

# *P* Values

## IFSPM Kolloquium

Leonhard Held  
Biostatistics Unit  
IFSPM Zurich

22. September 2010

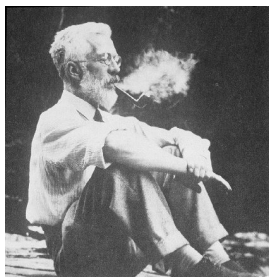


Universität  
Zürich<sup>UZH</sup>

## Definition of a $P$ Value

The  $P$  value is the probability, under the assumption of no effect (the null hypothesis  $H_0$ ), of obtaining a result equal to **or more extreme** than what was actually observed.

R.A. Fisher (1925)



Universität  
Zürich <sup>UZH</sup>

# A More Accessible Definition?

**$p$ -Wert:** Dieser bezeichnet die Wahrscheinlichkeit (probability) der Nullhypothese, d.h. die Wahrscheinlichkeit, dass das beobachtete Resultat durch Zufall zustande kam. Wird in Dezimalbrüchen angegeben (0,05 entspricht 5%).

**Forschung mit Menschen:** Ein Leitfaden für die Praxis (2009)  
Herausgegeben von der Schweizerische Akademie  
der Medizinischen Wissenschaften (SMAW)  
Glossar



## Another Attempt ...

Die von uns heiss geliebten  $p$ -Werte drücken aus, wie wahrscheinlich ein rein zufälliges Zustandekommen der Studienresultate aus statistischer Sicht ist. Für diesen Typ-I- oder “Alphafehler” werden in wissenschaftlichen Studien in der Regel Wahrscheinlichkeiten von 5% oder weniger akzeptiert. Somit dürfte deswegen im Schnitt jedes 20. signifikante statistische Analysenresultat nur zufällig zustande kommen.

Gnädingler and Marty, 2007, 318-324, Schweiz Med Forum



## A Mathematical View ...

Die Wahrscheinlichkeit eines Fehlers 1. Art ist mit dem Signifikanzniveau kontrolliert. [...] Es gibt also ein [Signifikanz-]Niveau, wo  $H_0$  “gerade noch” verworfen wird. Der  $p$ -Wert ist das kleinste Signifikanzniveau wo  $H_0$  verworfen wird.



# Some Recent Headlines ...

Open access, freely available online

Essay

## Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine 2005, 696-701

## Signifikanter Nonsens

Die meisten Studien seien falsch, sagt der Epidemiologe John Ioannidis. Denn trotz korrekter Statistik schummeln Forscher an anderer Stelle. Untersuchungen über Nutzen und Risiken von Lebensmitteln stimmen fast nie. *Von Andreas Hirstein*

20. Mai 2007, NZZ am Sonntag



Universität  
Zürich <sup>UZH</sup>

# The Earth is Round ( $p < .05$ )



Cohan (1994), 997-1003, *Am Psychologist*



Universität  
Zürich<sup>UZH</sup>

# Fundamental Controversies

Ronald Fisher  
(1890-1962)



Jerzy Neyman  
(1894-1981)



# A Brief History of the $P$ Value

- ▶ Fisher suggested the  $P$  value as an informal measure of **statistical evidence**.
- ▶ Neyman dismissed the  $P$  value as a measure of evidence and proposed the formal **hypothesis test** framework based on **error rates**.
- ▶ These two methods are **incompatible** but mistakenly regarded as part of a single, coherent approach to statistical inference.



# 1. Is the $P$ Value the Probability of the Null Hypothesis?

**$p$ -Wert:** Dieser bezeichnet die **Wahrscheinlichkeit (probability) der Nullhypothese**, d.h. die Wahrscheinlichkeit, dass das beobachtete Resultat durch Zufall zustande kam. Wird in Dezimalbrüchen angegeben (0,05 entspricht 5%).



*“The  $P$  value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false.”*



## 2. Is the $P$ Value the Probability that the Observed Data Occurred by Chance?

**$p$ -Wert:** Dieser bezeichnet die Wahrscheinlichkeit (probability) der Nullhypothese, d.h. die **Wahrscheinlichkeit, dass das beobachtete Resultat durch Zufall zustande kam.** Wird in Dezimalbrüchen angegeben (0,05 entspricht 5%).



*“The  $p$ -value is not the probability that the observed data occurred by chance, a circumlocution that actually means the probability of the null hypothesis.”*



### 3. Is the $P$ Value the Type-I Error Rate?

Die Wahrscheinlichkeit eines Fehlers 1. Art ist mit dem Signifikanzniveau kontrolliert. [...] Es gibt also ein [Signifikanz-]Niveau, wo  $H_0$  “gerade noch” verworfen wird. **Der  $p$ -Wert ist das kleinste Signifikanzniveau wo  $H_0$  verworfen wird.**



*“The interpretation of the  $P$  value as a form of post hoc type I error rate (“observed type-I error”) creates the powerful illusion that an error rate ( $\alpha$ ) and a measure of inferential strength ( $p$ ) are identical.”*



## Back to Fisher ...



*"It is to be feared that the principles of Neyman and Pearson's 'Theory of Testing Hypotheses' are liable to mislead those who follow them into much wasted effort and disappointment, and that its authors are not inclined to warn students of these dangers."*

Fisher (1959) *Statistical Methods and Scientific Inference*



Universität  
Zürich <sup>UZH</sup>

## 4. Is the $P$ Value the False Discovery Rate?

Die von uns heiss geliebten  $p$ -Werte drücken aus, wie wahrscheinlich ein rein zufälliges Zustandekommen der Studienresultate aus statistischer Sicht ist. Für diesen Typ-I- oder “Alphafehler” werden in wissenschaftlichen Studien in der Regel Wahrscheinlichkeiten von 5% oder weniger akzeptiert. **Somit dürfte deswegen im Schnitt jedes 20. signifikante statistische Analysenresultat nur zufällig zustande kommen.**



# Epidemiology of Clinical Trials

- ▶ Suppose 200 trials are performed, but only 10% are of truly effective treatments.
- ▶ Suppose each trial is carried out with a type I error of 5% and a type II error of 20%.

Trial conclusion	Treatment		Total
	truly ineffective	truly effective	
not significant	171	4	175
significant	9	16	25
Total	180	20	200

- ▶  $9/25 = 36\%$  of trials with significant results are in fact of truly ineffective treatments!
- ▶ In diagnostic testing terms, the **positive predictive value** is only 64%.



# Summary

## The $P$ value

- ▶ is **not** the probability of the null hypothesis.
- ▶ is **not** the probability that the observed data occurred by chance,
- ▶ is **not** the probability that you will make a type-I error if you reject the null hypothesis,
- ▶ is **not** the false discovery rate.



# Summary

## The $P$ value

- ▶ is **not** the probability of the null hypothesis.
- ▶ is **not** the probability that the observed data occurred by chance,
- ▶ is **not** the probability that you will make a type-I error if you reject the null hypothesis,
- ▶ is **not** the false discovery rate.



# Summary

## The $P$ value

- ▶ is **not** the probability of the null hypothesis.
- ▶ is **not** the probability that the observed data occurred by chance,
- ▶ is **not** the probability that you will make a type-I error if you reject the null hypothesis,
- ▶ is **not** the false discovery rate.



# Summary

## The $P$ value

- ▶ is **not** the probability of the null hypothesis.
- ▶ is **not** the probability that the observed data occurred by chance,
- ▶ is **not** the probability that you will make a type-I error if you reject the null hypothesis,
- ▶ is **not** the false discovery rate.



# Summary

## The $P$ value

- ▶ is **not** the probability of the null hypothesis,
- ▶ is **not** the probability that the observed data occurred by chance,
- ▶ is **not** the probability that you will make a type-I error if you reject the null hypothesis,
- ▶ is **not** the false discovery rate.



*“In fact, the  $P$  value is almost nothing sensible you can think of. I tell students to give up trying.”*



# Conclusion I



*“The hypothesis test approach offered scientists a Faustian bargain [i.e. a deal with the devil] – a seemingly automatic way to limit the number of mistaken conclusions in the long run, but only by abandoning the ability to measure evidence and assess truth from a single experiment.”*



## Conclusion II



*“Since  $P$  values are not likely to soon disappear from the pages of medical journals or from the toolbox of statisticians, the challenge remains how to use them and still properly convey the strength of evidence provided by research data.”*



# What a Mess. Can These Men Help?



The Wolf: I'm Winston Wolfe. I solve problems.

Jimmie: Good, we got one.

The Wolf: So I heard. May I come in?

Jimmie: Uh, yeah, please do.

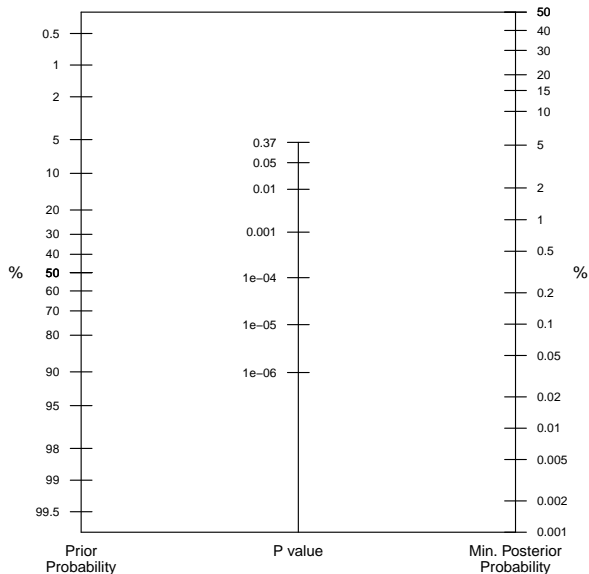


# Calibration of $P$ Values







- ▶ A calibration of  $P$  values to **minimum posterior probability of the null hypothesis** is possible.
- ▶ A universal finding is that the evidence against a simple null hypothesis is by far not as strong as the  $P$  value might suggest.
- ▶ However, the acceptance of this calibration in clinical fields is low due to
  - ▶ complicated calculations,
  - ▶ the need to specify a **prior probability** of the null hypothesis.



# A Nomogram for $P$ Values



# References

-  Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference in Psychological Research," *Psychological Review*, 70, 193–242.
-  Goodman, S. N. (1999a), "Towards Evidence-Based Medical Statistics. 1.: The  $P$  Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.
-  — (1999b), "Towards Evidence-Based Medical Statistics. 2.: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.
-  — (2005a), "Introduction to Bayesian methods I: measuring the strength of evidence," *Clin Trials*, 2, 282–290.
-  — (2005b), " $P$  Value," in "Encyclopedia of Biostatistics," Chichester: Wiley, pp. 3921–3925, second edition.
-  Held, L. (2010), "A nomogram for  $P$  values," *BMC Med Res Methodol*, 10, 21.

