Tutorials are not just exercises or applications.

They are intentionally designed to complement and strengthen the material covered in lectures, helping you develop a deeper understanding of key concepts.

Tutorial 2: Statistical Hypothesis Testing

The principles of statistical hypothesis testing: generating evidence-based conclusions without complete biological knowledge!

Statistics does not remove ***uncertainty*** - it formalizes it and communicates ***confidence***.

In statistics we quantify ***uncertainty*** to justify ***confidence***.

Understanding these two ideas is essential for developing lasting statistical reasoning.

They form the foundation of statistical inference and hypothesis testing.

Schematic summary of lecture 2

RESEARCH QUESTION

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

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?

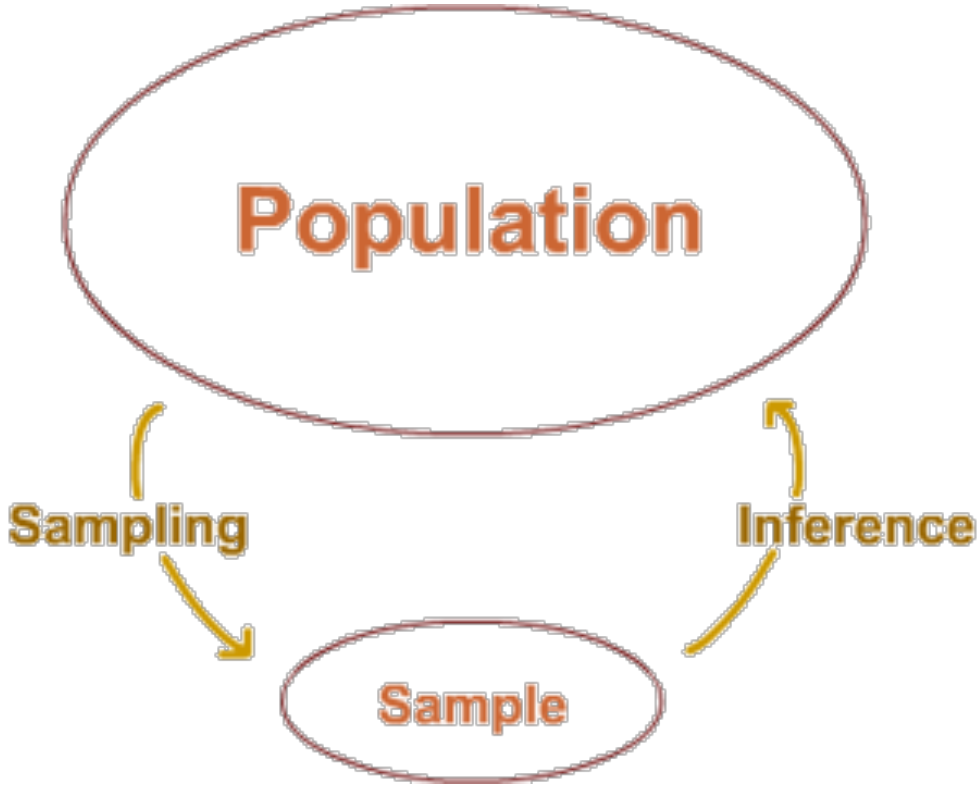
STATISTICAL INFERENTIAL APPROACH

→ **Statistical hypothesis testing.**

→ **Parameter estimation and confidence interval.**

Key statistical element → **Estimating *uncertainty* due to sampling variation**

Statistical inference: the process of quantifying *uncertainty* due to sampling variation in order to draw conclusions from samples (despite incomplete knowledge) about entire populations.



Sampling variation drives statistical uncertainty.
Observed proportions differ from the true population value simply because samples are random. Some samples will be closer to the true value than others, meaning the level of uncertainty varies across samples.

Remember: random sampling minimizes sampling error, uncertainty & inferential bias (i.e., how close or far sample values for the statistic of interest are from the true population value for that statistic)

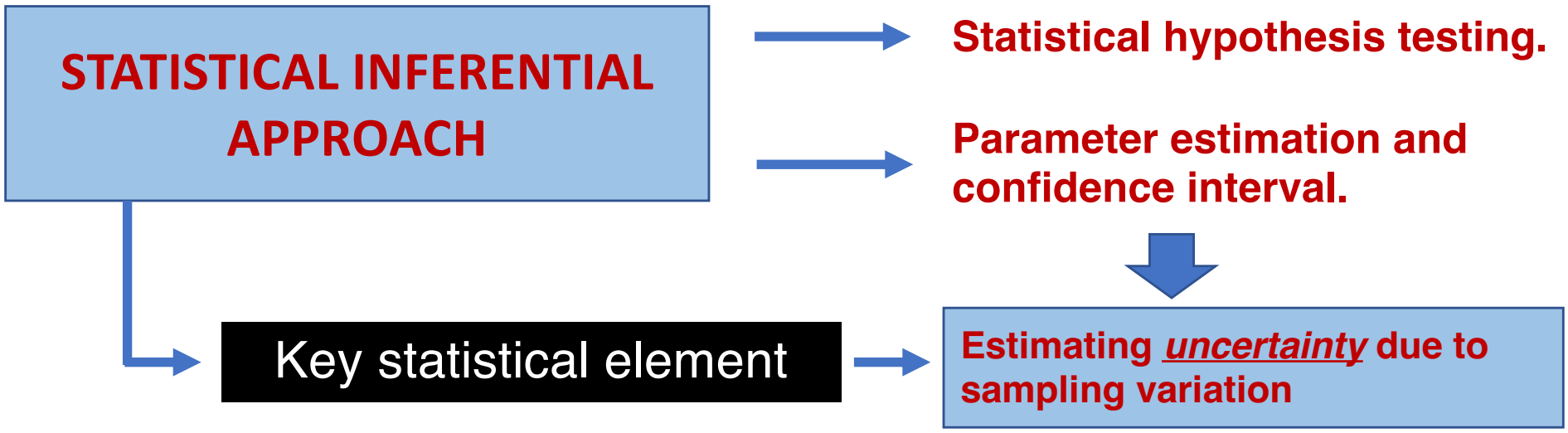
The common requirement for statistical inference is that data come from a **random sample**. A random sample is one that fulfills two criteria:

1) Every observational unit in the population (e.g., individual tree) have an **equal chance** of being included in the sample.

2) The selection of observational units in the population (e.g., individual tree) must be **independent**, i.e., the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

Samples are biased when some observational units of the intended population have lower or higher probabilities to be sampled.

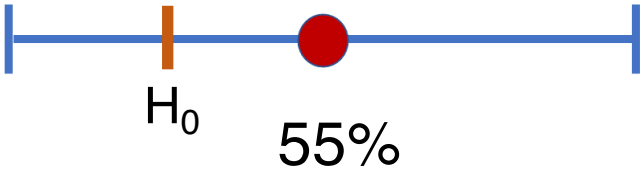
Statistical inference: We use sampling theory - either computationally based or calculus based (more common) - to derive the sampling distribution needed to quantify uncertainty around a sample estimate of interest (e.g., the proportion of right- and left-legged toads).



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

45% right-handed

65% right-handed



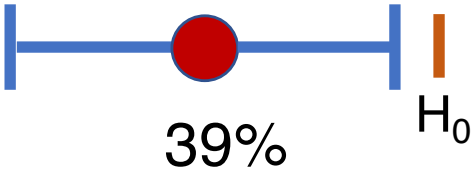
"Quantitative" statement

Don't reject H_0; $p > 0.05$

"Qualitative" statement

30% RH

48% RH



"Quantitative" statement

Reject H_0; $p < 0.05$

"Qualitative" statement

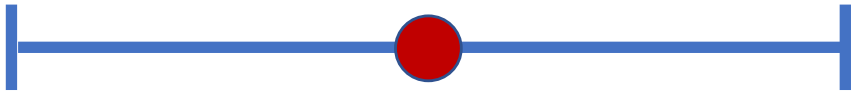
Statistical inferential process: The role of sampling theory in parameter estimate and estimating confidence intervals

"The purpose of statistical inference is to develop theory and methods to make inference on the unknown parameters based on observed data"
(Hong, 2017)

we can state that we have some confidence (say 95%) that the true parameter of interest (say number of right-legged frogs is between two values

95% confidence interval for the proportion of right-legged toads

52.4% right-legged 93.6% right-legged



77.8%

(14 right- and 4 left-legged toads, i.e., 14/18 = 77.8%)

A large confidence interval (e.g., 95% or 99%) provides a most plausible range for a parameter of interest (true population value). Values lying within the interval are most plausible, whereas values outside are less plausible, based on the sample data alone.

Statistical inference

Operationally simple, but conceptually essential. Understanding the underlying theory is what makes statistics intuitive rather than mechanical.

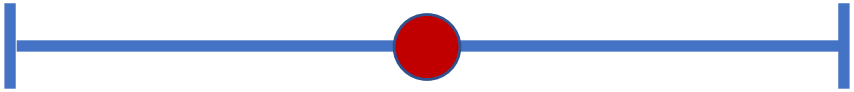
```
> binom.test(14,18,0.5,alternative="two.sided")

    Exact binomial test

data:  14 and 18
number of successes = 14, number of trials = 18, p-value = 0.03088
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5236272 0.9359080
sample estimates:
probability of success
             0.7777778
```

95% confidence interval for the proportion of right-legged toads

52.4% right-legged — 93.6% right-legged



77.8%

(14 right- and 4 left-legged toads, i.e., 14/18 = 77.8%)

Statistical inference: The role of sampling theory

If sampling is random and the model assumptions (e.g., normal distribution) are satisfied, two key statistical results follow:

[1] Across a very large number of repeated samples from the same population, 95% of the confidence intervals constructed in this way will contain the true population value.

[2] Because of this long-run (repeated sample) property, we can say that a confidence interval built from a single random sample is associated with 95% confidence of containing the true population value.

Statistical inference: The role of sampling theory

Why not a 100% confidence interval?

Because 100% certainty requires intervals so wide that they contain all possible values - and therefore tell us nothing useful.

In many cases, this would mean an interval from $-\infty$ to $+\infty$; for proportions, from 0% to 100%.

Statistics trades certainty for information: Narrow intervals (e.g., 95%) allow learning; 100% intervals guarantee coverage but eliminate inference.

As we accept slightly less certainty (e.g., 95% instead of 100%), the interval becomes narrower, allowing us to make more precise statements about where the true value is likely to lie. However, this precision comes with a small risk of being wrong.

In statistics, we deliberately accept a limited risk of error in order to gain informative, precise estimates. This balance—less certainty for more precision—is what makes statistical inference possible.

Statistical inference: The role of sampling theory

Why not a 100% confidence interval?

Absolute certainty requires intervals so wide that they convey no information.

Greater certainty requires wider intervals; greater precision requires accepting some uncertainty.

Precision (confidence interval width) can be increased by increasing sample size (among other features).

Imagine a weather forecast saying: “Tomorrow’s temperature will be between –50 °C and +50 °C.”

That forecast is certain, but completely useless.

A forecast like “between 14 °C and 18 °C” is far more informative, but it accepts that the true value might occasionally fall outside that range.

Inference is not about being certain—it is about managing uncertainty. Narrow intervals support decisions, provided we plan for the possibility of being wrong.

The intuition behind the frequentist framework of statistical hypothesis testing

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

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

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

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

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

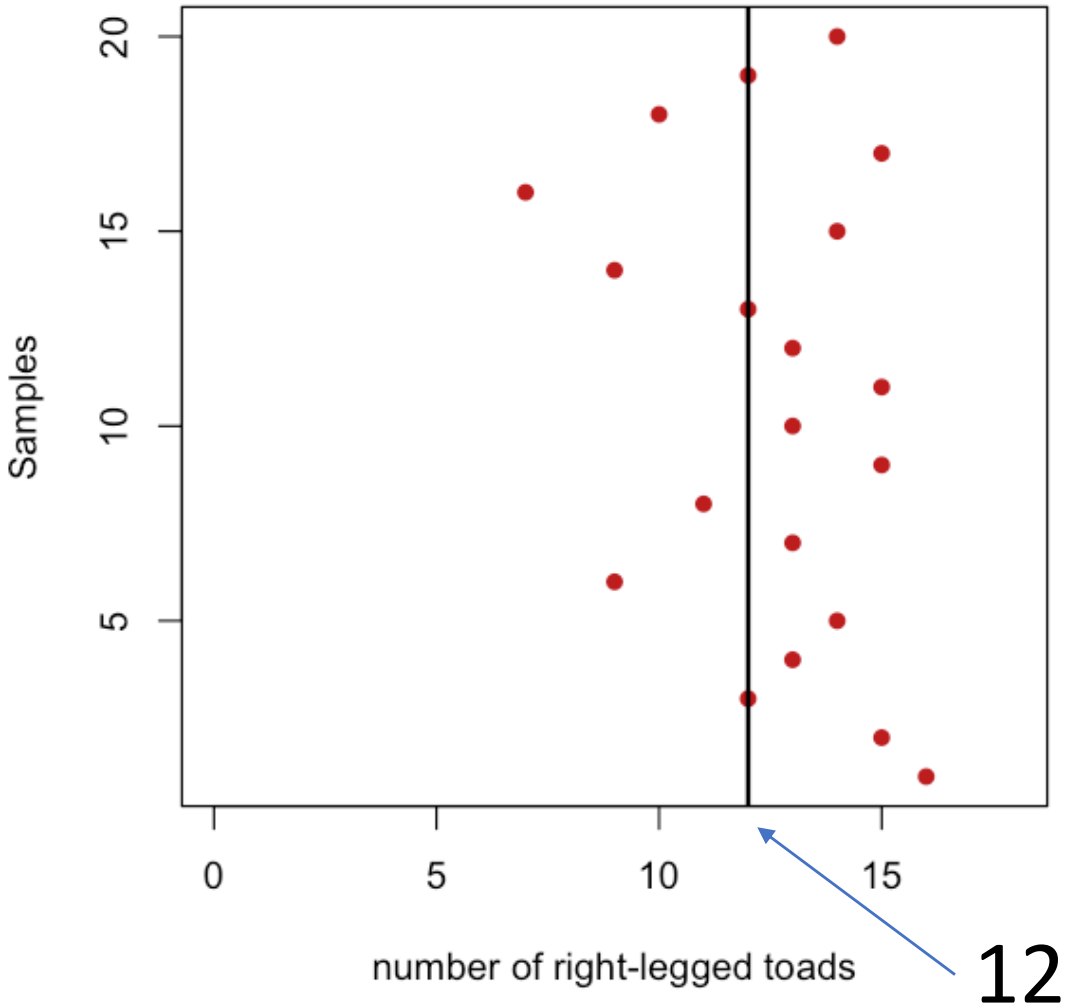
*Resampling (sampling with replacement) is important because it ensures that each selection of an observational unit (e.g., a piece of paper) is independent of the others. In other words, drawing one unit (L or R) does not affect the probability of subsequent draws.

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

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

Let’s imagine 20 samples of 18 toads, each using the “bag approach”



Assume a toad population where (for the sake of demonstration) 66.7% of observational units (toads) are right-legged and 33.3% left-legged. Assume this population to be mathematically infinite.

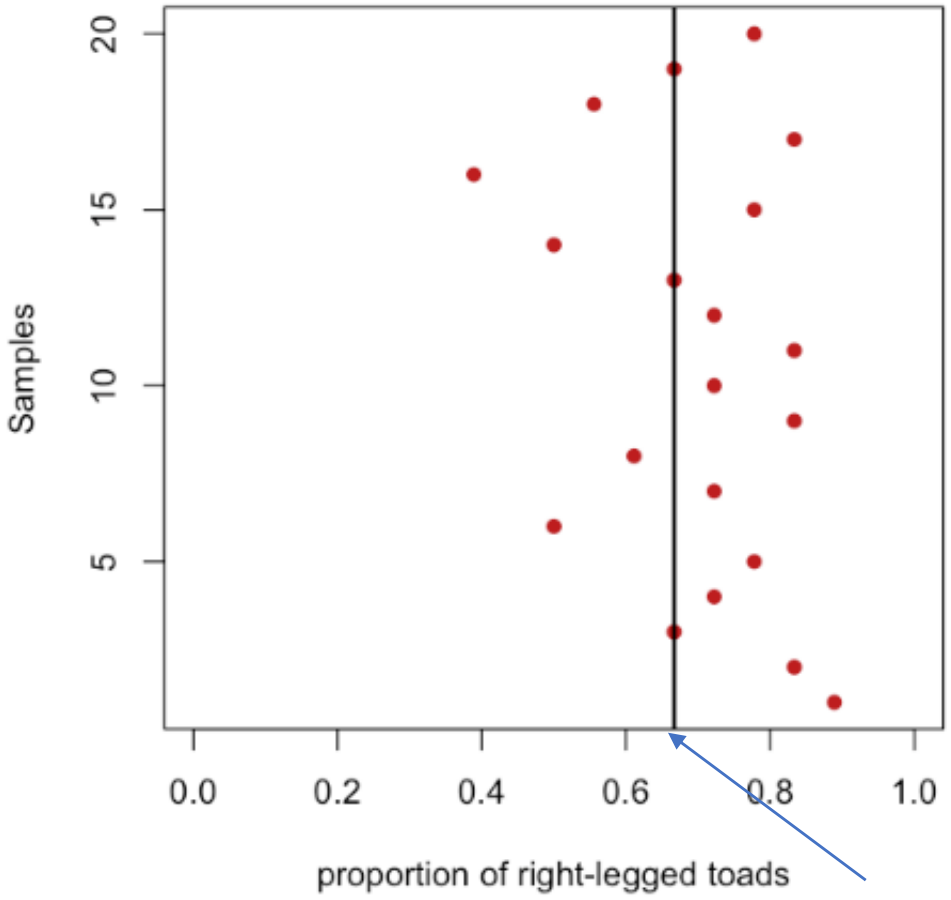
The vertical line line is the expected number of right-handed toads across all possible infinite samples of 18 toads (i.e., 12 toads are expected in average to be right-handed, i.e., 0.66667*18=12)

12 (66.7%)

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

More traditionally we use the proportion instead of the numbers
(easier to generalize mathematically)



Assume a toad population where (for the sake of demonstration) 66.7% of observational units (toads) are right-legged and 33.3% left-legged. Assume this population to be mathematically infinite.

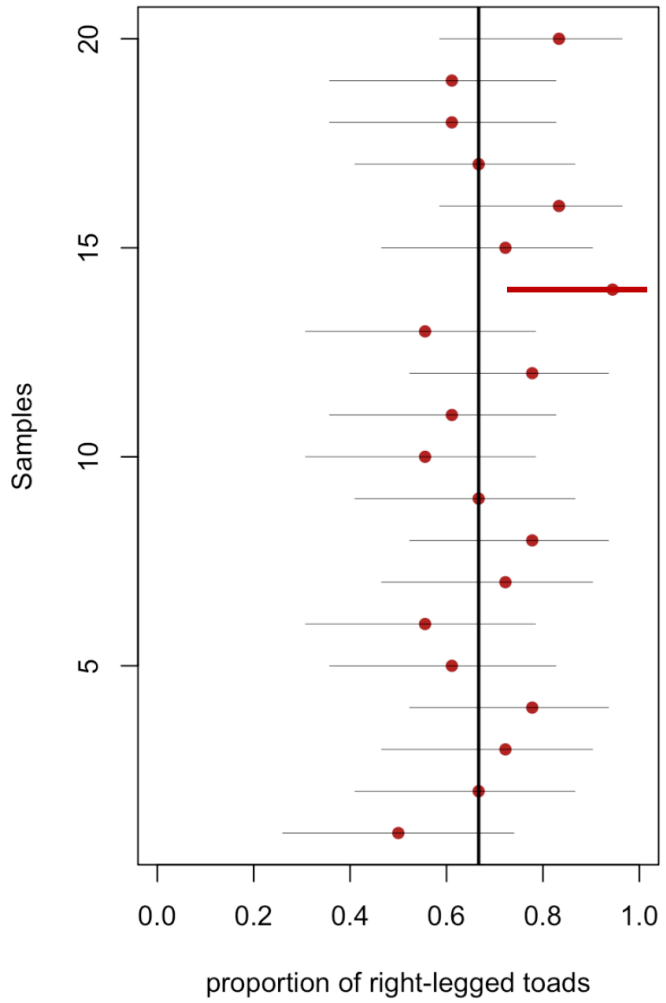
The vertical line line is the expected number of right-handed toads across all possible infinite samples of 18 toads (i.e., 12 toads are expected in average to be right-handed, i.e., 0.66667*18=12)

0.667

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

More traditionally we use the proportion instead of the numbers
(easier to generalize mathematically)



It is expected that 1 sample confidence interval over 20 does not include the true population proportion (i.e., 95%). This interval is in red.

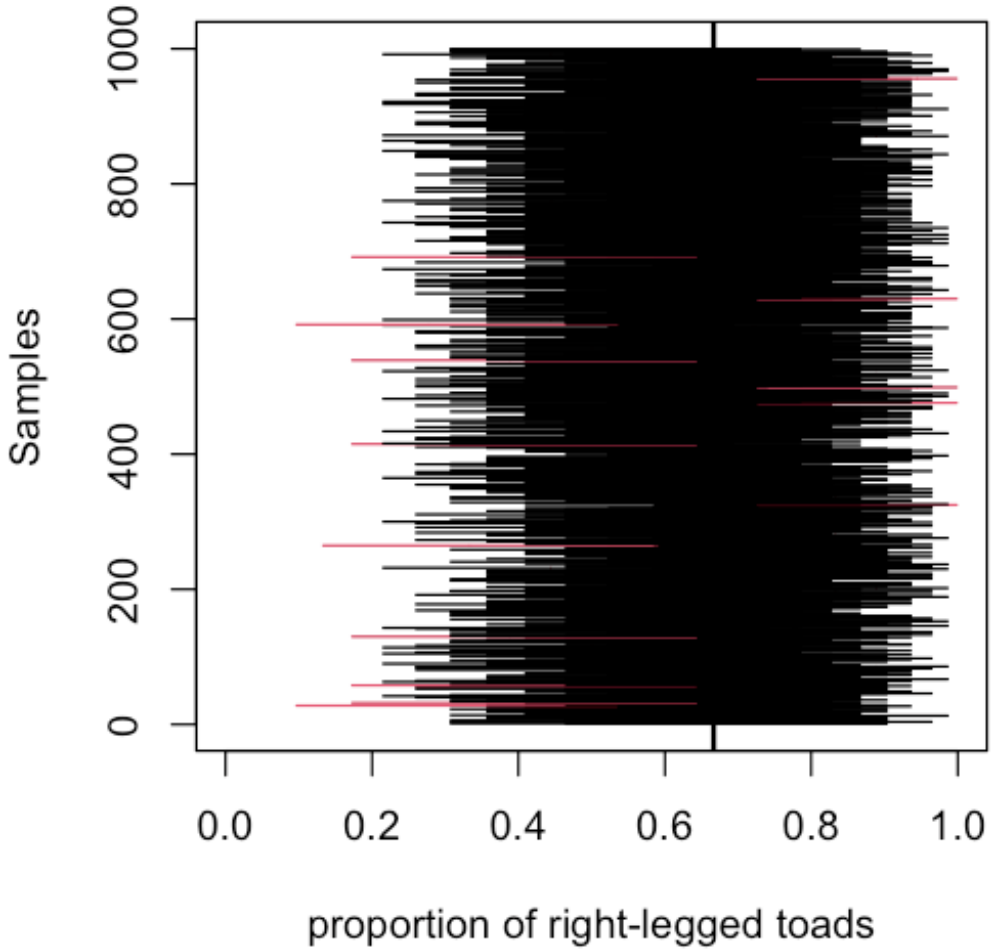
There are different ways to estimate the confidence interval for proportions (here we used the *exact* method which can produce non-symmetric intervals and tend to be wider than the asymptotic estimation).

For now, the estimation method is not important; what’s important is the rationale. We will see some more details later in this lecture.

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

Now 1000 estimates for the confidence intervals. The point here is that we can produce a really large number of intervals & 95% of them will contain the true population value



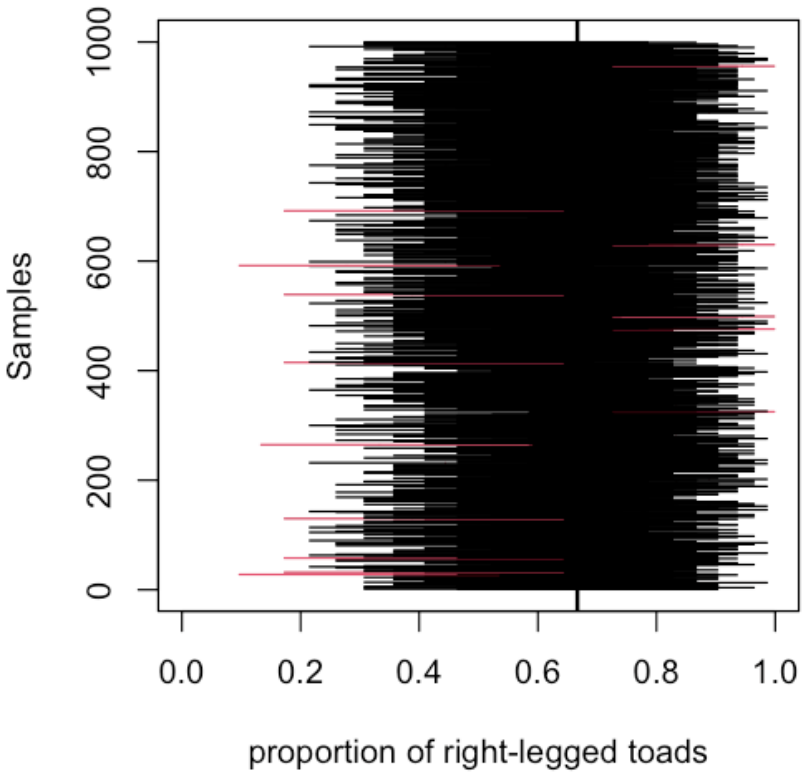
intervals not containing the true parameter are plotted in red (i.e., 5% of the intervals).

The role of sample size in increasing confidence and decreasing uncertainty

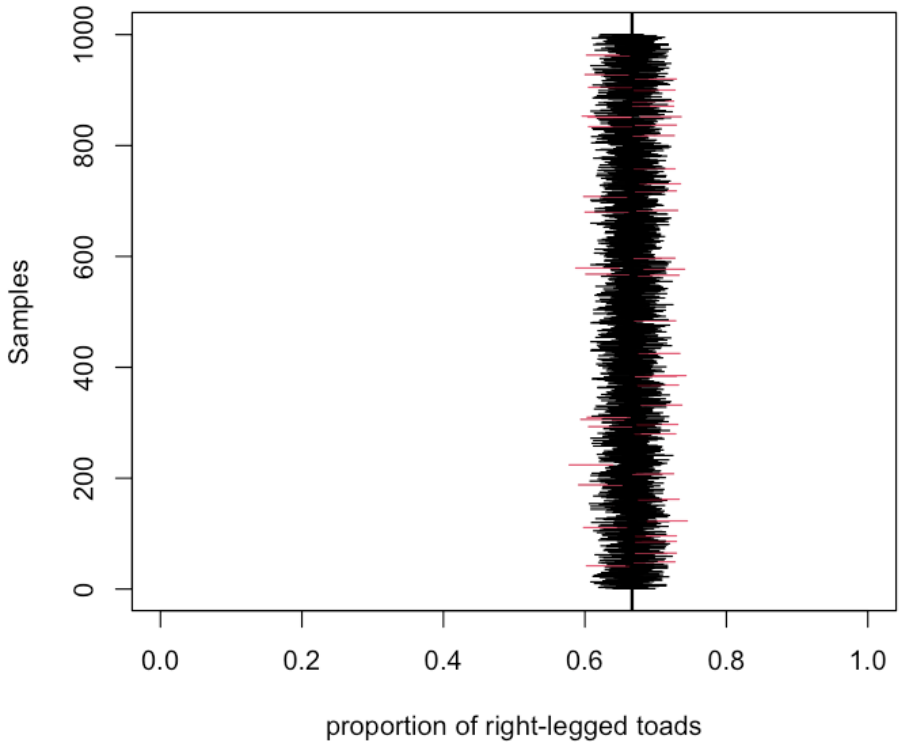
One way to improve estimation & reduce uncertainty is to increase sample size.

Remember though that increasing sample may be extremely timely and financially consuming, and even ethically irresponsible (e.g., collecting and manipulating too many individuals that can put biological populations and species at risk).

Sample size = 18



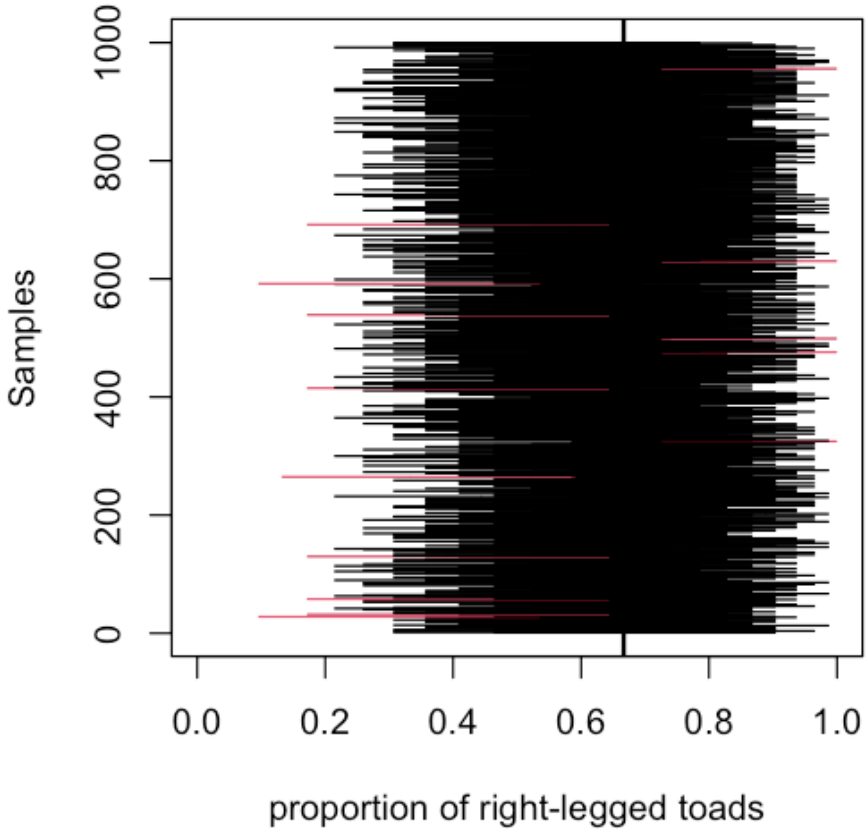
Sample size = 1000



Very important & often confusing!

For any given sample confidence interval, we can state that “we are 95% confident that the true population mean lies between the lower and upper limits of the interval”.

BUT, we cannot say that “there is a 95% probability that the true population mean lies within the confidence interval”.



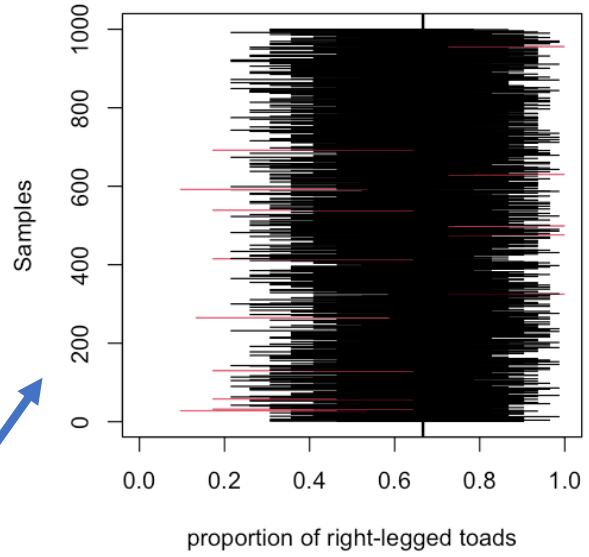
Very important & often confusing!

We cannot say that "there is a 95% probability that the true population mean lies within the confidence interval".

The population parameter (e.g., the true proportion) is a fixed but unknown value; but it does not change from sample to sample.

What does vary is the confidence interval, because it is built from a random sample. Across many repeated samples, some intervals will capture the true value, and some will not.

Once a specific interval is calculated from a single sample, the interval is fixed. At that point, the true parameter is either inside it or outside it, i.e., there is no randomness left.



Confidence intervals are not well grasped by a large number of users of statistics!

CANADIAN JOURNAL OF SCIENCE, MATHEMATICS
AND TECHNOLOGY EDUCATION, *14*(1), 23–34, 2014
Published with license by Taylor & Francis
ISSN: 1492-6156 print / 1942-4051 online
DOI: 10.1080/14926156.2014.874615

Routledge
Taylor & Francis Group

Confidence Trick: The Interpretation of Confidence Intervals

Colin Foster
School of Education, University of Nottingham, Nottingham, United Kingdom

Often stated by students and practitioners of statistics: "there is a 95% probability that the true population mean lies within the confidence interval". We can't state that; either the parameter is within the interval or not! So, no probability attached to this condition.

Let’s take a break - 1 minute

Confidence intervals are not well grasped by many (probably most) users of statistics! RECAP

Confidence intervals is a concept based on sampling theory.

Here, sampling theory relates to repeated sampling making certain assumptions about the statistical population.

We use the principle of repeated sampling to model the expectations of sampling variation. Under repeated sampling, if we were to estimate a confidence interval for each sample, 95% of them would contain the true population parameter.

As such, we can be confidence that one single sample confidence interval (i.e., we usually only have one sample) will most *likely* (but without a precise probability) contain the true population value.

A large confidence interval (e.g., 95% or 99%) provides a most *plausible* range for a parameter (true population value). Values lying within the interval are most *plausible*, whereas values outside are less plausible, based on the sample data alone.

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the "bag approach"

Let's go back to the original results of Bisazza et al. (1996) in which 14 toads were 14 right-legged and 4 were left-legged (i.e., 77.8% right- and 22.2% left-legged). Let's estimate its confidence interval based on the "bag approach"

BUT FIRST REMEMBER:

[1] 95% of the "infinite" (really large value) confidence intervals that could be built based on each possible sample from a given population will contain the true population value.

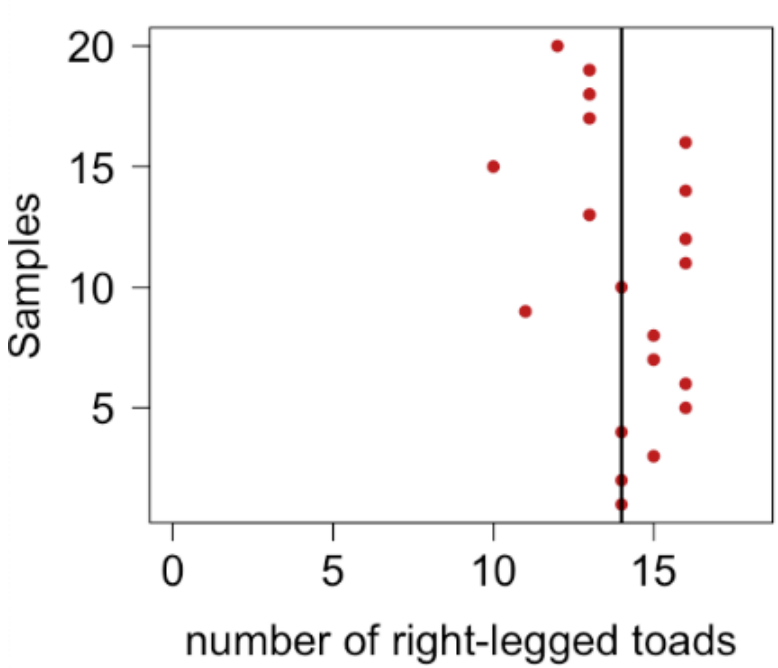
[2] Because of statement 1, we can be then 95% confident that the interval estimated based on a single sample contains the true population value of the population from where that sample was taken!

Confidence is about trust in the method, not a probability attached to the parameter. Once the interval is computed, the true value is either inside it or not.

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

20 random samples

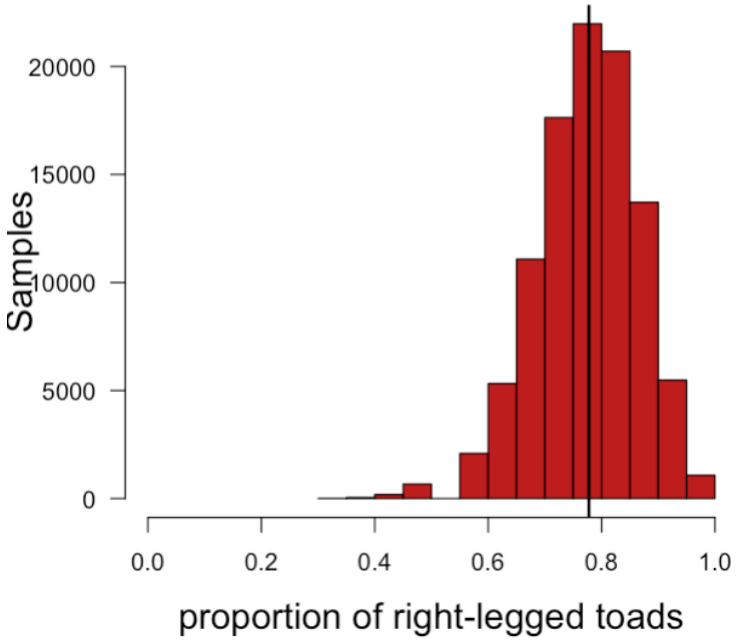


Assume that 77.7% (14/18 individuals) of toad population is right-legged and 22.2%(4/18 individuals) left-handed. Assume this population to be mathematically infinite.

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

100000 random samples



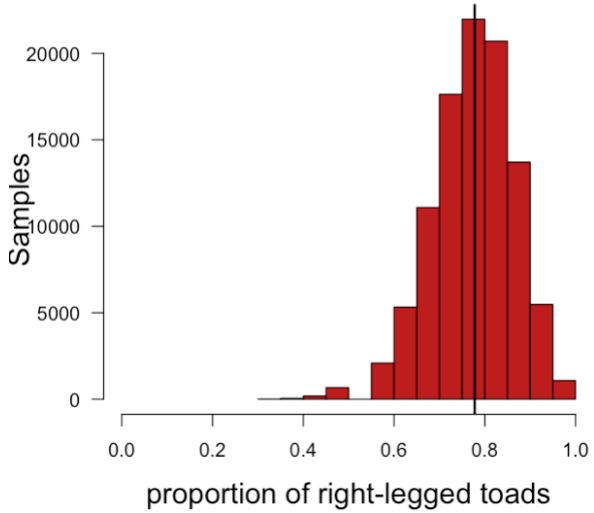
in each of the 100000 samples

Assume that 77.7% (14/18 individuals) of toad population is right-legged and 22.2%(4/18 individuals) left-handed. Assume this population to be mathematically infinite.

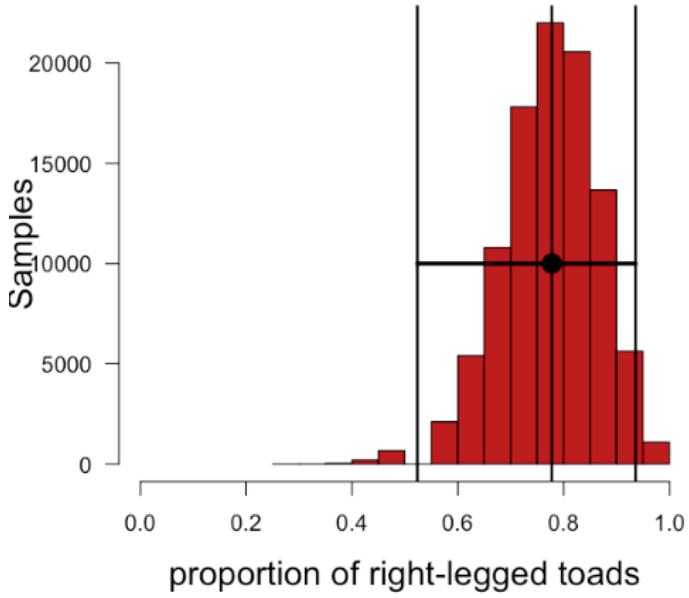
How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

100000 random samples



Assume that 77.7% (14/18 individuals) of toad population is right-legged and 22.2%(4/18 individuals) left-handed. Assume this population to be mathematically infinite.

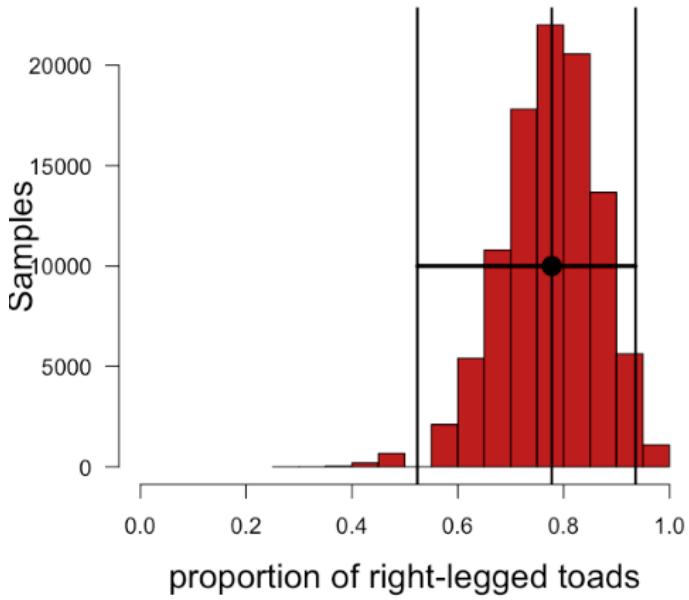


Calculate the 2.5% percentile and 97.5% percentile as the confidence Interval.

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

100000 random samples



Calculate the 2.5% percentile and 97.5% percentile as the confidence Interval.

Assume that 77.7% (14/18 individuals) of toad population is right-legged and 22.2%(4/18 individuals) left-handed. Assume this population to be mathematically infinite.

In this way, we can be then 95% confident that the interval estimated based on a single sample contains the true population value of the population from where that sample was taken!

How are confidence intervals computationally derived?

Building knowledge about sampling theory using the “bag approach”

For binomial distributions, i.e., distributions that have two possible outcomes (here right- and left-legged individuals), there are a few different ways to estimate confidence intervals – and they differ somewhat, particularly when sample sizes are smaller.

```
> binconf(14,18,method = "all")
           PointEst     Lower     Upper
Exact      0.7777778 0.5236272 0.9359080
Wilson     0.7777778 0.5478542 0.9099907
Asymptotic 0.7777778 0.5857194 0.9698362

> binconf(778,1000,method = "all")
           PointEst     Lower     Upper
Exact         0.778 0.7509431 0.8034091
Wilson        0.778 0.7512054 0.8026670
Asymptotic    0.778 0.7522419 0.8037581
```

This may happen because different methods are used to approximate the distribution of a variable.

Some confidence intervals for some distributions can be modelled exactly (e.g., normal) whereas others only approximated (e.g., binomial).

The exact method is better but can be computationally intense (impractical for large sample sizes) compared to the asymptotic (approximation via a normal approximation); the Wilson is better for small sample sizes than the asymptotic.

PREVIOUS SLIDE INTO "SOLID" WORDS

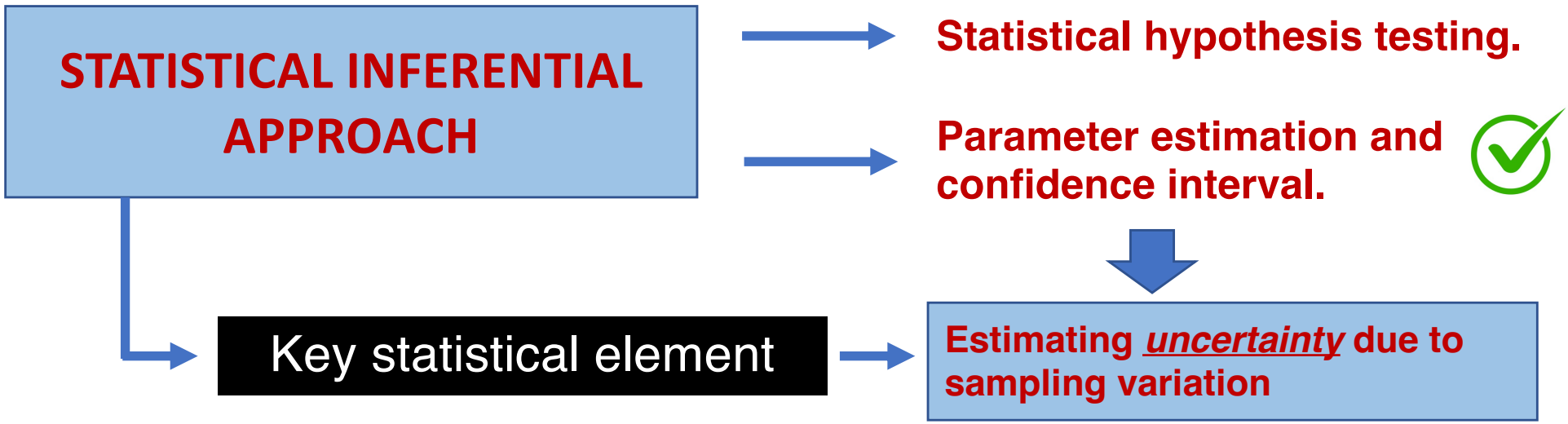
Statistical inferential process: The role of sampling theory?

Estimating uncertainty from sample-based values (information) that allows generalization to entire populations

The basic idea of statistical inference is to assume that the observed data (e.g., 77.8% of frogs were right-legged) is generated from a probability distribution (all possible sample values for the population of interest; e.g., frogs) which is modelled by a function in the form of a probability distribution (e.g., Exact, Wilson, Asymptotic).

Sampling theory is applied to predict sampling uncertainty from sample estimates that is then used to estimate uncertainty. The prediction is made by making assumptions about certain aspects of the sample or populations to estimate the sampling distribution for the value of interest (e.g., number of right- and left-legged toads).

Statistical inferential process: we use sampling theory to determine the sampling distribution required to estimate uncertainty around a sample value of interest (e.g., number of right- and left-legged toads).



Let’s take a break – 1 minute

Frequentist Statistical Inference: An Intimate Stranger - Routine in use, mysterious in meaning!

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: BUT 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 research hypothesis.

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.

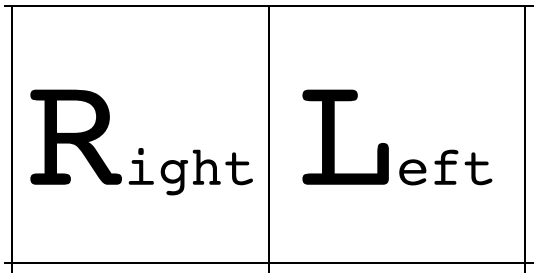
The intuition behind the framework of statistical hypothesis testing

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

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

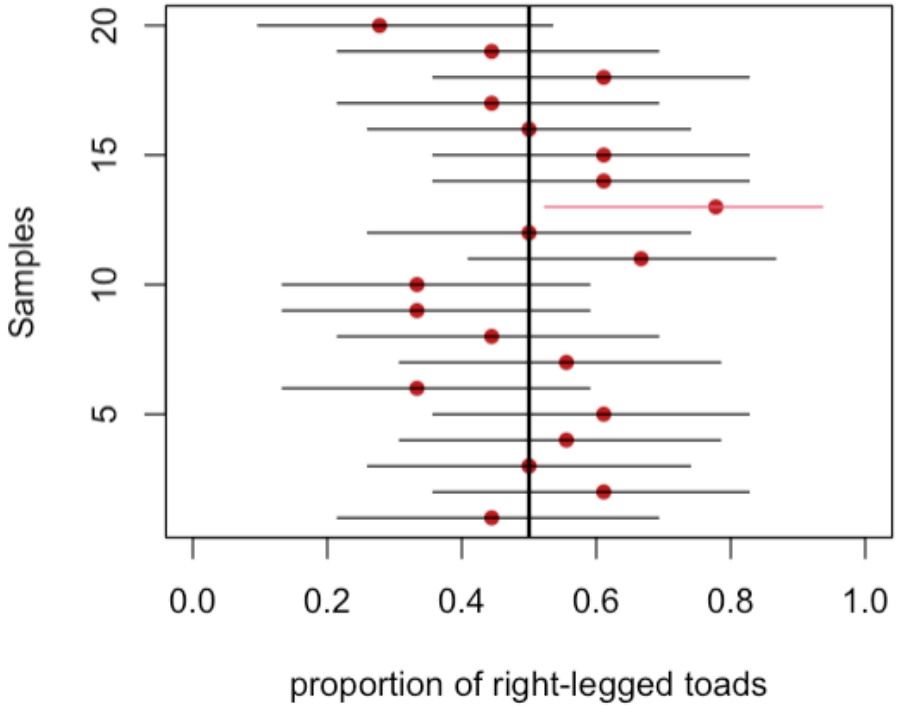
Alternative hypothesis (H_A): the proportion of right- and left-handed toads in the population ARE NOT equal.

It estimates a p-value (seen often as more "quantitative") than when contrasted to a threshold (alpha, i.e., significance level) it "forces" us into a yes (reject H_0) or no (don't reject H_0) answer.



Statistical hypothesis testing can be also understood as building the confidence interval for samples that come from a population of “no interest”

20 random samples



Assume that 50% of toad population is right-legged and 50% left-handed. In other words, assume that the null hypothesis is true.

Assume this population to be mathematically infinite.

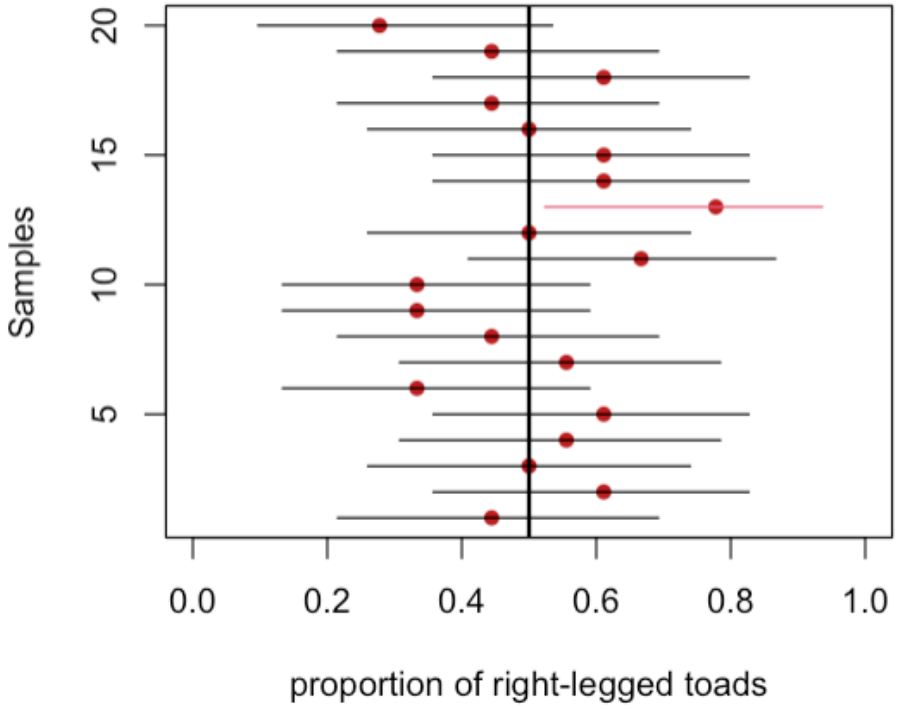
It is expected that 1 sample confidence interval over 20 does not include the true population proportion (i.e., 95% of intervals). This interval is in red.

H_0 true

Statistical hypothesis testing can be also understood as building the confidence interval for samples that come from a population of “no interest”

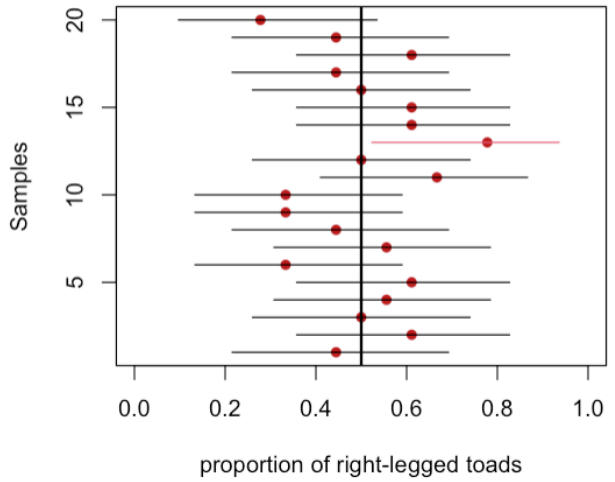
It is expected that 1 sample confidence interval over 20 does not include the true population proportion (i.e., 95% of intervals). This interval is in red.

The principle of statistical hypothesis testing is that if a sample confidence interval covers the value underlying the null hypothesis (here 50%), then we should reject the null hypothesis.



H_0 true

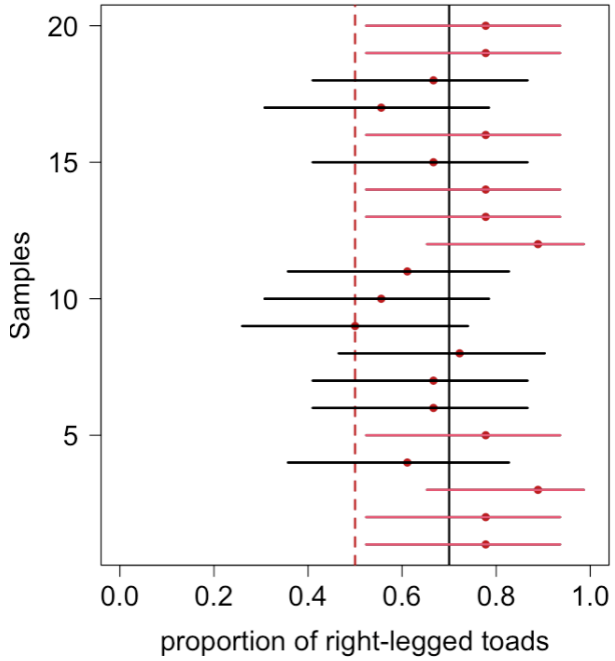
Statistical hypothesis testing can be also understood as building the confidence interval for samples that come from a population of “no interest”



H_0 true

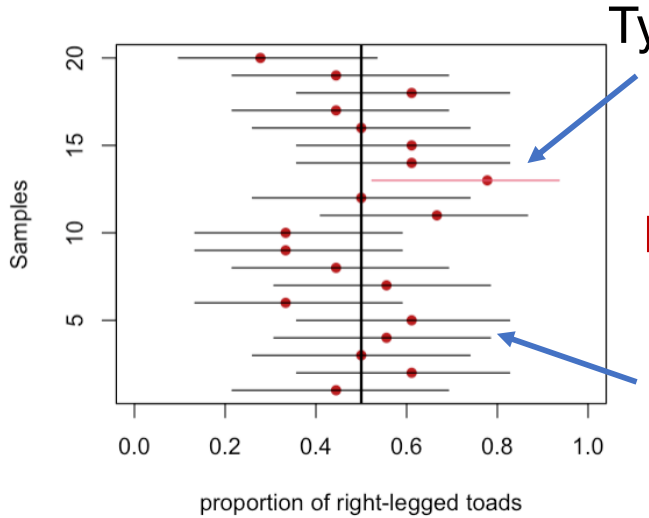
While only 5% of the confidence intervals based on the 50% right-legged population rejects H_0 when it should not (type I error); many intervals also do not reject when they should reject (type II error). Therefore, we can’t state that a H_0 is true; all we can say is that we have evidence to reject it.

Type II errors (probability off not rejecting a H_0 that is false) will depend here on the sample size and the value of the true population.



H_0 is false

Errors underlying statistical hypothesis testing – using confidence intervals to increase the ability to understand this notion



Type I error

Incorrect assessment

H_0 true

Correct assessment

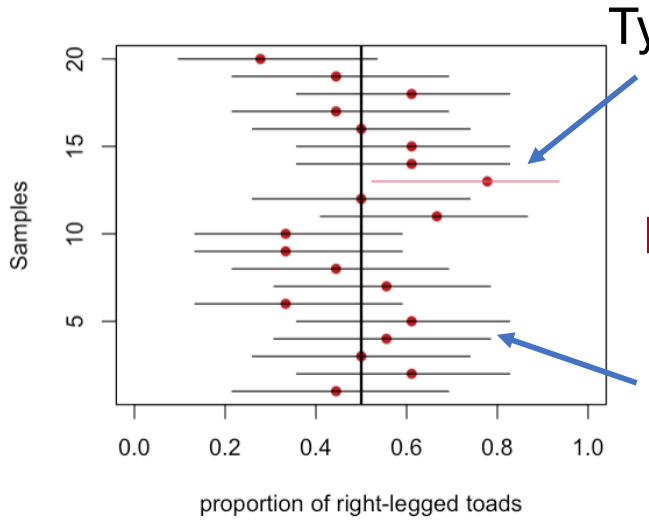
	Reality (unknown)	
Statistical decision	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

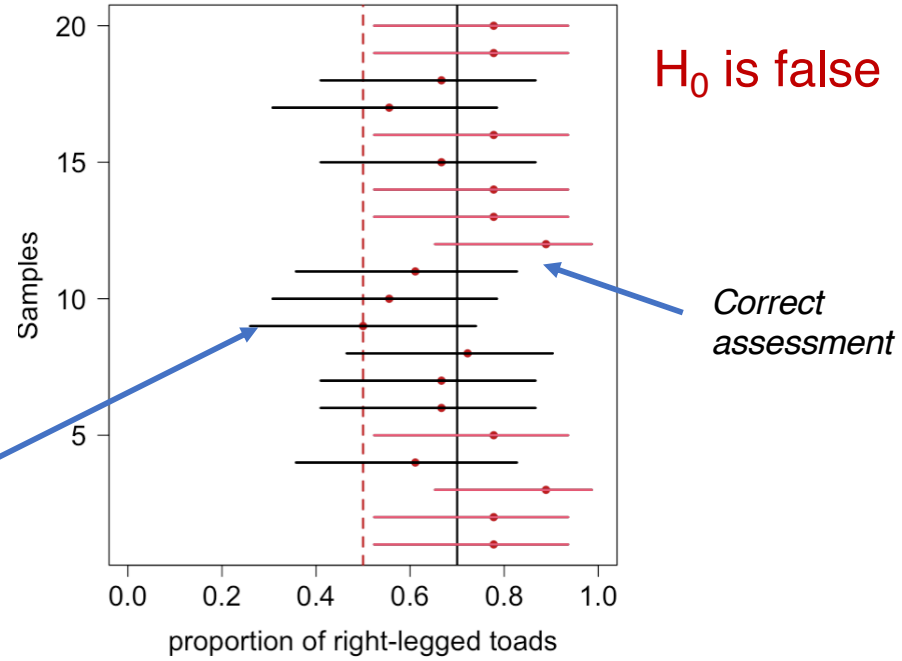
Inference is not about being certain—it is about managing uncertainty. Narrow intervals support decisions, provided we plan for the possibility of being wrong

Errors underlying statistical hypothesis testing – using confidence intervals to increase the ability to understand this notion

	Reality (unknown)	
Statistical decision	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

Inference is not about being certain—it is about managing uncertainty. Narrow intervals support decisions, provided we plan for the possibility of being wrong





Frequentist statistical hypothesis testing uses p-values to support statistical conclusions (e.g., do not reject vs reject the null hypothesis). And confidence intervals "always" (rarely not) agree with p-value-based decisions.

10000000000 random samples (~infinite)

Assume that 50% of toad population is right-legged and 50% are left-handed. Assume this population to be mathematically infinite.

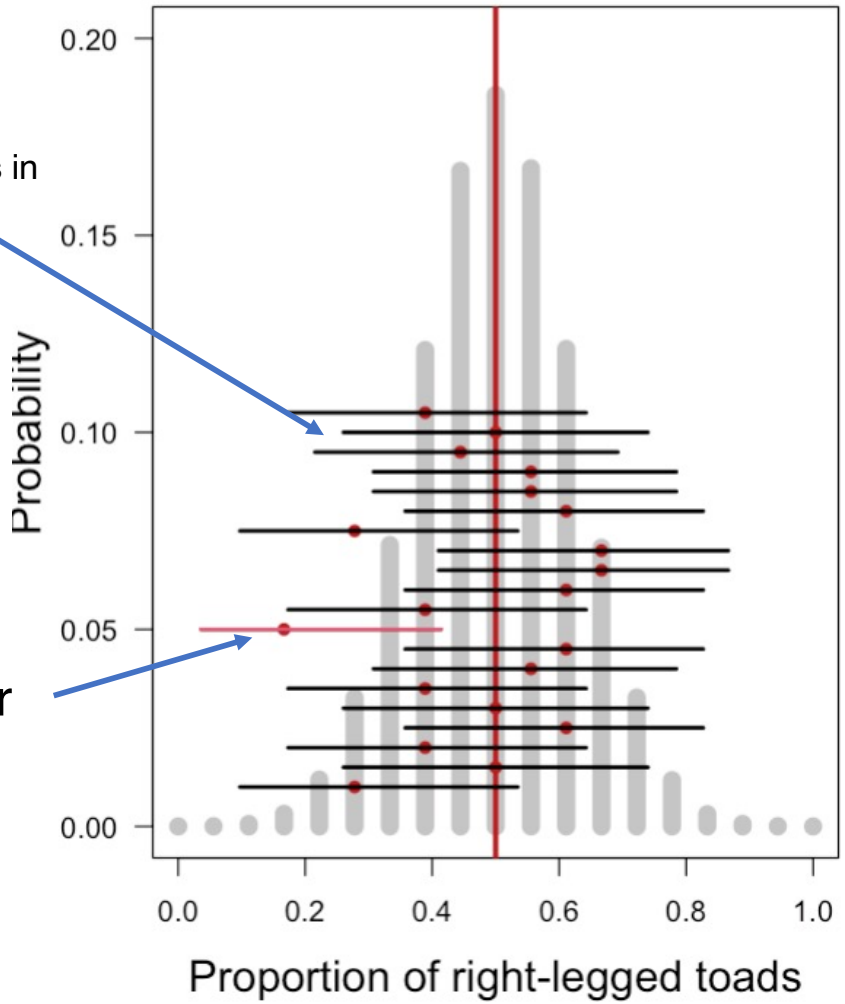
Number of right-handed toads	Probability of those samples
0	0.000004
1	0.00007
2	0.0006
3	0.0031
4	0.0117
5	0.0327
6	0.0708
7	0.1214
8	0.1669
9	0.1855
10	0.1669
11	0.1214
12	0.0708
13	0.0327
14	0.0117
15	0.0031
16	0.0006
17	0.00007
18	0.000004
Total	1.0

Contrasting the confidence intervals from samples from a population where H_0 is true against the null distribution assuming H_0 is true. Any wrong assessment is then obviously an error (wrong assessment).

Confidence intervals in black are correct assessments for those samples

Type I error

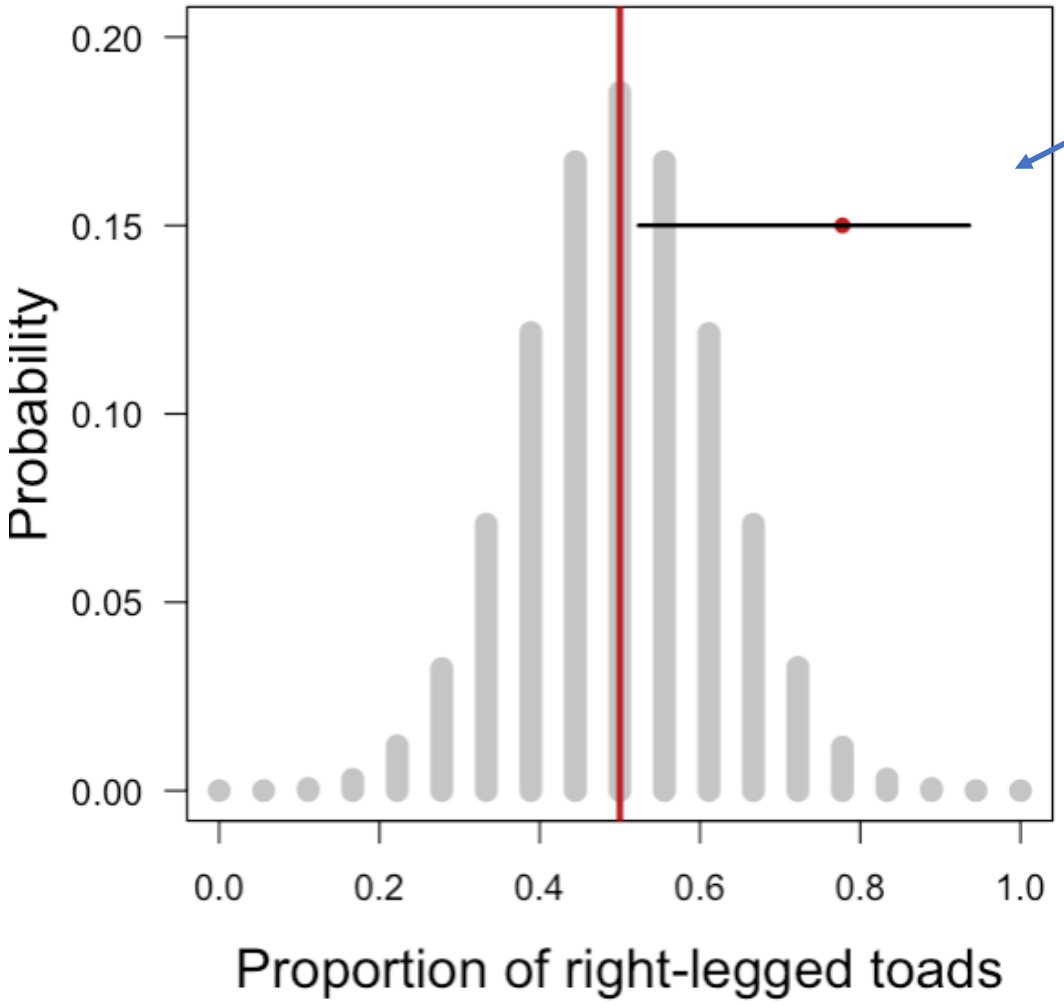
Incorrect assessment (in red)



In this example, we know which assessments are correct or incorrect because the samples were generated from a population in which the null hypothesis (H_0) is known to be true.

In real applications, we do not know whether our samples truly come from the population assumed under H_0. This setup was used purely for pedagogical purposes—to illustrate the fundamental properties of confidence intervals and statistical hypothesis testing.

Contrasting the confidence intervals from the observed sample (14 right-legged) against the distribution of values under the H_0



Confidence interval
for the observed value
14 right-legged (i.e., 77.8%)

Remember that we don't know reality because it's all based on a sample.

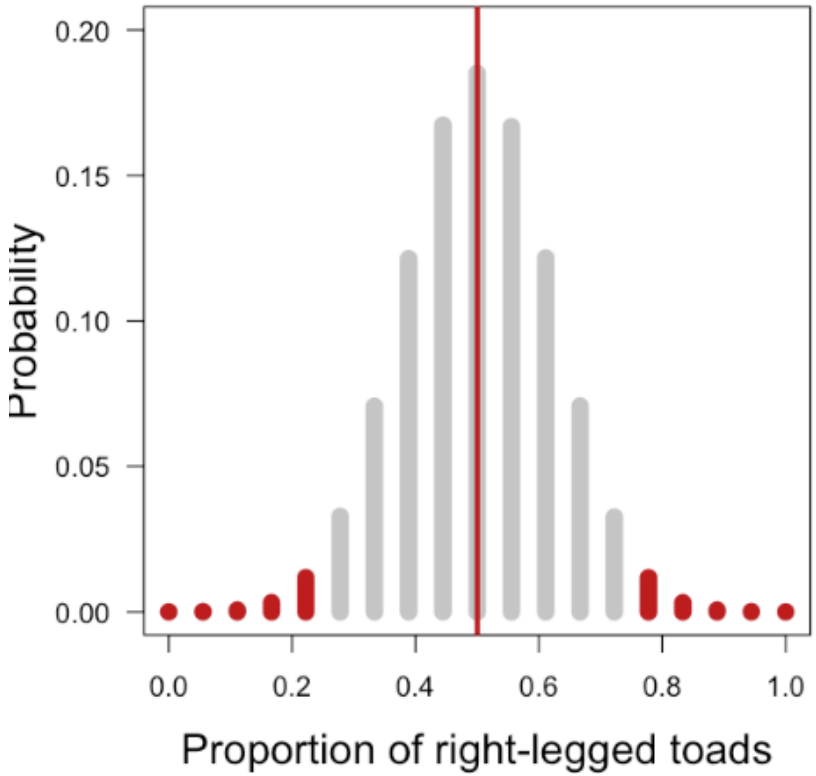
But we can say that we are 95% confident that this is NOT a type I error.

Because the answer (reject or not) always agree, we could estimate the p-value from the confidence interval for the sample and the null distribution. This is a bit advanced for our level right now.

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



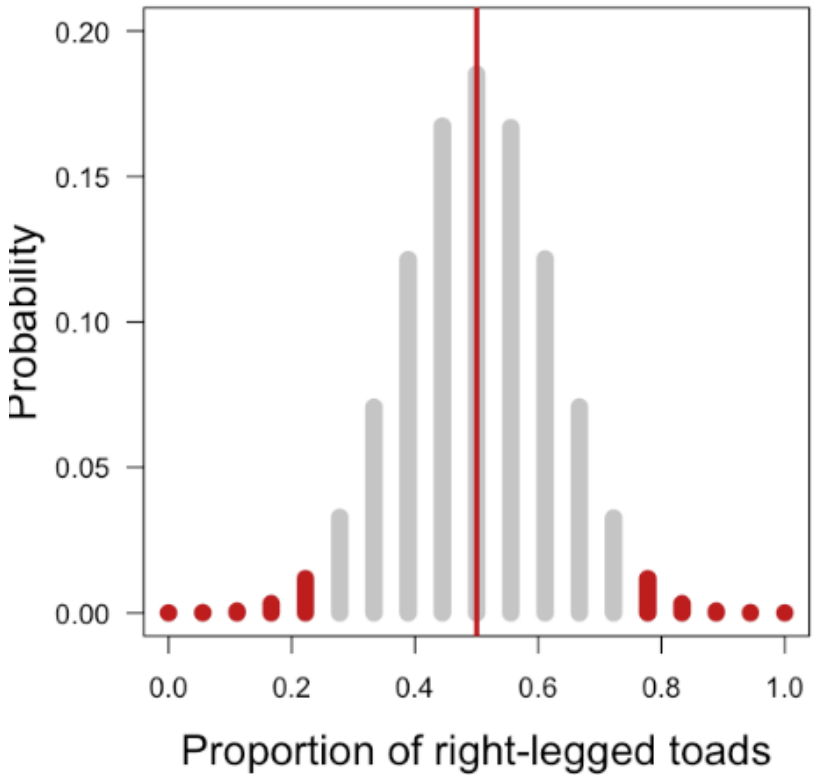
Pr[14 or more right-handed toads] = Pr[14] + P[15] + P[16] + P[17] + P[18] = 0.0155 x 2 (symmetric distribution) = **0.031**

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 p-value is NOT the probability that the null hypothesis is true. IT IS the probability of observing a value of the test statistic that is as or more extreme than what was observed in the sample, assuming the null hypothesis is true.

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)



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**

Let’s take a break - 1 minute

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

The **p-value** is the probability of the observed sample data assuming that the null hypothesis is true.

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 lack of handedness (i.e., 50%/%50) is unlikely.

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

P = 0.031

AGAIN, and VERY IMPORTANT; and also “confusing”:

We can say that we have evidence to reject (can’t say *not accept*) the null statistical hypothesis, but we cannot say that we have evidence to accept a specific alternative hypothesis.

This is because our decision is based entirely on the sampling distribution expected under chance alone, assuming the null hypothesis (H_0) is true (e.g., 50% right-legged and 50% left-legged toads).

By rejecting H_0, we conclude that the observed data are inconsistent with what would be expected under the null model (i.e., probability distribution assuming H_0 as true). This provides support for the research hypothesis (that something other than chance is at work), but not confirmation of any particular alternative, since infinitely many alternative hypotheses could also explain the deviation from H_0.

The process of statistical hypothesis testing:
SUMMARY OF 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 hypotheses 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 alternative hypothesis often is not.

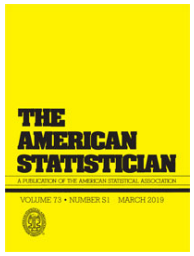
Research hypotheses cannot be proven right or wrong from the data. Hypotheses can be said to be either refuted (evidence is against the research hypothesis) or supported (evidence is in favour of the research hypothesis) by the data generated.

What does the significance level (α level) represent?

There is disagreement among statisticians and users about whether to reject or not reject (referred as to thresholding) 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!



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 hypothesis testing (Wasserstein et al. 2019)

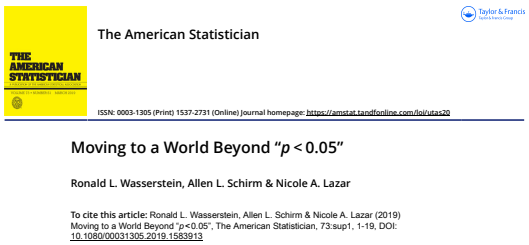
1. P-values indicate how incompatible the observed data are with a specified statistical model (e.g., the one assumed under H_0).

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) – (even though they currently are)

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

- There are other important don’ts that we will see later in the course.


THE AMERICAN STATISTICIAN

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

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

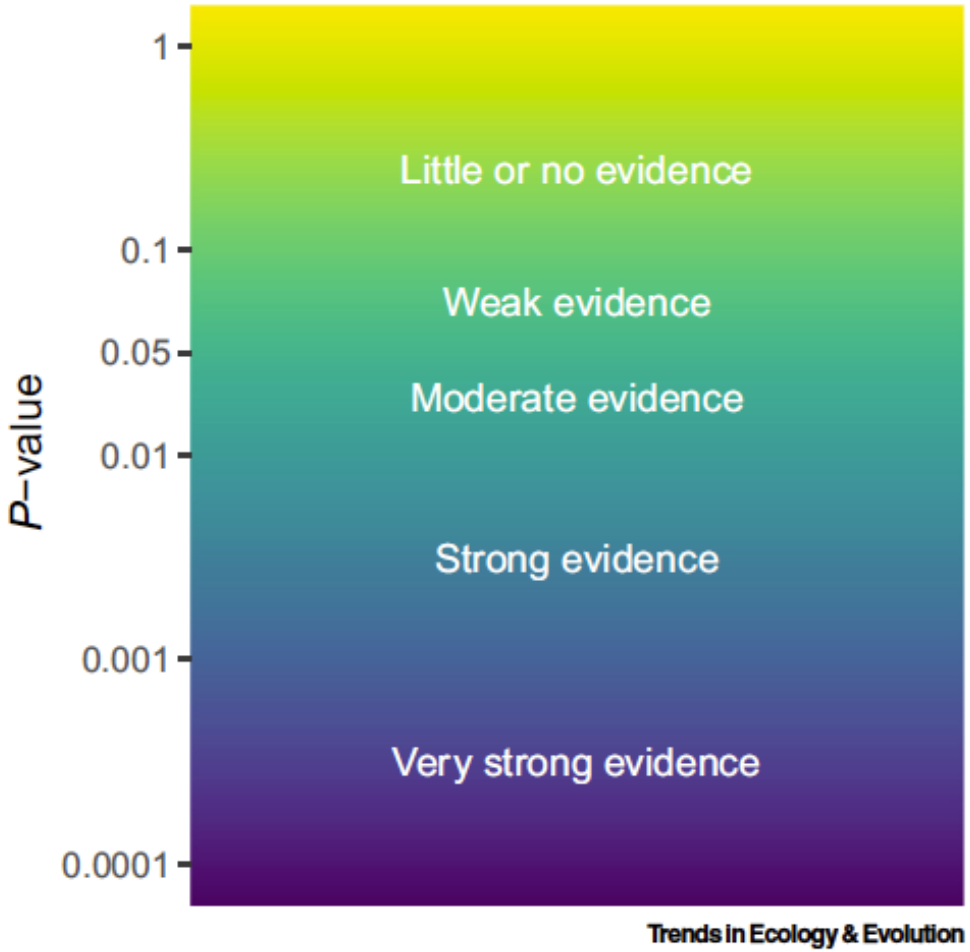


Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

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 H_0 (one fixed assumed parameter as true) but not accept H_A (many potential parameters can 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.

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.

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 our course but it’s hard to get away from this common culture in the statistical applications in biology and in most other fields.

Statistical inferential process: we use sampling theory to determine the sampling distribution required to estimate uncertainty around a sample value of interest (e.g., number of right- and left-legged toads).