P-value: use and misuse

Daniel Commenges

INSERM, Bordeaux Population Health Research Center, Biostatistics Team, Bordeaux
http://sites.google.com/site/danielcommenges/

6 November 2017

Organization of the talk

1. ASA statement
2. Theory of tests
3. Definition of p-value
4. Misuse
5. Confidence intervals
6. causal interpretation

The ASA statement on p-value

- The ASA's statement on p-value: context, process and purpose (editorial)
- ASA statement on statistical significance and p-value (The American Statistician, 2016, 70, 129-1633).
- 21 comments (supplemental material)

What is a p-value ?

- a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value
- 1. P-values can indicate how incompatible the data are with a specified statistical model.

Caveats

- 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
- 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold
- 4. Proper inference requires full reporting and transparency
- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
- 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

Classical theory of tests

- A test looks whether the data are compatible with a given hypothesis H_0
- it is based on a test statistic, T and a rejection region R_α
- a decision of rejection of H_0 is taken with $P_{H_0}(T \in R_\alpha) = \alpha$
  α is the Type I risk (or level, or size) of the test
- T is chosen to be sensitive to an alternative hypothesis H_1, so that the power, $P_{H_0}(T \in R_\alpha) = 1 - \beta$, is as high as possible

- Most remarks addressed to p-values are in fact addressed to significance testing
- The size of the test is not a p-value: it is decided in advance

Dissymmetry between H_0 and H_1

- Most often we have a model $\mathcal{M}_0$ and a model $\mathcal{M}_1$ such that $\mathcal{M}_0 \subset \mathcal{M}_1$
- $H_0 = \mathcal{M}_0$ and $H_0 = \mathcal{M}_1 / \mathcal{M}_0$
- Typically H_1 implies one or more additional parameters
- A decision to "reject" H_0 can be taken. The decision itself is a binary random variable D_α with $P_{H_0}(D_\alpha = 1) = \alpha$

It may be questioned whether "rejection or not" is a true "decision". In real life, decisions may be "publish or not a result" or "accept to deal a drug in the market"; not purely statistical decisions.

The null bias

Bias in favor of the null

- From the point of view of the manufacturer the error in asserting the carcinogenicity of A is more important to avoid than the error in asserting that A is harmless. Thus, for the manufacturer of A, the hypothesis tested may be:
  "A is not carcinogenic"
- For the prospective user of chemical A the hypothesis tested will be:
  "A is carcinogenic"
  We wish that the probability of error in rejecting this hypothesis be reduced to a very small value!

adapted from Neyman (1977) cited by Greenland

Definition of the p-value

Definition

- The smallest α for which the α-test would be significant
- p-value$= P_{H_0}(T > t_{obs})$
- Because t_{obs} is random, p-value is a random variable: under H_0 it has a uniform distribution on $[0, 1]$

Origin of the p-value

R Fisher: Statistical Methods for Research workers (1925)

"In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05, and consider that higher values of χ^2 indicate a real discrepancy."

Size-α tests or p-value ?

- The decision D_α is a dichotomization of the p-value
- Often practitioners use a categorisation of the p-value with cut-off points $0.05; 0.01; 0.001$
- Any categorisation results in a loss of information: 0.051 and 0.049 are closer than 0.049 and 0.0015

- Many methods are given in terms of controlling the Type-one error
- There are some pitfalls in manipulating p-values

Size-α tests or p-value ?

- The decision D_α is a dichotomization of the p-value
- Often practitioners use a categorisation of the p-value with cut-off points $0.05; 0.01; 0.001$
- Any categorisation results in a loss of information: 0.051 and 0.049 are closer than 0.049 and 0.0015

- Many methods are given in terms of controlling the Type-one error
- There are some pitfalls in manipulating p-values

Multiplying p-values ?

The case of Lucia de Berk

In 2003, she was sentenced to life imprisonment for four murders and three attempted murders of patients in her care. In 2004, after an appeal, she was convicted of seven murders and three attempts. Her conviction was controversial in the media and amongst scientists. In October 2008, the case was reopened by the Supreme Court. De Berk was freed, and her case was re-tried; she was exonerated in April 2010.

Flawed statistical analysis

"Elffer's method of combining wards by multiplying p-values is blatantly incorrect, since data from a large enough number of wards would make any nurse eventually guilty. Worse still, by disaggregating a fixed amount of data, the p-value can also be made almost arbitrarily small." (From Richard Gill: Ellfers method and Ellfer's mistake)

The intuitive Bayesian interpretation

- P-value does not measure the probability that the studied hypothesis is true
- This is because p-value is a frequentist quantity
- The probability of an hypothesis is a Bayesian concept
- In the Bayesian approach $P(\beta = 0)$ raises a problem: either it is zero or we have to put a mass on it: rather compute $P(\beta \in (a, b))$
- This may be related to a paradox of the test theory: one may argue that "$\beta = 0$" is never exactly true !

The issue of assumptions for computing the p-value

- The p-value is a measure of how much the observation are compatible with the model under H_0; the model makes other assumptions
- Even permutation tests make the exchangeability assumption which is stronger than it seems
- Example: Poisson model in presence of overdispersion: example of influence of phase of the moon on birth rate

The issue of the size of the effect

- The p-value does not indicate the magnitude of the effect
- A small p-value can be associated to a small effect if the sample size is large
- Inversely, a non-significant p-value does not prove that the effect is small

Estimation

- If a test of a zero effect is significant, then one has to look at the estimate of the effect $\hat{\beta}$
- A standard error is also estimated
- A confidence interval for the parameter can be constructed: $(\hat{\beta}_{inf}; \hat{\beta}_{sup})$

Tests and confidence intervals

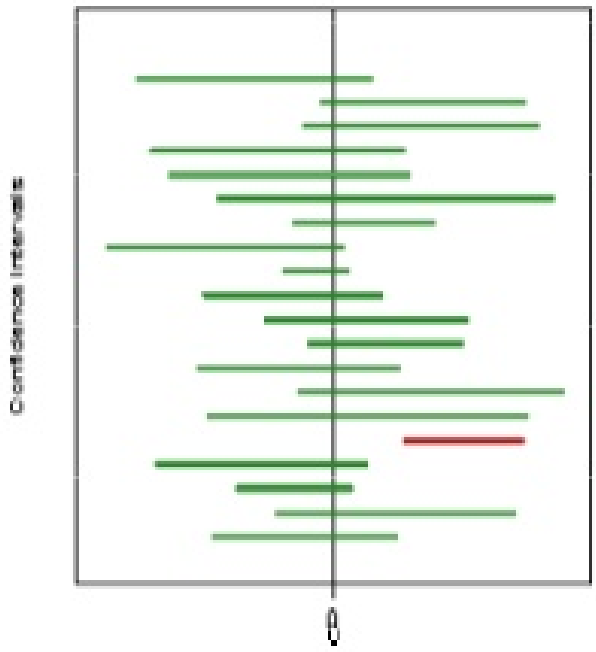
- No antinomy between tests an confidence intervals
- A $(1-\alpha)$-confidence interval is the set of all values of the parameters which are not rejected at the α-level: inversion of a test
- The confidence interval $(\hat{\beta}_{inf}; \hat{\beta}_{sup})$ is the set of values not rejected by the Wald test

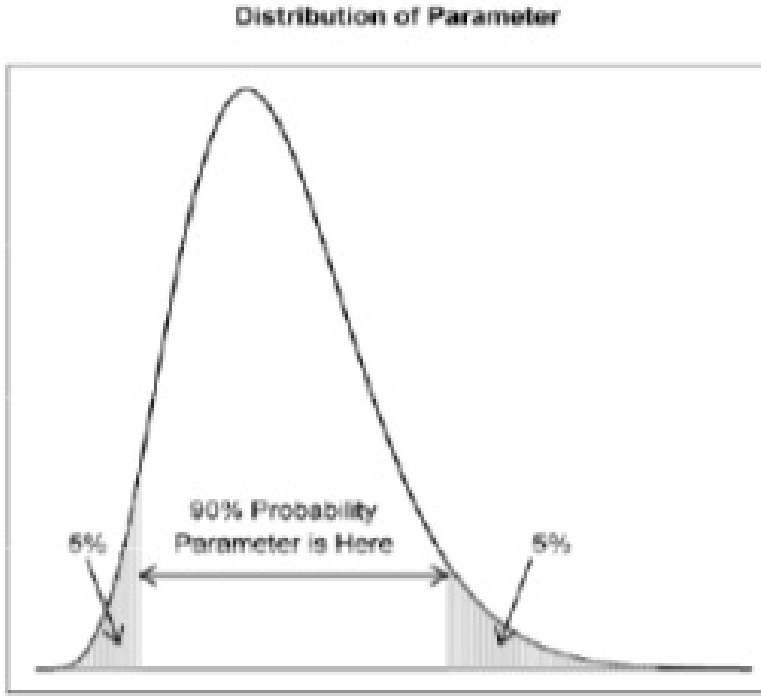
Confidence intervals and credible intervals

- Bayesian credible intervals: $P(\beta \in (\beta_{inf}; \beta_{sup})) = 1 - \alpha$
- Gives the probability that β^* is in this interval
- A proportion $1 - \alpha$ of $(\hat{\beta}_{inf}; \hat{\beta}_{sup})$ contains β^*
- With large *n* the two intervals tends to be the same

Confidence vs. Credibility Intervals

- Frequentist: A collection of intervals with 90% of them containing the true parameter
- Bayesian: An interval that has a 90% chance of containing the true parameter.





Distribution of $\hat{\beta}$, posterior distribution

- Distribution of $\hat{\beta}$
- Posterior distribution
- With large *n* both distributions tend to the same normal distribution
- In practice the estimated distribution of $\hat{\beta}$ and the posterior distribution of β will be very similar for *n* large

The multiplicity issue

- Family-wise type-one error: probability to reject one null hypothesis if all are true; m independent tests:
  $P_{H_0}(\textit{rejection}) = 1 - (1 - \alpha)^m$
  for $m = 10$ $P_{H_0}(D_{0.05} = 1) = 0.40$
- Bonferroni rule: make each test at level α/m: conservative (but not so much for moderately large m)
- other methods may be less conservative (Holm)
- or take into account the correlation between tests: example of interim analyses by Armitage
- Caveat: computing correctly very small p-value is not easy (asymptotic results, simulations)

Transforming size α tests into p-value

- The p-value is the smallest α at which H_0 is rejected
- Bonferroni rule for a test of H_0 "all individual hypotheses are true": make each test at level α/m
- The p-value for rejecting H_0 is $m\times$ p-value$_{min}$

Sources of multiplicity

The good old time

- In an epidemiological study many associations are tested
- Several cutoff points of an explanatory factor are tested
- In a clinical trial multiple outcomes, multiple side effects are tested
- In a clinical trial interim analyses are done

The big data era

- Genomic analysis: linkage studies, Genome-wise association studies (GWAS), micro-arrays, Next-generation sequencing...
- all omics

Managing multiplicity

Controlling multiplicity ?

- Avoid it: distinction between exploratory and confirmatory analyses
- Do nothing ; tell it or not ! A major factor of non-reproducibility of results
- Control the family-wise type-one error
- For high multiplicity, control the false discovery rate (FDR) (Benjamini-Hochberg, 1995): proportion of rejection of true null hypotheses among rejected
- The Benjamini-Hochberg procedure uses p-values (of the different hypotheses) but does result in global a p-value

Causal interpretation

- Correlation is not causation
- In randomized trials a causal interpretation is possible, but there are limitations to randomized trials
- A statistically established correlation between two factors may indicate a causal mechanisms underlying it
- The Bradford Hill criteria may help for causal interpretation; the size of the effect is a major element

Conclusion

- The p-value is an indicator of whether the observations are incompatible with an hypothesis, provided other assumptions are true:
  check the assumptions
- Neither small nor large p-values give directly an indication on the size of an effect:
  check the size of the effect by confidence intervals
- Multiplicity is a major problem is test theory:
  limit it or control the family-wise error or the FDR