Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 8: Interpretation, causality, cautionary notes

27./28. April 2017

Overview (todo: check)

- *P*-values: Interpretation and (mis-)use

- Statistical significance vs biological relevance

- Relative importance of regression terms

- Causality vs correlation

- Bradford-Hill criteria for causal inference

- Experimental vs observational studies

Course material covered today

- todo

Optional reading:

- todo

Recap of Last week

- todo

P-values

Recap:
P-values are commonly used for *statistical testing*, e.g. by checking if $p < 0.05$.

Examples:

- T-test for a difference between two samples.
- χ^2-test for independence of two discrete distributions.
- Test if a regression coefficient $\beta_x \neq 0$ in a regression model.

Such tests are useful whenever a **decision** needs to be made (e.g., in clinical trials, intervention actions in ecology etc.).

P-values in regression models

In regression modelling, the p-value is often used as an indicator of covariate importance. Remember the mercury example:

	Coefficent	95%-confidence interval	p-value
Intercept	-0.68	from -0.88 to -0.47	< 0.0001
log10(Hg_soil)	0.033	from -0.05 to 0.11	0.42
vegetables	0.07	from -0.03 to 0.17	0.18
migration	-0.036	from -0.19 to 0.12	0.65
smoking	0.27	from 0.06 to 0.48	0.012
sqrt(amalgam)	0.33	from 0.24 to 0.42	< 0.0001
age	-0.042	from -0.06 to -0.02	0.0004
mother	-1.03	from -1.70 to -0.35	0.003
sqrt(fish)	0.079	from 0.03 to 0.13	0.004
last_fish	0.30	from 0.15 to 0.45	< 0.0001
age:mother	0.055	from 0.03 to 0.08	0.0002

A common practice is to look only at the p-value and use $p < 0.05$ to decide whether a variable has an influence or not ("is significant or not").

P-values criticism

P-value **criticism is** as **old** as statistical significance testing (1920s!). Issues:

- The sharp line $p < 0.05$ is arbitrary and significance testing according to it may lead to *mindless statistics* (Gigerenzer, 2004).

- *P*-hacking / dada dredging: Search until you find a result with $p < 0.05$.

- Publication bias: Studies with $p < 0.05$ are more likely to be published than "non-significant" results.

- Recent articles in *Science*, *Nature* or a statement by the *American Statistical Associaton (ASA)* in March 2016 show that the debate still continues (Claridge-Change and Assam, 2016; Goodman, 2016; Wasserstein and Lazar, 2016).

- Model selection using *p*-values may lead to a model selection bias (see last week).

P-values even made it into NZZ (April 2016)

Note: R.A. Fisher, the "inventor" of the p-value (1920s) didn't mean the p-value to be used in the way it is used today (which is: doing a single experiment and use $p < 0.05$ for a conclusion)!

From Goodman (2016):

> *Fisher used "significance" merely* **to indicate that an observation was worth following up, with refutation of the null hypothesis justified only if further experiments "rarely failed" to achieve significance.** *This is in stark contrast to the modern practice of making claims based on a single demonstration of statistical significance.*

The misuse of p-values has led to a reproducibility crisis in science!

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

(Ioannidis, 2005)

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2
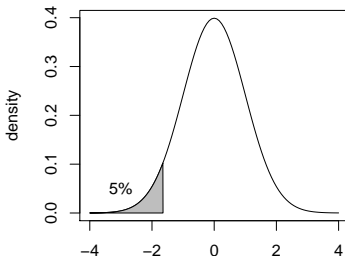
What is the problem with the p-value?

Many applied researchers do not really understand what the p-value actually is.

> The **formal definition of p-value** is the probability of an observed data summary (e.g., an average) and its more extreme values, given a specified mathematical model and hypothesis (usually the "Null", indicating "no effect").
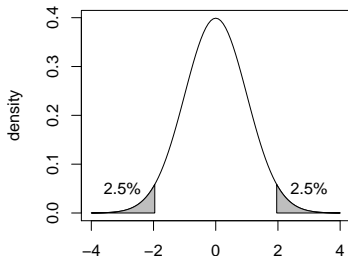
Two–sided p–value   One–sided p–value

Test yourself: Klicker-Exercise

http://www.klicker.uzh.ch/bkx

+ Discussion of the results!

What is the problem with the p-value? II

- The p-value is often used to classify results into "significant" and "non-significant". Typically: $p < 0.05$ vs $p \geq 0.05$.

- However, this is often too crude!

- It is much better to have a more gradual interpretation of the p-value (see slide 18).

Quote from ASA statement:

In February, 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach p = .05?
A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use p = 0.05?
A: Because that's what they were taught in college or grad school.

Significance vs relevance

In regression models:

- A low *p*-value does not automatically imply that a variable is "important".
- "Is there an effect?" v.s. "How much of an effect is there?".

(from Goodman, 2008)
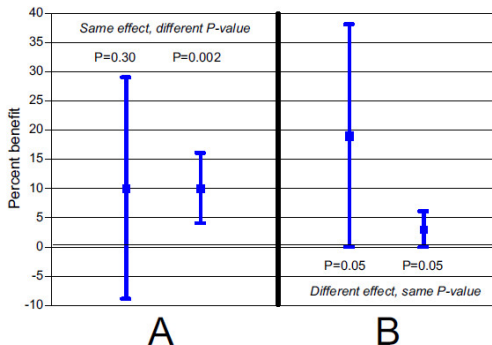
In addition:

A large p-value (e.g., $p > 0.05$) does not automatically imply that a variable is "unimportant".

Absence of evidence is not evidence of absence (Altman and Bland, 1995).

In other words:

One cannot prove the Null Hypothesis!!

Several reasons may lead to large p-values:

- Low sample size ($\rightarrow$ low power).
- The truth is not "far" from the null hypothesis.
  Example: Small effect sizes in regression models.

Shall we abolish p-values?

No: p-values are not "good" or "bad". They contain important information, and they have have **strengths** and **weaknesses**.

Suggestions:

1. Use p-values, but don't over-interpret them, use them properly.

2. Look at effect sizes and confidence intervals.

3. Look at relative importances of covariates.

4. **Don't use p-values for model selection.**

Suggestion 1: Proper interpretation of p-values

Rather than a black-and-white decision ($p < 0.05$), Martin Bland suggests to regard p-values as continuous measures for statistical evidence (Introduction to Medical Statistics, 4th edition, Oxford University Press):

$p > 0.1$	little or no evidence against the null hypothesis
$0.1 > p > 0.05$	weak evidence
$0.05 > p > 0.01$	evidence
$0.01 > p > 0.001$	strong evidence
$p < 0.001$	very strong evidence

But: The level of significance must also depend on the context!

In the Hg example:

	Coefficent	95%-confidence interval	p-value
Intercept	-0.68	from -0.88 to -0.47	< 0.0001
log10(Hg_soil)	0.033	from -0.05 to 0.11	0.42
vegetables	0.07	from -0.03 to 0.17	0.18
migration	-0.036	from -0.19 to 0.12	0.65
smoking	0.27	from 0.06 to 0.48	0.012
sqrt(amalgam)	0.33	from 0.24 to 0.42	< 0.0001
age	-0.042	from -0.06 to -0.02	0.0004
mother	-1.03	from -1.70 to -0.35	0.003
sqrt(fish)	0.079	from 0.03 to 0.13	0.004
last_fish	0.30	from 0.15 to 0.45	< 0.0001
age:mother	0.055	from 0.03 to 0.08	0.0002

- **Little or no evidence:** Hg soil, vegetables from garden, migration background
- **Weak evidence:** Smoking
- **Strong evidence:** Mother, monthly fish consumption
- **Very strong evidence:** Amalgam, age, last fish ($>$ or $<$ 3 days), interaction of age and mother

Suggestion 2: Report effect sizes....

Ask: **Is the effect size relevant?**

Example

WHO recommendation concerning smoking and the consumption of processed meat. Both, smoking and meat consumption, appear to be carcinogenic.

- 50g processed meat per day increases the risk for colon cancer by a factor of 1.18 (+18%).
- Smoking increases the risk for cancer by a factor of 3.6 (+260%).

Thus: Although both, meat consumption and smoking, are carcinogenic ("significant"), their **effect sizes are vastly different**!

...and 95% CIs

Ask: **Which range of true effects is statistically consistent with the observed data?**

Example

Body fat example, slide 39 of week 1.

The effect estimate for the effect of BMI on body fat is given as $\hat{\beta}_{BMI} = 1.82$, 95% CI from 1.61 to 2.03.

Interpretation: for an increase in the bmi by one index point, roughly 1.82% percentage points more bodyfat are expected, and all true values for β_{BMI} between 1.61 and 2.03 are **compatible with the observed data**.

However...

- The choice of the 95% is again somewhat arbitrary. We could also go for 90% or 99% or any other interval, but 95% has established as a commonly accepted range.

- The 95% CI should **not be misused for simple hypothesis testing** in the sense of

  "Is 0 in the confidence interval or not?"

  Because this boils down to checking whether $p < 0.05$ ...

Suggestion 3: Look at relative importances of covariates

- Ultimately, the popularity of *p*-values is based on the wish to judge which covariates are **relevant** in a model, particularly in observational studies.

- The problem with this: Low *p*-values do not automatically imply high relevance (Cox, 1982).

- Alternative: **relative importances** of explanatory variables that measure the proportion (%) of the responses' variability explained by each variable.

Relative importance: Decomposing R^2

Remember: R^2 indicates the proportion of variance explained by **all** covariates in a model

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_2 x_i^{(m)} + e_i \ .$$

The aim of **relative importance** is to decompose R^2 such that

- each variable $x^{(j)}$ is attributed a fair share r_j.
- the sum of all importances sums up to R, that is, $\sum_{j=1}^{p} r_j = R^2$.

Further, it is required that

- all shares are ≥ 0.

Question: How would you define/calculate relative importance?

- **Idea 1:** Fit simple models including only one covariate at the time, *i.e.*:

$$y_i = \beta_0 + \beta_j x_i^{(j)} + e_i$$

for each variable $x^{(j)}$ and use the respective R^2 as r_j.

- **Idea 2:** Fit the linear model twice, once with and once without the covariate of interest, and then take the increase of R^2 as r_j.

Problem: In practice, regressors $x^{(j)}$ are *always correlated*, thus both ideas lead to $\sum_j r_j \neq R^2$!

To understand the problem of ideas 1 and 2, let us fit three models for $\log(Hg_{\text{urine}})$ with

- $x^{(1)} = \sqrt{\text{Number of monthly fish meals}}$
- $x^{(2)} = $ binary indicator if last fish meal was less than 3 days ago.

These two variables are correlated (people who consume a lot of fish are more likely to have it consumed within the last 3 days).

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + + e_i \qquad R^2 = 0.12 \qquad (1)$$

$$y_i = \beta_0 + \beta_2 x_i^{(2)} + + e_i \qquad R^2 = 0.08 \qquad (2)$$

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + e_i \qquad R^2 = 0.14 \qquad (3)$$

Note: The R^2 of the model with both covariates is much less than the sum of the R^2 from models (1) and (2)!

$\Rightarrow$ The increase of R^2 upon inclusion of a covariate depends on the covariates that are already in the model!

A better way to calculate relative importance?

Various proposals to calculate relative importance (R^2 decomposition) have been proposed. The (currently) most useful is given by the following idea, called **LMG** (**L**indemann, **M**erenda and **G**old 1980):

- Fit the model for all possible orderings of the covariates.

- Record the incrase in R^2 each time a variable is included.

- Average over all orderings of the covariates.

Luckily, the `R`-package `relaimpo` (Groemping 2006) contains the function `calc.relimp()` that does this for us!

Hg results

Which proportion (%) of variance in $\log(Hg_{\text{urine}})$ is explained by each covariate? Interpret the table below:

```
> library(relaimpo)
> lmg.hg <- calc.relimp(r.lm.hg)$lmg
```

Variable	Rel. imp. (%)	*p*-value
$\log(Hg_{\text{soil}})$	0.10	0.42
Vegetable	0.46	0.18
Migration	0.43	0.65
Smoking	1.21	0.012
Amalgam	19.69	<0.0001
Age	1.25	0.0004
Mother	1.08	0.0031
Fish	7.26	0.0042
Last fish	7.34	<0.0001
Age:mother	6.56	0.0002

Several variables have very low *p*-values, but their relative importance
differs clearly.

⇒ Relative importance gives intuitive complementary information to
p-values, effect sizes and confidence intervals!

Does relative importance solve all the problems?

Unfortunately not...

Relative importance should be understood as a complement to standard statistical output.

There are several limitations to it:

- Rel.imp. of a variable may heavily depend on the other variables included in the model, especially when there are strongly correlated variables (see slide 32).

- Hard to generalize to other, non-linear regression models.

Groemping 2007:

"...a request for a decomposition of R^2 is often driven by a desire to prioritize intervention actions with the intention to influence the response. It is important to notice that any intervention bears the risk [...] of not only influencing the targeted regressor but also the correlation structure among regressors. Thus, unexpected results may occur regarding changes of the response's variance. In this way, the benefit of the concept of decomposing R^2 is more limited than the typical user might realize."

Example

Compare the estimated relative importance for the variable `fish` (monthly fish meals) for two cases:

Model 1
Original Hg model.

Model 2
Model without the indicator variable `last_fish`.

- **Case 1:** Relative importance of `fish`: 7.26% (see slide 28).

- **Case 2:** Relative importance of `fish`: 10.75% .

Interpretation: If one of two correlated variables is removed, the other absorbs some of the importance from it.

Ev give another example where rel. imp is calculated, e.g. from previous weeks.

Causality vs correlation

In explanatory models the ultimate goal usually is to reveal causal relationships between the covariates and the response.

Examples:

- Does Hg in the soil influence Hg-levels in humans?

- Does inbreeding negatively affect population growth of Swiss Alpine ibex (Steinbock)?

- Does exposure to Asbest lead to illness or death?

- ...

However: Regression models actually only reveal associations, that is, correlations between **x** and **y**!

Example: Breakfast eating and obesity

Please read the following article and answer the questions below:
http://www.webmd.com/diet/news/20080303/
eating-breakfast-may-beat-teen-obesity

Questions:

- Does the cited study show that teens that eat breakfast are generally less obese?

- Does this automatically imply that eating breakfast **leads to** less obesity among teens?

Bradford-Hill-Criteria for causal inference I

In 1965 the Epidemiologist Bradford Hill presented a list of criteria to assess whether there is some causality or not. However, he wrote "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required sina qua non."

Bradford-Hill Criteria:

1. **Strength:** A causal relationship is likely when the observed association is strong.

2. **Consistency:** A causal relationship is likely if mutiple independent studies show similar associations.

3. **Specificity:** A causal relationship is likely when a covariate x is associated only with one potential outcome y and not with other outcomes.

4. **Temporality:** The effect has to occur after the cause.

Bradford-Hill-Criteria for causal inference II

⑤ Biological gradient: Greater exposure should generally lead to greater incidence of the effect.

⑥ Plausibility: A plausible mechanism is helpful.

⑦ Coherence: Coherence between findings in the lab and in the field / population increases the likelihood of an effect.

⑧ Analogy: Similar factors have a similar effect.

⑨ Experiment: Evidence from an experiment is valuable.

Experimental vs observational studies

Experimental studies are relevant in biology and even more so in medicine, e.g., in the context of clinical trials where novel drugs are tested.

The teen obesity study was an **observational study**:

- All study participants only had to report their behaviour.
- None of them was assigned to a treatment group.
- There was **no intervention**.

An observed effect is more likely to be *causal* if participants were *randomly assigned* to a group, here: breakfast eating yes/no.

Observational study ("Erhebung"):

- Obseration of subjects / objects in a real-world (existing) situation.

- Variables are usually correlated.

- More variables that can be included in the model.

- **Examples**: Influence of pollutans (mercury) on humans, studies of wild animal populations, epidemiological studies,...

Experimental study:

- Obseration of subjects / objects in a constructed (experimental) situation.

- Variables are controlled and uncorrelated (given a good study design!).

- Usually all variables enter the model, **no model selection**.

- **Examples**: Field experiments; clinical studies; psycological or pedagogical experiments,...

	Experiment	Observational study
Situation	Artificial, designed	Existing, cannot be influenced
Analysis	Simple, no model selection	Difficult (see model selection issues)
Interpretation	Clear, "proofs" causal relationship	Difficult, especially w.r.t. causality

Causality considerations for model selection

It is **widely unknown** that a model can be broken by the inclusion of a "wrong" covariate, which is causally associated in the wrong direction:
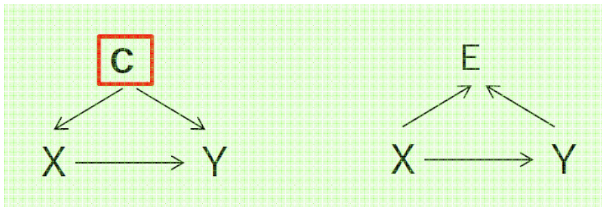
Remember: Avoid to include covariates in your model that are caused by the outcome!

Example: ...

Summary

- Try to understand the definition and the meaning of p-values.

- Correct understanding, use and interpretation of p-values: Do not use the "mindless" $p < 0.05$ criterion!!

- Statistical significance vs biological relevance: Ask for the effect size and confidence interval, and reflect what it means, instead of only reporting p-values.

- The p-value is not "bad", but it has to be used properly
  $\rightarrow$ 3 suggestions or alternatives (gradual interpretation on p-values, effect sizes and CIs, relative importances).

- Correlation should not be mistaken for causality.

- Experimental studies are better suited to reveal causality than observational studies!

References:

Altman, D. G. and J. M. Bland (1995). Absence of evidence is not evidence of absence. *British Medical Journal 311*, 485.

Claridge-Change, A. and P. N. Assam (2016). Estimation statistics should replace significance testing. *Nature 13*, 108–109.

Cox, D. R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology 14*, 325–331.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics 33*, 587–606.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science 352*, 1180–1182.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine 2*, e124.

Wasserstein, R. L. and N. A. Lazar (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.