

# Lecture 9: Interpretation, causality, cautionary notes

## BIO144 Data Analysis in Biology

Stephanie Muff, Owen Petchey, Erik Willems

University of Zurich

29 April, 2024

# Recap of previous lecture

## 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 some of the literature you will find on OLAT (o.a. “The essential guide to effect sizes”, by Ellis)

# *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 groups
- ▶  $\chi^2$ -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 linear models, the  $p$ -value is often used as an indicator of explanatory model importance. Remember the mercury example:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
age	-0.041	0.013	-3.289	0.001
mother	-1.087	0.373	-2.912	0.004
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000

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 totally **arbitrary** and significance testing solely based on this leads 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 Association* (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 feature in the media (occasionally)

**NZZ am Sonntag** 9. April 2016 | **Wissen** | **61**

## Überschätzte Statistiken

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

**K**linische Studien können nicht mit Menschen, sondern auch statistischen Größen. Das gilt besonders für den sogenannten p-Wert, mit dem jeder Mittelwert verglichen wird. Dies kommt in Kontakt kommt, vor allem aber jeder, der im wöchentlichen Sinn etwas mit Statistik zu tun hat. Inzwischen ist der p-Wert Index auf dem Cover des *New England Journal of Medicine*, der modernen Statistik, der breite Gürtel. Peter Jüni, der seit Kurzem am Applied Health Research Centre der Universität Toronto in Kanada lehrt, warnt: 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 direkt: Der p-Wert misst die Wahrscheinlichkeit, dass etwas dann passiert, was man von ihm erwartet, erklärt der Berater Epidemiologe Peter Jüni, der seit Kurzem am Applied Health Research Centre der Universität Toronto in Kanada lehrt. 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 direkt: Der p-Wert misst die Wahrscheinlichkeit, dass etwas dann passiert, was man von ihm erwartet, erklärt der Berater Epidemiologe Peter Jüni, der seit Kurzem am Applied Health Research Centre der Universität Toronto in Kanada lehrt.

Die Hypothese in einem Patientenstudium könnte zum Beispiel lauten, dass ein Medikament A gegen eine Nierenerkrankung besser wirkt als ein Medikament B. Bei der Hypothese besagt dann das Gegenstell, dass Medikament A nicht besser wirkt als das Medikament B. Beim Test berechnet der forscher die P-Werte, 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 richtig ist. Je geringer er ist, desto weniger spricht für die Null-Hypothese. Ein p-Wert von 0,05 bedeutet, dass es festgestellt werden kann, dass es sich um einen wahrscheinlichen Unterschied handelt. Die Hypothese mit einer Wahrscheinlichkeit von lediglich 5 Prozent zustande kommen kann – und nicht, dass eine wahrscheinliche Hypothese auf einer Sicherheit von 95 Prozent beruht.

Der eigentlich unterschiedliche Hypothesen kann der p-Wert nur indirekt etwas aussagen, weil er über zwei Ecken gedacht ist. Das ist der Grund, warum es keinen umfassenden Beweis für einen gezielteren Unterschied oder Zusammenhang. «Der p-Wert ist eine bedingte und nicht eine absolute Wahrscheinlichkeit», erklärt Peter Jüni. «Doch genau das verhindert, dass man sicher ist, ob seine Aussage richtig ist.» Hinzu kommt, dass die Signifikanzgrenzen von 5 Prozent viel Prozent historisch entstanden sind, um die falsche Signifikanzrate zu begrenzen. «Das ist eine Menge von Falschen geblieben», sagt Leemann Hei vom Institut für Epidemiologie, Biostatistik und Prävention der Universität Zürich.

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

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 «nicht signifikant» zu bezeichnen, sondern im Kontext der gesamten Untersuchung und anhand anderer Ergebnisse zu interpretieren. Genügt hat das herzlich wenig, wie das Team von John Ioannidis und anderen herausgefunden hat. Dennoch sind in den vergangenen 25 Jahren in der biomedizinischen Forschung immer mehr Studien entstanden, die p-Werte angeben, die nicht immer klar aufeinander abfallen. Gleichzeitig werden Zusatzinformationen zu den festgestellten Effekten immer seltener (*JAMA*, Bd. 315, S. 134).

### 5 %

**Kleiner als diese Wert muss der p-Wert in einem statistischen Test sein, dann gelten die Daten aus einer Studie als signifikant. Doch die Grenze ist willkürlich gewählt. (spm)**



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

**Note:** R.A. Fisher, the “inventor” of the  $p$ -value (1920s) didn’t intend the  $p$ -value to be used in the way it is today (which is: doing a single experiment and use  $p < 0.05$  to arrive at 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!

## Essay

# 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.

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.

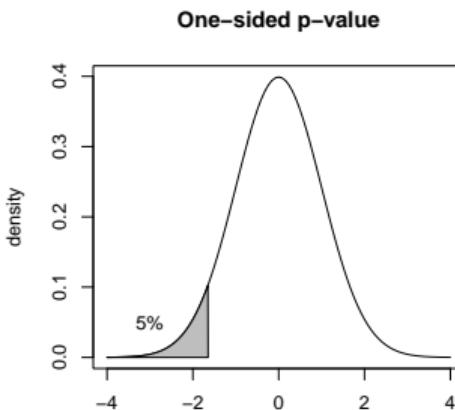
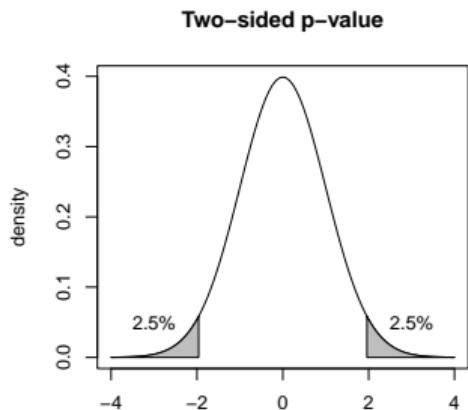
should be interpreted based only on  $p$ -values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations.

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,  $\alpha$ . Assuming that  $c$  relationships are being probed in the field, the expected values of the  $2 \times 2$  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 (FPR). According to the 9

# 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 the *p*-value** is the probability to observe a summary statistic (e.g., an average) that is at least as extreme as the one observed, given that the Null Hypothesis is correct.



## 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 better to have a more **nuanced 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.

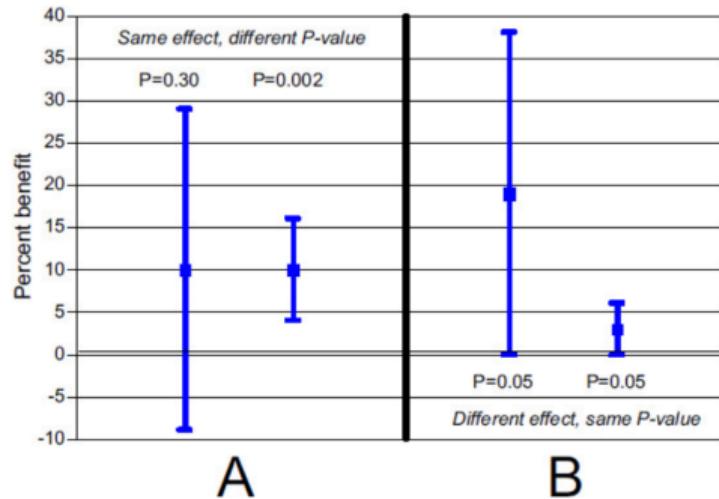
**Q:** Why do so many people still use  $p < 0.05$ ?

**A:** Because that's what they were taught in college or grad school.

# Significance vs. relevance

In linear models:

- ▶ A low  $p$ -value does not automatically imply that a variable is "important".
- ▶ "Is there an effect?" vs. "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!!**

## Some causes of non-informative $p$ -values

Many factors can contribute to large  $p$ -values:

- ▶ Low sample size ( $\rightarrow$  low power).
- ▶ The truth is not "far" from the null hypothesis. (e.g. small effect sizes in linear models)
- ▶ Collinear explanatory variables
- ▶ Incorrect fitting (e.g. non-linear explanatory variables)

## So, shall we just abolish *p*-values?

**No:** *p*-values are not “good” or “bad”. They contain important information, and 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 associated **confidence intervals**
3. One could additionally look at the **relative importance** of explanatory variables
4. **NEVER use *p*-values for model selection**

## Suggestion 1: Graded 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:

## Redefine statistical significance

We propose to change the default  $P$ -value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

(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**.

Another recent suggestion, signed by > 800 researchers:



## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Amrhein et al., 2019, Nature: Do not use the term “statistical significance” at all.

In the Hg example:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
age	-0.041	0.013	-3.289	0.001
mother	-1.087	0.373	-2.912	0.004
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000

- ▶ **Little or no evidence:** Hg soil, vegetables from garden, migration background
- ▶ **Moderate evidence:** Monthly fish consumption
- ▶ **Strong evidence:** Smoking
- ▶ **Very strong evidence:** Amalgam, last fish, interaction between 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, “significantly” increase the probability to get cancer.

- ▶ 50g processed meat per day increases the risk of 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  $\beta_{BMI}$  between 1.61 and 2.03 are **compatible with the observed data**.

However...

- ▶ The choice of the **95% is again arbitrary**. We could also go for 90% or 99% or any other interval!
- ▶ 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 explanatory variables

- ▶ Ultimately, the popularity of  $p$ -values in regression models is based on the wish to judge which explanatory variables 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 relevance** of explanatory variables that measure the proportion (%) of the responses' variability explained by each variable.

## Relative importance: Decomposing $R^2$

**Remember:**  $R^2$  quantifies the proportion of variance explained by **all** explanatory variables in a model

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

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  $r_j \geq 0$ .

## How would you define/calculate relative importance?

- ▶ **Idea 1:** Fit simple models including only one explanatory variable 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 explanatory variable of interest, and then take the **increase** of  $R^2$  as  $r_j$ .

**Problem:** In practice, regressors  $x^{(j)}$  are always correlated *to some extent*, 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

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

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

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \epsilon_i \quad R^2 = 0.10 \quad (1)$$

$$y_i = \beta_0 + \beta_2 x_i^{(2)} + \epsilon_i \quad R^2 = 0.09 \quad (2)$$

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \epsilon_i \quad R^2 = 0.13 \quad (3)$$

**Note:** The  $R^2$  of model (3) with both explanatory variables is much less than the sum of the  $R^2$  from models (1) and (2)!

⇒ The increase of  $R^2$  upon inclusion of an explanatory variable depends on the explanatory variables already in the model!

## A better way to calculate relative importance?

Various solutions to calculate relative importance ( $R^2$  decomposition) have been proposed. The (currently) most useful is given by the following idea, called **LMG** (Lindemann, Merenda and Gold 1980):

- ▶ Fit the model for **all possible orderings of the explanatory variables**.
- ▶ Record the increase in  $R^2$  each time a variable is included.
- ▶ **Average** over all orderings of the explanatory variables.

The R-package '**relaimpo**' (Groemping 2006) contains the function `calc.relimp()` that does this for you.

## Hg results

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

Variable	Rel. imp. (%)	p-value
$\log(Hg_{\text{soil}})$	0.077	0.49
Vegetable	0.29	0.31
Migration	0.52	0.88
Smoking	2.00	0.003
Amalgam	13.56	<0.0001
Age	1.10	0.0013
Mother	1.01	0.0042
Fish	5.94	0.02
Last fish	7.49	<0.0001
Age:mother	6.53	0.0004

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

⇒ Relative importance gives **complementary information** to  $p$ -values, effect sizes and confidence intervals. . .

# Does relative importance solve all problems?

Of course 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 variable last\_fish**.

- ▶ **Model 1:** Relative importance of fish: 5.94% (see slide 31).
- ▶ **Model 2:** Relative importance of fish: 9.07%

**Interpretation:** If two (somewhat) correlated variables feature in the same model, they “share” their chunk of relative importance.

## Causality vs. correlation

In **explanatory models** the ultimate goal is to reveal **causal relationships** between the explanatory variables 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 Asbestos lead to illness or death?
- ▶ ...

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

## Example: Breakfast eating and teen obesity

Please read the NZZ article on OLAT (Unit 9, Lecture 9), and answer the questions below.

### Questions:

- ▶ Does the article show that people that eat breakfast are generally less obese?
- ▶ Does this automatically imply that eating breakfast **leads to** less obesity?

Look at a regression model including explanatory variable  $x$  and response  $y$ . If the coefficient  $\beta_x$  is “significant”, there are several plausible scenario’s:

1.  $x$  is a **cause** for  $y$ . Written as:  $x \rightarrow y$

**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  $y \rightarrow x$ .

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

In that case, the model is not correctly specified!

3. There is another (unmeasured) variable  $z$  that influences both  $x$  and  $y$

$$z \rightarrow x \quad \text{and} \quad z \rightarrow y .$$

$\rightarrow x$  and  $y$  **covary**, but there is no causal relationship between the two.

In the obesity example, all three scenario's are possible, perhaps even at the same time!

**Ideas:**

- ▶ No breakfast ( $x$ ) → Obesity ( $y$ )
- ▶ Obesity ( $y$ ) → No breakfast ( $x$ )
- ▶ Large dinner ( $z$ ) → Obesity ( $y$ )

*and*

Large dinner ( $z$ ) → No breakfast ( $x$ )

Many other scenario's are possible...

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

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

(More about this 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 more likely when the observed association is strong.
2. **Consistency:** A causal relationship is more likely if multiple independent studies show similar associations.
3. **Specificity:** A causal relationship is more likely when an explanatory variable  $x$  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.

However, these criteria are perhaps a little outdated by now, and the study of causality and causal inference from statistical models is a very active field of current research.

## Experimental vs. observational studies

**Experimental studies** are relevant in biology as well as medicine, perhaps especially in the context of clinical trials in which novel medication is tested.

Many obesity studies are **observational**:

- ▶ Study participants self-report on their behaviour.
- ▶ Are often not assigned to a treatment group.
- ▶ There is **no intervention**.

An observed effect is more likely to be *causal* if participants are *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 pollutants (mercury) on humans, studies of wild animal populations, epidemiological studies, ...

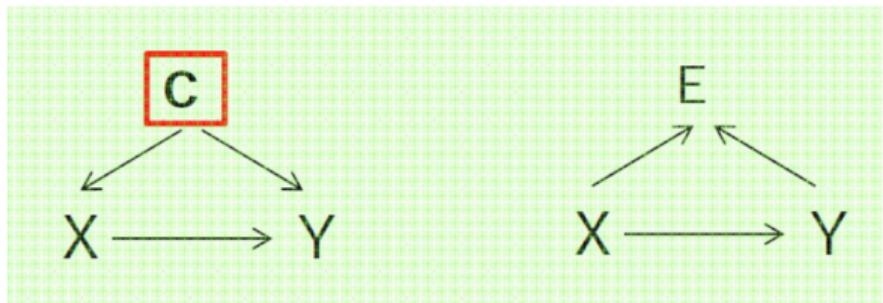
## Experimental study:

- ▶ 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, ...

	<b>Observational study</b>	<b>Experiment</b>
<b>Situation</b>	Existing, cannot be influenced	Artificial, designed
<b>Analysis</b>	Difficult (model selection issues)	Simple <b>no model selection</b>
<b>Interpretation</b>	Difficult, especially w.r.t. causality	Clear, <i>if well-designed</i> causal inference is possible

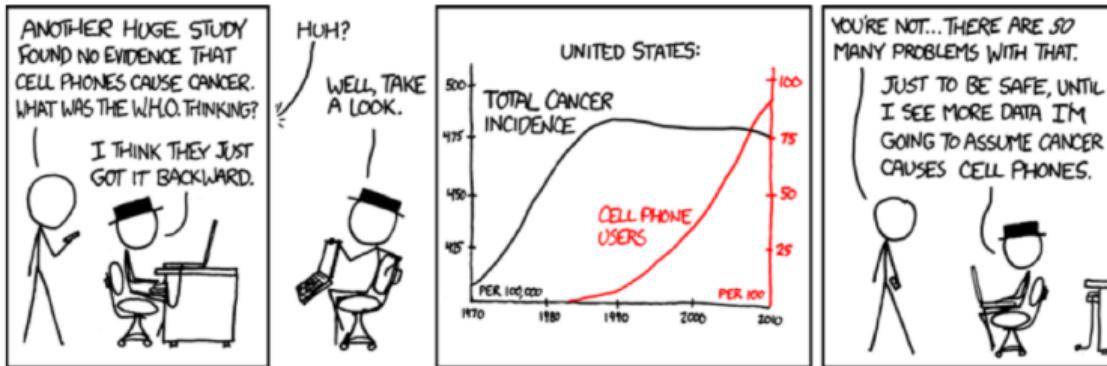
# Causality considerations for model building

It is **widely unknown** that a model can be broken by the inclusion of a “wrong” explanatory variable, which is causally associated in the wrong direction:



**Remember:** Avoid including explanatory variables in your model that are **caused** by the outcome!

**Example:** ...



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 on what it means, instead of only reporting *p*-values
- ▶ The *p*-value is not "bad", it contains useful information, but it has to be used correctly
  - 3 suggestions (gradual interpretation of *p*-values, and report effect sizes, CIs, and relative importance)
- ▶ **Correlation** does not imply **causation**
- ▶ **Experimental studies** are often better suited to infer 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*.