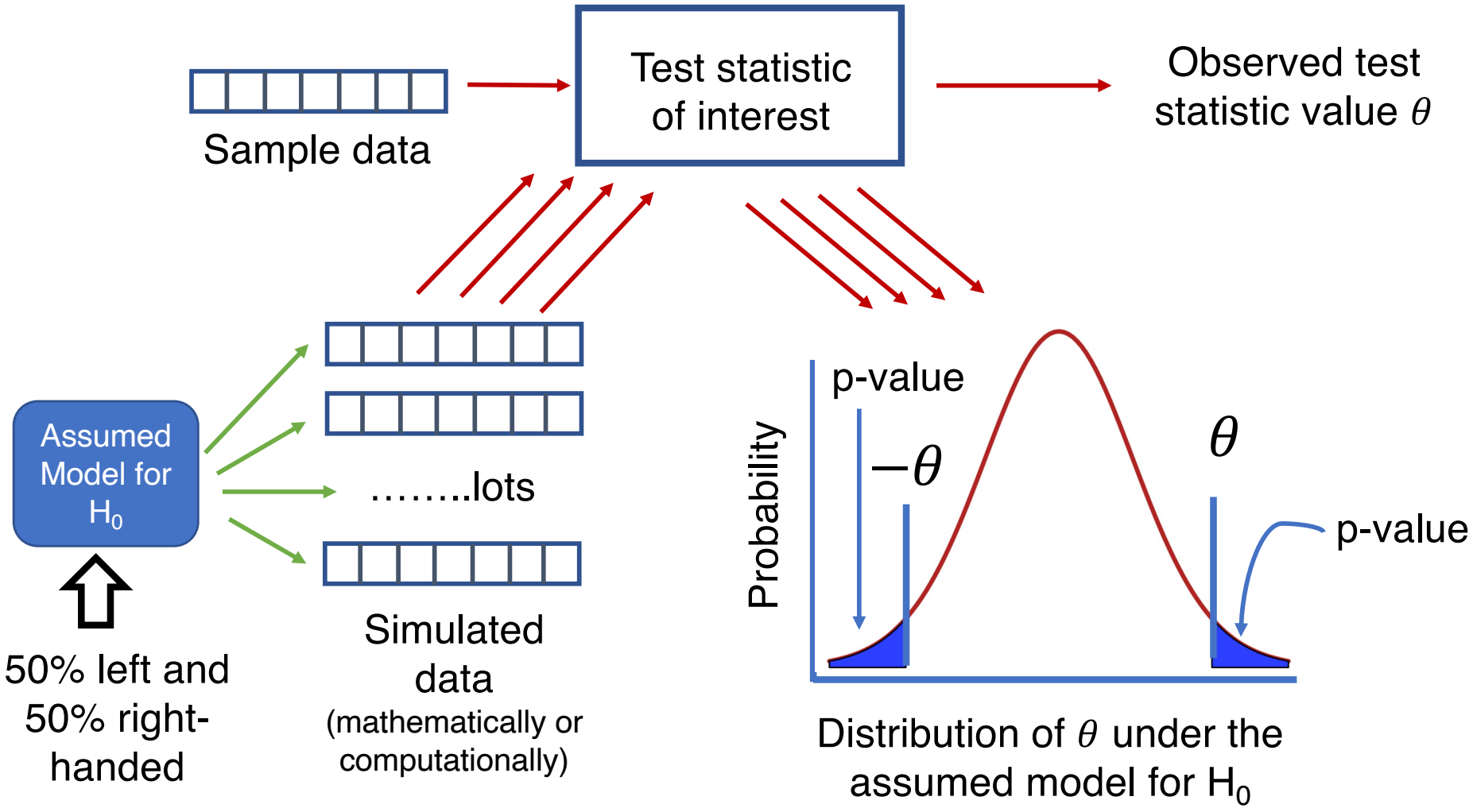


Pr[14 or more right-handed toads] = Pr[14] + P[15] + P[16] + P[17] + P[18] = 0.0155 x 2 (symmetric distribution) = **0.031**

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

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

The “machinery” behind the framework of statistical hypothesis testing



θ = #R-handed toads

$-\theta$ = #L-handed toads

= #L-handed toads

Adapted from https://moderndive.com/

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

How do we use the sampling distribution from the theoretical population where proportion is 50%/50% to generate evidence *of handedness in toads?*

The probability of generating a value equal to the sample data or values more unusual (smaller or greater) than the one for the sample data in the sampling distribution based on assuming that the theoretical population is true was:

P = 0.031

It is either ***likely*** or ***unlikely*** that the researcher would have collected the evidence (i.e., observed value) given the initial assumption (theoretical population with equal number of individuals with right- and left-handed).

The ”levels” of likely or unlikely is estimated by the contrast between the observed value and the null distribution (generated from assuming a certain theoretical population and its associated parameter).

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 of support for or against the research hypothesis (e.g., generate evidence for or against handedness in toads).

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

That's not to say that handedness is true OR false but rather that we have strong evidence to say that handedness (i.e., 50%/%50) is unlikely.

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

P = 0.031

AGAIN - VERY IMPORTANT but “confusing”:

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).

AGAIN: hypotheses cannot be proven right or wrong from the data. Hypotheses can be said to be either refuted or supported by the data generated.

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

The p-value is a measure of the evidence against the null hypothesis, calculated from the sample data.

It represents the probability of obtaining a test statistic as extreme or more extreme than the one observed, under the assumption that the null hypothesis is true.

A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis and in favor of the alternative hypothesis.

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

The p-value is a measure of the evidence against the null hypothesis, calculated from the sample data.

It represents the probability of obtaining a test statistic as extreme or more extreme than the one observed, under the assumption that the null hypothesis is true.

A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis and in favor of the alternative hypothesis.

The p-value is **a measure of consistency** between the sample data and the theoretical hypothesis assumed when stating the parameter for a theoretical population of no interest (null hypothesis, e.g., toads have equal number of individuals right and left-handed)

The process of statistical hypothesis testing: critical details

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

A null hypothesis is specific (generally assuming one value; e.g., 50%/50%); an alternate hypothesis is not (any proportion different from 50%/50%).

As such, by rejecting the null hypothesis we learned something new!

Let’s take a break - 2 minutes

Statistical hypothesis testing versus estimation

Statistical hypothesis testing, like estimations, uses sample data to make inferences about the population from which the sample was drawn.

Estimations put bounds on the value of a population parameter, whereas hypothesis testing asks only whether the parameter differs from a specific "null" expectation (also called theoretical population or initial assumption).

Statistical hypothesis testing versus estimation

Statistical hypothesis testing, like estimations, uses sample data to make inferences about the population from which the sample was drawn.

Estimations put bounds on the value of a population parameter, whereas hypothesis testing asks only whether the parameter differs from a specific "null" expectation (also called theoretical population or initial assumption).

Estimation asks - What is the value of the parameter? (e.g., how many frogs are right handed?).

Hypothesis testing asks a yes or no question (e.g., "Do right-handed and left-handed toads occur with H_0: equal frequency in the toad population, or H_A is one type more frequent than the other?")

Statistical hypothesis testing versus estimation

Statistical hypothesis testing, like estimations, uses sample data to make inferences about the population from which the sample was drawn.

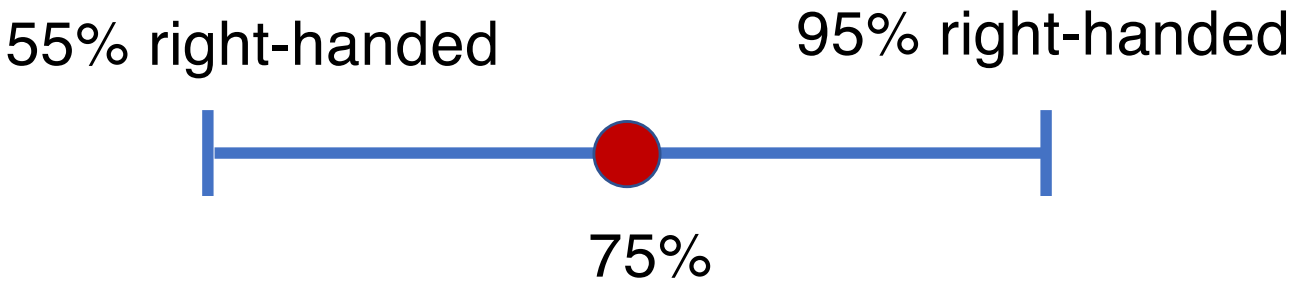
Estimations put bounds on the value of a population parameter, whereas hypothesis testing asks only whether the parameter differs from a specific "null" expectation (also called theoretical population or initial assumption).

Estimation asks - What is the value of the parameter? (e.g., how many frogs are right handed?).

Hypothesis testing asks a yes or no question (e.g., "Do right-handed and left-handed toads occur with H_0: equal frequency in the toad population, or H_A is one type more frequent than the other?")

Statistical hypothesis testing *versus* estimation

Often we don't have enough large sample sizes to produce good estimates and intervals that allows us to be confident on what the population values could be. But we can generate a different type of question as to whether we can generate evidence to reject the hypothesis that the population is 50%/50%.



Estimation thinking: The 95% confidence interval above is not very informative about the true population proportion. It is likely that is between 55% and 95% of right-handed (informative but not with great accuracy).

Statistical hypothesis thinking: But we are confident that is not likely to be in equal proportion (50% right- and 50% left-handed).

If we can't state what it is likely, at least we can try to state what is likely not!

Estimation versus Statistical hypothesis testing

Do the conclusions from the two statements below differ?
How?
Which one you prefer?

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

OR

The proportion is different from 0.5

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

We may not be able to generate a good estimate (e.g., large confidence intervals) of what the value of left/right handed proportion (e.g., sample size of 18 is perhaps too small to generate good estimates).

But we can do something different: generate evidence for or against that the common toad use their limbs in equal proportion.

Here we can quantify how unusual the observed sample data (4/18 left or 14/18 right) are in regards to the assumption that they are 50%/50% (i.e., contrast the observed number of right-handed against a sampling distribution of number of right-handed toads for a theoretical statistical population where the proportion is truly 50%)

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?

Statistical hypothesis testing

Instead of estimating what value for the parameter is likely [within an interval], under a statistical hypothesis framework we estimate how unlikely a particular parameter value of no interest is.

If we can't state what it is likely, at least we can try to state what is likely not!

In the toad study. Perhaps we can't state with certainty that the true proportion is likely between a narrow confidence interval [e.g., 75%-79%; which may require a very large sample size] but we can state that it is not likely 50%. That's the attempt of statistically hypothesis testing; with the hope that this would suffice as evidence for handedness [many philosophical and probabilistic discussions on this topic though; more later].