STATISTICAL SIGNIFICANCE IN BIOLOGY

Conventions, Abstractions, and the Future

Erik Kusch

erik.kusch@uni-leipzig.de

Behavioural Ecology Research Group
University of Leipzig

18/06/2019

1 The Reproducibility Crisis
- What crisis?
- Why are we in this crisis?

2 The p-Value Conundrum
- Background
- Alternatives

3 Finding A Solution
- Summary
- Discussion

Outline

1. The Reproducibility Crisis
   - What crisis?
   - Why are we in this crisis?
2. The p-Value Conundrum
   - Background
   - Alternatives
3. Finding A Solution
   - Summary
   - Discussion

Irreproducible research

Reproducibility analyses have shown that only a surprisingly small portion of studies can be replicated.

∼ Nuzzo (2015). FOOLING OURSELVES. Nature.

This manifests in:

- Large sample-to-sample variations of the p-value

  ∼ Halsey et al. (2015). The fickle P value generates irreproducible results. Nature Methods.

- Ambiguity in data handling procedures

  ∼ Peng & Leek (2015). P values are just the tip of the iceberg. Nature.

- Difficulty in establishing meta-analyses

  ∼ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Thus, our studies become **solitary glances behind the curtain**.

Irreproducible research

Reproducibility analyses have shown that only a surprisingly small portion of studies can be replicated.

∼ Nuzzo (2015). FOOLING OURSELVES. Nature.

This manifests in:

- Large sample-to-sample variations of the p-value

  ∼ Halsey et al. (2015). The fickle P value generates irreproducible results. Nature Methods.

- Ambiguity in data handling procedures

  ∼ Peng & Leek (2015). P values are just the tip of the iceberg. Nature.

- Difficulty in establishing meta-analyses

  ∼ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Thus, our studies become **solitary glances behind the curtain**.

Irreproducible research

Reproducibility analyses have shown that only a surprisingly small portion of studies can be replicated.

$\sim$ Nuzzo (2015). FOOLING OURSELVES. Nature.

This manifests in:

- Large sample-to-sample variations of the p-value

  $\sim$ Halsey et al. (2015). The fickle P value generates irreproducible results. Nature Methods.

- Ambiguity in data handling procedures

  $\sim$ Peng & Leek (2015). P values are just the tip of the iceberg. Nature.

- Difficulty in establishing meta-analyses

  $\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Thus, our studies become **solitary glances behind the curtain**.

Irreproducible research

Reproducibility analyses have shown that only a surprisingly small portion of studies can be replicated.

$\sim$ Nuzzo (2015). FOOLING OURSELVES. Nature.

This manifests in:

- Large sample-to-sample variations of the p-value

  $\sim$ Halsey et al. (2015). The fickle P value generates irreproducible results. Nature Methods.

- Ambiguity in data handling procedures

  $\sim$ Peng & Leek (2015). P values are just the tip of the iceberg. Nature.

- Difficulty in establishing meta-analyses

  $\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Thus, our studies become **solitary glances behind the curtain**.

Irreproducible research

Reproducibility analyses have shown that only a surprisingly small portion of studies can be replicated.

$\sim$ Nuzzo (2015). FOOLING OURSELVES. Nature.

This manifests in:

- Large sample-to-sample variations of the p-value

  $\sim$ Halsey et al. (2015). The fickle P value generates irreproducible results. Nature Methods.
- Ambiguity in data handling procedures

  $\sim$ Peng & Leek (2015). P values are just the tip of the iceberg. Nature.
- Difficulty in establishing meta-analyses

  $\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Thus, our studies become **solitary glances behind the curtain**.

Reasons for the reproducibility crisis

Some select phenomena that brought us here:

Dichotomy of p-values

- Arbitrary $.05$ significance cut-off generates a false dichotomy of *'false'* or *'true'* conclusions
- Significant effect are not necessarily biologically relevant

∼ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

Peer-review shortcommings

- Reluctancy to make corrections
- No clear guidelines on where to direct criticism towards
- No standard process for data and code access

∼ Allison (2016). A tragedy of errors. Nature.

Research integrity

- Research questions often formulated post-hoc leading to *multiple testing* issue
- Sloppy reporting of data handling procedures
- Lack of data and code repository guidelines
- Lack in pre-specification of research

∼ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Reasons for the reproducibility crisis

Some select phenomena that brought us here:

Dichotomy of p-values

- Arbitrary $.05$ significance cut-off generates a false dichotomy of *'false'* or *'true'* conclusions
- Significant effect are not necessarily biologically relevant

$\sim$ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

Peer-review shortcommings

- Reluctancy to make corrections
- No clear guidelines on where to direct criticism towards
- No standard process for data and code access

$\sim$ Allison (2016). A tragedy of errors. Nature.

Research integrity

- Research questions often formulated post-hoc leading to *multiple testing* issue
- Sloppy reporting of data handling procedures
- Lack of data and code repository guidelines
- Lack in pre-specification of research

$\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Reasons for the reproducibility crisis

Some select phenomena that brought us here:

Dichotomy of p-values

- Arbitrary $.05$ significance cut-off generates a false dichotomy of *'false'* or *'true'* conclusions
- Significant effect are not necessarily biologically relevant

$\sim$ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

Peer-review shortcommings

- Reluctancy to make corrections
- No clear guidelines on where to direct criticism towards
- No standard process for data and code access

$\sim$ Allison (2016). A tragedy of errors. Nature.

Research integrity

- Research questions often formulated post-hoc leading to *multiple testing* issue
- Sloppy reporting of data handling procedures
- Lack of data and code repository guidelines
- Lack in pre-specification of research

$\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Reasons for the reproducibility crisis

Some select phenomena that brought us here:

Dichotomy of p-values

- Arbitrary $.05$ significance cut-off generates a false dichotomy of *'false'* or *'true'* conclusions
- Significant effect are not necessarily biologically relevant

$\sim$ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

Peer-review shortcommings

- Reluctancy to make corrections
- No clear guidelines on where to direct criticism towards
- No standard process for data and code access

$\sim$ Allison (2016). A tragedy of errors. Nature.

Research integrity

- Research questions often formulated post-hoc leading to *multiple testing* issue
- Sloppy reporting of data handling procedures
- Lack of data and code repository guidelines
- Lack in pre-specification of research

$\sim$ Cumming (2014). The New Statistics: Why and How. Psychological Science.

Outline

1 The Reproducibility Crisis
- What crisis?
- Why are we in this crisis?

2 The p-Value Conundrum
- Background
- Alternatives

3 Finding A Solution
- Summary
- Discussion

What is the p-value and why is it isnufficient?

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the $p$value will always be below the $.05$ cut-off, no matter the magnitude of the effect

What is the p-value and why is it isnufficient?

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the $p$value will always be below the $.05$ cut-off, no matter the magnitude of the effect

What is the p-value and why is it isnufficient?

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the $p$value will always be below the $.05$ cut-off, no matter the magnitude of the effect

Effect sizes

"A measure of the magnitude of a statistical effect within the data (i.e. values calculated from test statistics)."

$\sim$ Nakagawa & Cuthill (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. Biological Reviews.

- **Intuitive** to interpret and often what we are interested in
- Three types for most situations:
  - r statistics (correlations)
  - d statistics (comparisons of values)
  - OR (odds ratio) statistics (risk measurements)
- These are **point estimates**
- Need to be reported alongside some information of credibility
- These are usually *standardised* thus enabling meta-studies

In `R`: https://cran.r-project.org/web/packages/compute.es/compute.es.pdf and https://cran.r-project.org/web/packages/effsize/effsize.pdf

Confidence Intervals

"Confidence intervals (CIs) answer the questions: 'How strong is the effect' and 'How accurate is that estimate of the population effect'."

∼ Halsey (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biology Letters.

- **Intuitive** to interpret
- Answers the questions we are most interested in
- Does not require additional information of statistical certainty
- Combines **point estimates** and **range estimates**
- Removes some of the pressure of the *"file drawer problem"*
- Shares the same mathematical framework as the p-value calculation
- Especially useful in **data visualisation**

In R, many functions come with in-built ways of establishing CIs.

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a indicator of model fit.

~ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

- Used for **model selection and comparison**
- Lower AICs indicate better model fit
- One can establish contrasting models adhering to different hypothesis and identify which model suits the data best
- A proper hypothesis selection tool
- Model selection often comes with some degree of uncertainty
- Can be misused in step-wise model building procedures

In `R`, most model outputs can be assessed using the `AIC()` function.

Bayes Factor

" The minimum Bayes factor is simply the exponential of the difference between the log-likelihoods of two competing models."

$\sim$ Goodman (2001). Of P-Values and Bayes: A Modest Proposal. Epidemiology.

- **Intuitive** to interpret (Bayes Factor of 1/10 means that our study decreased the relative odds of the null hypothesis being true tenfold)
- Uses prior information to establish expected likelihoods thus enabling a progression in science

In `R`: https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf or direct Bayesian Statistics using JAGS or STAN (for example)

Outline

1 The Reproducibility Crisis
- What crisis?
- Why are we in this crisis?

2 The p-Value Conundrum
- Background
- Alternatives

3 Finding A Solution
- Summary
- Discussion

Research Integrity

1. Distinguish between **prespecified** (answering a question) and **exploratory** (formulating a question) studies.
2. Express **research question in terms of expectations** of effect sizes
3. Identify the **effect sizes best suited to answer** these **questions**
4. **Report full study plan before commencing data collection**
5. **Calculate measures** of statistical meaning that **enable meta-studies** (e.g. effect sizes and CIs)
6. Make sure to **correctly interpret the results** outside of the p-value dichotomy of true and false
7. **Report the findings in a meta-analytic context**

Where do we go from here?

"Treat statistics as a science, and not a recipe"
$\sim$ Andrew Vickers

"The numbers are where the scientific discussion should start, not end!"
$\sim$ Regina Nuzzo