Fisher, Neyman and Bayes: Part I

Philosophical excursion

André Meichtry

Departement of Health Professions
Bern University of Applied Sciences

April 5, 2023

1. Probability
   - Uncertainty versus long-run frequency

2. The Frequentist Approach
   - Fisher and Neyman-Pearson
   - Null Hypothesis Significance Test Procedure NHST

Probability

- A measure of uncertainty (very general)
- A measure of long-run frequency (classical statistics)

Axioms of probability

- Events or propositions A and B:
  1. Non-negativity: $\Pr(A) \geq 0$ for any event A
  2. Certain event: $\Pr(\textit{certain event}) = 1$
  3. $\Pr(A \textit{ or } B) = \Pr(A) + \Pr(B)$ for disjoint A and B
- Very simple! All Bayesian statistics is based on these axioms.

Subjective probability

- We can think of probability as a measure of degree of belief.
- This is not thought of as something measured by strength of feeling, but in terms of betting behaviour.

Subjective probability

For me to give 0.7 degree of belief to there being rain tomorrow is, roughly:

- for me to regard 0.7 units as the fair price for a bet
- that returns
  - 1 unit if it rains tomorrow
  - and nothing if it does not.

Subjective probability

- Ramsey, de Finetti, Savage, etc.
- Measuring the evidence in favour of a proposition A
- How much would **You** bet about the truth of A?
- What odds O are **You** willing to give or receive for a fair bet?
- **Your** probability

$$\Pr(A) = \frac{O}{1+O}$$

Coherence and rational behavior*

- **Your** odds $O = 2 : 8$, so probability=0.2.
- **You** are willing to give 2, receive 8 (if A turns out to be true).
- Expected gain: $0.2 \cdot (+8) + 0.8 \cdot (-2) = 0$.

When the expected gain is zero, we have a fair bet, and this definition of probability assumes rational behavior.

Coherence and rational behavior*

Assume that an expert knows that the success probability of his therapy is $p = 0.6$.

- Scenario 1:
  - He bets $O = 9 : 1$ overstating the effect.
  - Expected gain: $0.6 \cdot (+1) + 0.4 \cdot (-9) = -1$
- Scenario 2:
  - He bets $O = 3 : 7$ understating the effect.
  - Expected gain: $0.6 \cdot (+7) + 0.4 \cdot (-3) = 3$
- Better for him to be coherent.

When the expected gain is zero, we have a fair bet, and this definition of probability assumes rational behavior.

Fisherian test of significance

Inductive evidence

- Only one hypothesis, the "null", H_0, the hypothesis "to be nullified"
- "Proof" by contradiction (not absolute). Inference. Model validation.
- Fundamental quantity: A posteriori p-value quantifying the evidence against the null from a single experiment.
- p represents the probability of seeing something as weird or weirder than you actually saw, if the null is true. No sampling interpretation.
- α is secondary! and technically a decision rule.

Example: Fisherian test of significance

Probability of data x under some parameter $\theta = \theta_0$, that is, under the null model, $p(x \mid \theta = \theta_0)$:

x	1	2	3	4
$p(x \mid \theta = \theta_0)$	.980	.005	.005	.010
p-value	1	.01	.01	.02

Table: Probability distribution of X under H_0

An $\alpha = 0.01$ Fisherian Test of $H_0 : \theta = \theta_0$ rejects for $x = 2, 3$, with p-value$= 0.01$ in each case.

Neyman-Pearson hypothesis test

Inductive behavior

- Additionally: alternative hypothesis H_A and the concept of power.
- Based on a priori fixed long run error rates, *Type I* and *Type II*. [1]
- The most powerful test at a specified α-level is the one maximizing the likelihood (Neyman-Pearson Lemma[2]).
- Roots in deductive philosophy and mathematics.
- Decision problem.
- $(1-\alpha)$-"confidence regions" as the long run probability of these regions including the true parameter.

[1] $\alpha = \Pr(\textit{reject } H_0 \mid H_0) \quad \beta = \Pr(\textit{not reject } H_0 \mid H_A)$
[2] Fundamentallemma der mathematischen Statistik

Neyman-Pearson hypothesis test

Probabilities $p(x \mid \theta)$ under $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_1$

x	1	2	3	4
$p(x \mid \theta = \theta_0)$	.980	.005	.005	.010
$p(x \mid \theta = \theta_1)$	.098	.001	.001	.900
Likelihood Ratio LR [3]	.1	.2	.2	90

Table: Probability distribution of X under H_0 and H_A

- The most powerful (or maximal likelihood ratio) $\alpha = 0.01$ NP-test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta = \theta_1$ rejects for $x = 4$.
- Result is different from the Fisher test!

[3] $LR = \frac{L(\theta_1)}{L(\theta_0)} = \frac{p(x|\theta_1)}{p(x|\theta_0)}$

Neyman-Pearson hypothesis test

Probabilities $p(x \mid \theta)$ under $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_1$

x	1	2	3	4
$p(x \mid \theta = \theta_0)$	.980	.005	.005	.010
$p(x \mid \theta = \theta_1)$	.098	.001	.001	.900
Likelihood Ratio LR [4]	.1	.2	.2	90

Table: Probability distribution of X under H_0 and H_A

- The rejection region for the $\alpha = 0.02$ NP-test of includes $r = 2, 3$, even though 2 and 3 are five times more likely under the null hypothesis than under the alternative.

[4] $LR = \frac{L(\theta_1)}{L(\theta_0)} = \frac{p(x|\theta_1)}{p(x|\theta_0)}$

Neyman-Pearson hypothesis test

Probabilities $p(x \mid \theta)$ under $H_0 : \theta = \theta_0$ and $H_{A2} : \theta = \theta_2$

x	1	2	3	4
$p(x \mid \theta = \theta_0)$	.980	.005	.005	.010
$p(x \mid \theta = \theta_2)$	.100	.200	.200	.500
Likelihood Ratio LR	.1	40	40	50

Table: Probability distribution of X under H_0 and H_{A2}

- NP testing cannot appeal to the idea of proof by contradiction!
- The most powerful $\alpha = 0.01$ NP test would reject for $r = 4$, even though $r = 4$ is the most probable value for the data under the null hypothesis!

First Bayesian intermezzo: From Prior to Posterior

			x=1	x=2	x=3	x=4
	θ		Likelihood: $p(x \mid \theta)$			
	θ_0		.980	.005	.005	.010
	θ_1		.098	.001	.001	.900
Prior odds	θ	Prior prob: $p(\theta)$	Posterior: $p(\theta \mid x)$			
1:1	θ_0	1/2	.91	.83	.83	.01
	θ_1	1/2	.09	.17	.17	.99

Table: Posterior probabilities with uninformative prior odds. Decision based on higher posterior probability.

Simple versus composite hypothesis*

Assume the parameter space $\Theta = \{\theta_0, \theta_1, \theta_2\}$. We want to test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$

x	1	2	3	4
$p(x \mid \theta = \theta_0)$	.980	.005	.005	.010
$p(x \mid \theta = \theta_1)$	.098	.001	.001	.900
$p(x \mid \theta = \theta_2)$	.100	.200	.200	.500

Table: Probability distribution of X under H_0 and H_A

- Because the most powerful tests of the alternatives $H_A : \theta = \theta_1$ and $H_A : \theta = \theta_2$ are identical ($x = 4$), this is the uniformly most powerful (UMP) $\alpha = 0.01$-test.
- Fisher: not forbidden to test individually different null models: $H_0 : \theta = \theta_0, \quad H_0 : \theta = \theta_1, \quad H_0 : \theta = \theta_2$

Beyond UMP*

- UMP tests exist for one-parameter models from exponential family (i.e. one-sided t-test)
- UMP tests do not exist for two-sided tests and vector parameters.
- The lack of availability of UMP tests has led to the search for tests under less stringent requirements of optimality.
  - Likelihood Methods:
    - Locally most powerful tests, score test (most powerful for small deviations)
    - Generalized Likelihood ratio test
    - Wald-Test
  - Many others...

Null Hypothesis Significance Test Procedure (NHST)

- A combined approach has emerged.
- One follows Neyman-Pearson formally, but Fisher philosophically.
- *p*-values are measures of evidence and long run error rates.
- Planning of experiments: more Neyman-Pearson; analysis stage, observational studies: more Fisherian.
- The initial protagonists of the approaches would never have accepted today's practice...
- The distinction between evidence (*p*-values) and error (α's) were not semantic sophistry for Fisher and NP!

Null Hypothesis Significance Test Procedure (NHST)

- (Apparent) separation of evidence from subjective factors.
- Ease of computation, availability of software.
- “Wide acceptability” and “established criteria” for “significance”.
- (Apparent) relevance for regulatory agencies.

What humans – by nature – ask for

Definition (p-value)

The p-value is the probability that any value of a statistic generated from the null hypothesis according to the intended sampling process has magnitude greater than or equal to the magnitude of the observed value of the statistic. [a]

[a] $\Pr(T \geq t \mid H_0)$, for a test statistic T and observed statistic t.

- That is a conditional probability of data, given an hypothesis.
- Does not reply to the very question human minds by nature ask for, the probability of H_0, given observed data.

Why attacking a straw-man?

Philosophy of Science

June, 1967

THEORY-TESTING IN PSYCHOLOGY AND PHYSICS: A METHODOLOGICAL PARADOX*

PAUL E. MEEHL[1]

Minnesota Center for Philosophy of Science

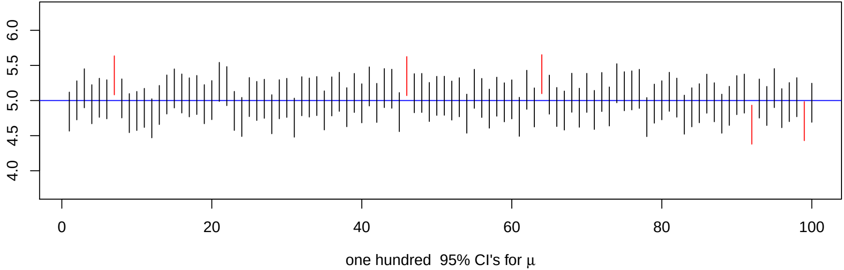
- Theories are expressed very weakly, confirmed by “any” magnitude of increase.
- “Statistical significance” plays a logical role in psychology precisely the reverse of its role in physics.
- Reason: Straw-man argument, nil-nulls such as H_0 : "$Effect = 0$", "$Correlation = 0$" etc.

p-values do not depend only on data*

- *p*-values depend on sampling intentions.
- NHST has 100% false alarm rate in sequential testing. sampling to reach a foregone conclusion (e.g., Anscombe, 1954).
- *p*-values violate the so called likelihood principle: all information from the data should be in the likelihood function. [5]
- *p*-values are inherently subjective!

[5] $L(\theta) = p(x \mid \theta)$

Intermediate solution: confidence intervals



- A 95% CI on a parameter is the range of parameter values that would not be rejected at $\alpha = 0.05$ by the observed data.
- They do not carry distributional information.
- Nevertheless, people – almost invariably – interpret "confidence" as Bayesian posterior probability.

References

[Chr05] Ronald Christensen. Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59(2):121–126, 2005.

[GCS+13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 3 edition, 11 2013.

[KM15] John K. Kruschke and Mike Meredith. *BEST: Bayesian Estimation Supersedes the t-Test*, 2015. R package version 0.3.0.

[Kru10] John Kruschke. An open letter. `http://www.indiana.edu/~kruschke/AnOpenLetter.htm`, 2010. Accessed: 2015-09-30.

[Kru14] John Kruschke. *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press, 2 edition, 11 2014.

[R C22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[SAM04] David J. Spiegelhalter, Keith R. Abrams, and Jonathan P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, 1 edition, 1 2004.

[ZM08] Stephen T. Ziliak and Deirdre N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (Economics, Cognition, and Society)*. University of Michigan Press, 1st edition, 2 2008.