Kurs Bio144: Datenanalyse in der Biologie

Lecture 9: Interpretation, causality, cautionary notes

Stefanie Muff (Lecture) & Owen L.Petchey (Practical)

University of Zurich

31 December, 2020

Recap of previous lecture

- Model selection is difficult.
- **Predictive** vs **explanatory** models.
- Information criteria for predictive models: AIC, AIC$_c$ and BIC
  $\rightarrow$ **model fit vs model complexity**
- Automatic model selection is inappropriate for explanatory models!
- Types of explanatory models
  - confirmatory
  - exploratory
- Strategies to fit explanatory models.

Overview

- 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

The lecture material of today is based, in part, on the literature that you studied during the self-study week (before Easter).

P-values

Recap:

P-values are often 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 might be 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 modeling, the p-value is often used as an indicator of covariate importance. Remember the mercury example:

```
## [1] "MISSING HG DATA"
```

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 / data 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 (Goodman, 2016; Wasserstein and Lazar, 2016; Amrhein et al. 2019).
- Model selection using *p*-values may lead to a model selection bias (see last week).

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

NZZ am Sonntag 3. April 2016 | Wissen | 61

Überschätzte Statistiken

Daten-Analysen entscheiden heute darüber, ob ein Medikament als wirksam gilt. Bloss verstehen viele Forscher die Bedeutung dieser Berechnungen gar nicht. **Von Patrick Imhasly**

Daten werden meist von Leuten analysiert, die nicht dafür ausgebildet sind.

Karriere machen können nicht nur Menschen, sondern auch statistische Grössen. Das gilt besonders für den sogenannten p-Wert, mit dem jeder Mittelschüler und jede Studentin in Kontakt kommt, vor allem aber jeder, der im weitesten Sinn etwas mit Statistik zu tun hat. Inzwischen ist der p-Wert indes auf die schiefe Bahn geraten. Denn was der Vater der modernen Statistik, der britische Genetiker Ronald Fisher, 1920 als eine Art informelles Sensorium für die Aussagekraft von Daten entwickelte, ist in der Praxis oftmals zu einem simplen Lackmustest verkommen.

Ergibt die statistische Analyse von Daten einen p-Wert < 0,05 (5 Prozent) oder noch besser < 0,01 (1 Prozent), gelten diese als signifikant – den Daten wird dann automatisch Beweiskraft zugestanden. Das entscheidet etwa darüber, ob ein neues Medikament als wirksam eingestuft wird oder ob ein Forscher seine Studie in einem angesehenen Fachblatt publizieren kann. «Der p-Wert war aber nie dazu gedacht, wissenschaftliches Denken ausser Kraft zu setzen», hat sich Ron Wasserstein, der Direktor der Amerikanischen Statistischen Vereinigung (ASA), jüngst öffentlich beklagt.

Wissenschafter verwenden den p-Wert immer häufiger, ohne zu verstehen, was er bedeutet – das fördert schlechte Forschung und untergräbt die Glaubwürdigkeit der Wissenschaft. Der Mediziner und Epidemiologe John Ioannidis von der Universität Stanford sprach in einem Kommentar von «Drogenabhängigen»: «Der falsche Gebrauch des p-Wertes ist dermassen einfach und erfolgt so automatisiert, dass manche süchtig werden danach – vor allem wenn sie mit Forschungsgeldern und Publikationen belohnt werden.» Angesichts der Missstände sah sich die ASA jetzt veranlasst, zum ersten Mal in ihrer fast 180-jährigen Geschichte Empfehlungen zu veröffentlichen, wie man mit einer statistischen Grösse vernünftig umgeht.

Wider die Null-Hypothese

«Der p-Wert sagt nicht das aus, was man gemeinhin von ihm erwartet», erklärt der Berner Epidemiologe Peter Jüni, der seit kurzem am Applied Health Research Centre der Universität Toronto tätig ist. Das bedeutet: Der p-Wert misst nicht die Wahrscheinlichkeit, ob eine bestimmte Hypothese zutrifft, und auch nicht, ob ein bestimmtes Resultat zufällig zustande gekommen ist, wie die ASA festhält. Vielmehr dienen p-Werte dazu, die sogenannte Null-Hypothese zu testen und möglichst zu verwerfen.

Die Hypothese in einer Patientenstudie könnte zum Beispiel lauten, dass ein Medikament A gegen Nierenprobleme besser wirkt als ein Medikament B. Die Null-Hypothese besagt dann genau das Gegenteil davon, nämlich dass das Medikament A nicht besser wirkt als das Medikament B. Beim Test berechnet der Forscher im Prinzip, wie gross die Wahrscheinlichkeit für das Auftreten eines tatsächlich festgestellten oder noch grösseren Unterschieds zwischen den beiden untersuchten Medikamenten ist – unter der Annahme, dass die Null-Hypothese stimmt. Diese Wahrscheinlichkeit ist der p-Wert, und je geringer er ist, desto weniger spricht für die Null-Hypothese. Ein p-Wert von 0,05 bedeutet, dass das festgestellte (oder ein noch extremeres) Resultat unter den Bedingungen der Null-Hypothese mit einer Wahrscheinlichkeit von lediglich 5 Prozent zustande kommen kann – und nicht, dass eine bestimmte Hypothese mit einer Sicherheit von 95 Prozent wahr ist.

«Studien führen heute zu dermassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von p-Werten kommt.»

Über die eigentlich untersuchte Hypothese kann der p-Wert nur indirekt etwas aussagen, weil er über zwei Ecken gedacht ist. Der ominöse Wert liefert also keine ultimativen Beweise für einen postulierten Unterschied oder Zusammenhang. «Der p-Wert ist eine bedingte und nicht eine absolute Wahrscheinlichkeit», erklärt Peter Jüni. «Doch genau das verstehen viele Forscher nicht, und es interessiert sie auch nicht.»

Hinzu kommt, dass die Signifikanzgrenzen von 5 Prozent bzw. 1 Prozent historisch entstanden und keineswegs klar definierte Werte sind. Ronald Fisher, der Erfinder des p-Werts, überliess es jedenfalls ausdrücklich der Interpretation des einzelnen Forschers, ab welcher Grösse ein p-Wert in einer Untersuchung Aussagekraft haben soll. «Trotzdem haben sich die willkürlich gewählten Signifikanzgrenzen ins Gehirn von Generationen von Forschern gebrannt», sagt Leonhard Held vom Institut für Epidemiologie, Biostatistik und Prävention der Universität Zürich.

Die britischen Statistiker Jonathan Sterne und George Davey Smith haben schon vor 15 Jahren im «British Medical Journal» dazu aufgerufen, die Resultate von medizinischen Studien nicht mehr als «signifikant» oder «nichtsignifikant» darzustellen, sondern im Kontext der gesamten Untersuchung und anhand anderer Ergebnisse zu interpretieren. Genützt hat das herzlich wenig, wie das Team von John Ioannidis in einer soeben erschienenen Studie festgestellt hat. Demnach sind in den vergangenen 25 Jahren in der biomedizinischen Forschung immer mehr Studien erschienen, die p-Werte angeben, die zudem immer klarer signifikant ausfallen. Gleichzeitig werden Zusatzinformationen zu den festgestellten Effekten immer seltener («JAMA», Bd. 315, S. 1141).

5%

Kleiner als dieser Wert muss der sogenannte p-Wert in einem statistischen Test sein, dann gelten die Daten aus einer Studie als aussagekräftig. Doch die Grenze ist willkürlich gewählt. (pim.)

Es gibt Alternativen

Das Problem ist dabei nicht nur, dass der p-Wert ein trügerisch einfaches statistisches Instrument ist. «Studien führen heute zu dermassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von p-Werten kommt», erklärt Leonhard Held. «Der eine oder andere fällt dann bestimmt signifikant aus, auch wenn kein Effekt vorhanden ist.» Leonhard Held und Peter Jüni verlassen sich deshalb bei der Planung und Auswertung von Studien schon lange nicht mehr nur auf die p-Werte.

Jüni empfiehlt, die Resultate von Studien mindestens mit Vertrauensintervallen zu versehen, die spezifische Aussagen über die Unsicherheit einer Schätzung machen. Und Held erforscht alternative Masse für statistische Evidenz – zum Beispiel Bayes-Faktoren. Mit ihrer Hilfe lässt sich die Wahrscheinlichkeit einer Hypothese anhand der Daten anpassen, statt dass diese wie beim p-Wert nach einem Schwarz-Weiss-Schema angenommen oder abgelehnt wird.

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!

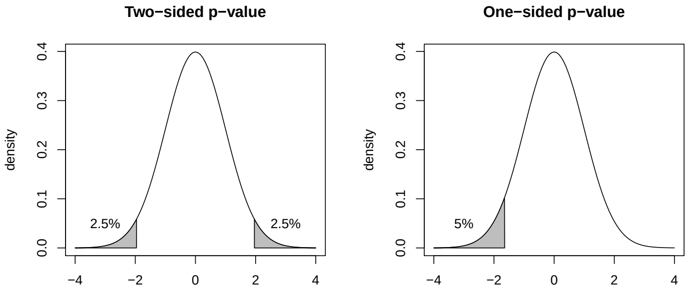
MISSING IMAGE Ioannidis2.png

(Ioannidis, 2005)

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 to observe a data summary (e.g., an average) that is at least as extreme as the one observed, given that the Null Hypothesis is correct.



Test yourself: Klicker-Exercise

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).

Probably the most important point to remember:

The p-value is **not** the probability that the Null Hypothesis is true!!!

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 $p = 0.05$?

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.
- Collinear covariates.

Shall we abolish *p*-values?

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

Suggestions:

1. Use *p*-values, but don't over-interpret them, use them properly.
2. Also look at effect sizes and confidence intervals.
3. Also look at relative importances of covariates.
4. **NEVER 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
$0.05 > p > 0.01$	moderate 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!

A suggestion from 2017 by 72 authors in the field:

(Benjamin et al., 2017, Nature Human Behaviour)

Their suggestion: replace $p < 0.05$ by $p < 0.005$. More precisely:

- Use $p < 0.005$ for **statistical significance.**
- Use $0.005 < p < 0.05$ as **suggestive evidence**.

The most recent suggestion, signed by >800 researchers:

(Amrhein et al., 2019, Nature)

Their suggestion: Do not use the term "statistical significance" at all.

In the Hg example:

[1] "Missing data"

- **Little or no evidence:** Hg soil, vegetables from garden, migration background
- **Moderate 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, are "significantly" increasing the probability to get cancer.

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

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

Paul D. Ellis writes in his book *The Essential Guide to Effect Sizes* (2010, chapter 2):

> *Indeed, statistical significance, which partly reflects sample size, may say nothing at all about the practical significance of a result. [....] To extract meaning from their results [...] scientists need to look beyond p values and effect sizes and* **make informed judgments about what they see**.

. . . and 95% CIs

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

Example

Body fat example, slide 40 of lecture 3.

The estimate for the slope of BMI in the regression for 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 in regression models 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

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

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, $\sum_{j=1}^m r_j = R^2$.

Further, it is required that

- all shares are ≥ 0.

How would you define/calculate relative importance?

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

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

  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).

```
## [1] "missing data!"
```

Note: The R^2 of model (3) 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:

```
## [1] "missing DATA"
```

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

$\Rightarrow$ 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 34).
- Hard to generalize to other, non-linear regression models.

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`.

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

Causality vs correlation

In **explanatory models** the ultimate goal 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 teen 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?

Look at a regression model including covariate x and response y. If the coefficient β_x is "significant", there are several possible reasons for this:

1. x is a **cause** for y. Write: $x \to y$

   Example: x is fish consumption and y is mercury concentration in the urine.

   This is the desired situation!

2. y (partially) causes x, that is $y \to x$.

   Example: x is *IQ* and y is *school education*.

   In that case, the model is not correctly specified!

3. There is another covariate z that both influences x and y

$$z \to x \quad \text{and} \quad z \to y \; .$$

   $\to$ x and y **covary**, but do not cause each other.

In the teen obesity example, all three reasons are possible – perhaps even at the same time!

Ideas:

- No breakfast (x) $\rightarrow$ Obesity (y)
- Obesity (y) $\rightarrow$ No breakfast (x)
- Large dinner (z) $\rightarrow$ Obesity (y)

  and

  Large dinner (z) $\rightarrow$ No breakfast (x)

Many other ideas are possible. . .

In fact, see a recent article in NZZ am Sonntag (temporarily available from OpenEdX):

On the following website you find many “spurious correlations”, where the **causality is very obviously missing**:

http://www.tylervigen.com/spurious-correlations

(More about it in the BC material of this unit!)

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"):

- Observation of subjects / objects in a real-world (existing) situation.
- Variables are usually correlated.
- Often more variables than can be included in the model.
- **Examples**: Influence of pollutans (mercury) on humans, studies of wild animal populations, epidemiological studies,...

Experimental study:

- Observation 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; psychological or pedagogical experiments,...

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

Causality considerations for model building

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: ...

Some further reading on a very recent case of a food researcher that conducted "questionable" science:

- http://www.timvanderzee.com/the-wansink-dossier-an-overview/

From D. Randall & C. Welser (2018). "The irreproducibility crisis of modern science", NAS report.

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 alone.
- The p-value is not "bad", it contains useful information, but it has to be used properly.
  $\rightarrow$ 3 suggestions or alternatives (gradual interpretation of 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.
- Amrhein, V., S. Greenland, and B. McShane (2019). Retire statistical significance. *Nature 567,* 305-307.
- 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.*