P-arvo ja tilastollinen merkitsevyys – onko niillä tulevaisuutta?

21.5.2019

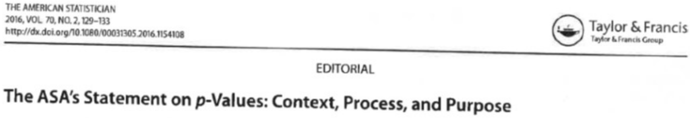
Tilastotieteen keskus/Turun yliopisto
Kari Auranen

Sisältö

- P-arvon määritelmiä ja tulkintoja
- Ongelmia ja väärintulkintoja
- ASA statement on statistical significance and P values
- Keskustelua

Teesi: Tieteellisen näytön tilastollisiin menetelmiin perustuva arviointi pitää vapauttaa dikotomisesta päätöksenteosta ja erityisesti ns. nollahypoteesin testaamisesta.

It doesn't matter if the p-value doesn't mean what the people think it means; it becomes valuable because of what it buys (Goodman)

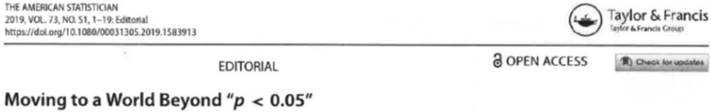


THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose



THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

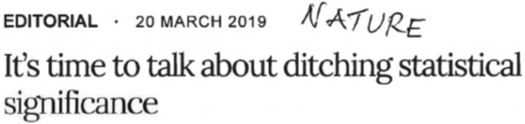
Taylor & Francis
Taylor & Francis Group

EDITORIAL

OPEN ACCESS

Check for updates

Moving to a World Beyond "$p < 0.05$"



EDITORIAL · 20 MARCH 2019 NATURE

It's time to talk about ditching statistical significance

P-arvon määritelmiä (1)

- In statistical hypothesis testing, the p-value or probability value or significance is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary (such as the absolute value of the sample mean difference between two compared groups) would be greater than or equal to the actual observed results. Wikipedia
- Informally, a p value is the probability under a specified statistical model that a statistical summary of the data (e.g.,the sample mean difference between two compared groups) would be equal to or more extreme than its observed value. Wasserstein et al. 2016

Määritelmiä (2)

- P value is the probability of having observed our data (or more extreme data) when the null hypothesis is true (Altman)

  > If P lies below the chosen cut-off point, we reject the null hypothesis and accept a complementary alternative hypothesis. If the P value exceeds the critical values we do not reject the null hypothesis. However, we cannot say that we believe the null hypothesis is true, but only that there is not enough evidence to reject it.

- We imagine a large number of repetitions of the study with the parameter equal to its null value and define the p-value as the proportion of these studies which provide less suppport for the null value than the data actually observed (Clayton and Hills)

  > If the p-value is small that data are at odds with the null hypothesis and the finding is said to be statistically significant

Määritelmiä (3)

- An upper one-sided P-value is the probability under the test hypothesis that the corresponding quantity computed from the data, the *test statistics* (such as a *t*-statistics or a χ^2 statistics), would be greater or equal to its observed value, assuming there are no sources of bias in the data collection or analysis processes.

  A P value is a continuous measure of the compatibility between an hypothesis and data. Although the utility as such a measure can be disputed, a worse problem is that it is often used to force a qualitative decision about rejection of an hypothesis

Rothman ja Greenland: Modern Epidemiology

Määritelmiä (4)

- P-arvo on todennäköisyys, että testisuure saa havaitun tai vielä enemmän vastahypoteesiin viittaavan arvon, kun nollahypoteesi on tosi. Merkitsevyystestien avulla pyritään saamaan selville, onko aineisto yhteensopiva testattavan nollahypoteesin kanssa vai tarjoaako vastahypoteesi havainnoille paremman selityksen. Tyyppi I virheen (α-virheen) todennäköisyys ilmoitetaan lähes kaikissa tutkimuksissa laskemalla P-arvo. (Uhari ym.)

- Hylkäämisvirheen (tyypin 1 virheen eli väärän positiivisen päätöksen) todennäköisyydestä käytetään nimitystä p-arvo (käytetään myös merkintöjä significance ja sig.).

  Saatu p-arvo osoittaa, kuinka suuri on väärän johtopäätöksen todennäköisyys, jos nollahypoteesi hylätään. Tämän voidaan ajatella myös olevan todennäköisyys sille, että vaihtoehtoinen hypoteesi on väärä. (Nummenmaa ym.)

Lähtökohta

- “The P value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it” (Greenland)
  - Measure of compatibility of the observed data with a specified statistical model: the smaller the p value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold.
- P-arvon määritelmä ei sisällä ajatusta testaamisesta
  - Ns. (nolla)hypoteesin testaaminen on silti täysin tavanmukaista (”null hypothesis significance testing”)
  - Havaittu ero on ”tilastollisesti merkitsevä” (statistically significant), jos $p < 0.05$

Ongelmia ja väärintulkintoja

- Tulkinnan riippuvuus otoskoosta
- Väärät tulkinnat*:
  - P-arvo on todennäköisyys, että (nolla)hypoteesi on totta
  - Nollahypoteesia vastaava P-arvo on todennäköisyys, että havaittu ero on syntynyt vain sattumalta ("chance alone produced the observed association")
  - Jos P-arvo on pieni, testattava nollahypoteesi ei ole totta ja pitää hylätä
  - Jos P-arvo on suuri, testattava hypoteesi on totta ja se pitää hyväksyä
- Merkitsevyystason käyttö sekä p-arvon että tyypin I virhetason (α-tason) nimityksenä
- Tilastollisen evidenssin esittämisen ja dikotomisen päätöksen teon sekoittuminen/sekoittaminen

$\ast$ Greenland ym.: Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. 2016

Otoskoon merkitys

- Isommassa tutkimuksessa saatu p-arvo on voimakkaampaa näyttöä nollahypoteesia vastaan
  - “A given P-value in a large trial is usually stronger evidence that the treatments really differ than the same P value in a small trial of the same would be” (Peto et al., 1976)
- Pienemmässä tutkimuksessa saatu p-arvo on voimakkaampaa näyttöä nollahypoteesia vastaan

> ”the interpretation to be placed on the phrase ‘significant at 5%’ depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large sample” (Lindley and Scott, 1984)

> The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect ... and if the p-value is the same, the probability of committing a Type I error remains the same. Thus can be more confidence with a small N than a large N (Bakan, 1970)

Royall 1986

I don’t mean to sound critical, but I am

at the edge of significance (p=0.055)

at the limit of significance (p=0.054)

at the limits of significance (p=0.053)

at the margin of significance (p=0.056)

at the margin of statistical significance ($p < 0.07$)

at the verge of significance (p=0.058)

at the very edge of significance (p=0.053)

barely below the level of significance (p=0.06)

barely escaped statistical significance (p=0.07)

barely escapes being statistically significant at the 5% risk level ($0.1 > p > 0.05$)

barely failed to attain statistical significance (p=0.067)

barely fails to attain statistical significance at conventional levels ($p < 0.10$)

barely insignificant (p=0.075)

barely missed statistical significance (p=0.051)

barely missed the commonly acceptable significance level ($p < 0.053$)

barely outside the range of significance (p=0.06)

barely significant (p=0.07)

below (but verging on) the statistical significant level ($p > 0.05$)

Hypoteesin testaus (Neyman-Pearson)

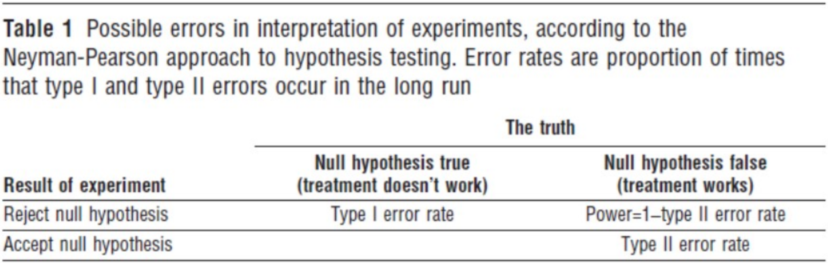


Table 1 Possible errors in interpretation of experiments, according to the Neyman-Pearson approach to hypothesis testing. Error rates are proportion of times that type I and type II errors occur in the long run

	The truth	
Result of experiment	**Null hypothesis true (treatment doesn't work)**	**Null hypothesis false (treatment works)**
Reject null hypothesis	Type I error rate	Power=1–type II error rate
Accept null hypothesis		Type II error rate

Sterne and Smith, BMJ 2001

Iso ongelma: Nollahypoteesin testauksen ja Neyman-Pearson -tyyppisen hypoteesin testauksen sekoittaminen

Merkitsevyystaso (significance level)

- P-arvo = havaittu merkitsevyystaso
  - "However, there are tradeoffs to consider. citrus performs feature selection but does not provide significance levels, such as p-values, for the strength of associations."
- Tyypin I virhe = ennalta valittu merkitsevyystaso
  - "Tyypin I -virheen todennäköisyyttä kutsutaan merkitsevyystasoksi"
- Kaksi eri merkitsevyystason käyttöä:
  - P-arvo on aineistosta laskettu ja aineistosta riippuva suure
  - α-taso on ennalta annettu, aineistosta riippumaton luku (esim. 0.05)
- Tilastollisen testisuureen ja 'efektin' sekoittaminen: "We have/have not found evidence of a statistically significant effect"

Kuinka usein hypoteesi on totta?

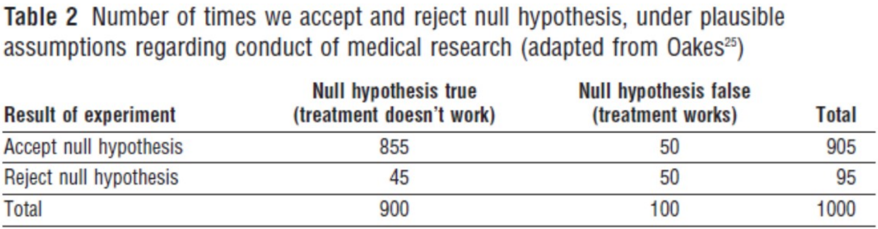


Table 2 Number of times we accept and reject null hypothesis, under plausible assumptions regarding conduct of medical research (adapted from Oakes[25])

Result of experiment	Null hypothesis true (treatment doesn't work)	Null hypothesis false (treatment works)	Total
Accept null hypothesis	855	50	905
Reject null hypothesis	45	50	95
Total	900	100	1000

- $45/95 = 47\%$ of statistically significant results (”reject the null hypothesis”) are such that the decision is a false alarm, I.e. is not correct

Sterne and Smith, BMJ 2001

Esimerkki jatkuu

- Prevalence of disease = proportion of false null hypotheses = 0.10
- Sensitivity of test = power = 0.50
- Specificity of test = $1 - \alpha = 0.95$

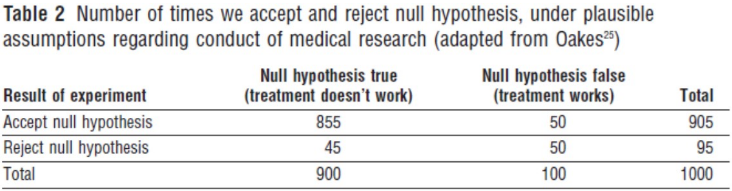


Table 2 Number of times we accept and reject null hypothesis, under plausible assumptions regarding conduct of medical research (adapted from Oakes[25])

Result of experiment	Null hypothesis true (treatment doesn't work)	Null hypothesis false (treatment works)	Total
Accept null hypothesis	855	50	905
Reject null hypothesis	45	50	95
Total	900	100	1000

Lehtien ohjeistusta

- Confidence intervals should be reported instead of P values for estimated parameters; P values should be reported only for relevant tests. Authors are encouraged to avoid the pitfalls associated with the misuse of P values as measures of significance (please refer to this statement from the American Statistical Association). AJE
- Significance Testing: For estimates of causal effects, we strongly discourage the use of categorized P-values and language referring to statistical significance (see discussion of this topic). We prefer instead interval estimation, which conveys the precision of the estimate with respect to sampling variability. We are more open to testing with respect to modeling decisions, such as for tests of interaction (see editorial) and for tests for trend, and with respect to studies using high-dimensional testing, such as genome-wide association or other genomic platforms. Epidemiology.

Ohjeistusta (2)

- Avoid solely reporting the results of statistical hypothesis testing, such as P values, which fail to convey important quantitative information. For most studies, P values should follow the reporting of comparisons of absolute numbers or rates and measures of uncertainty (eg, 0.8%, 95% CI 0.2% to 1.8%; P=.13). P values should never be presented alone without the data that are being compared. JAMA
- Statistical methods must be described and the program used for data analysis, and its source, should be stated. Summary statistics should define whether standard deviation (SD), variability of the sample, or standard error of the mean (SEM), uncertainty about the average, is being used. AM J Resp Crit Care Med

ASA statement on statistical significance and P-values (2016)

1 P-values can indicate how incompatible the data are with a specified statistical model.

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. Data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

Wasserstain ja Lazar, 2016

Moving to a world beyond "$p < 0.05$"

- Don't base your conclusions solely on whether an association or effect was found to be statistically significant (i.e., the p-value passed some arbitrary threshold such as p $<$ 0.05).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Wasserstain ym., 2019

Moving to a world beyond "$p < 0.05$"

- Don't base your conclusions solely on whether an association or effect was found to be statistically significant (i.e., the p-value passed some arbitrary threshold such as p < 0.05).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Don't say statistically significant

Ja vielä: siis miksi ei...?

- “Using bright-line rules for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.”
- ”No p-value can reveal the plausibility, presence, truth, or importance of an association or effect.

  Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important.”
- “For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.”
- ”Moving beyond ‘statistical significance opens researchers to the real significance of statistics, which is the science of learning from data, and of measuring, controlling, and communicating uncertainty.”

Mitä tilalle: ATOM

- Accept uncertainty. Be Thoughtful, Open, and Modest.

- Epävarmuuden hyväksyminen, arviointi ja esittäminen
- Välien estimointi
  - N.B. Ei saa enää tulkita tilastollista merkitsevyyttä sen mukaan, onko nollahypoteesin mukainen arvo välillä vaiko ei (!)
- Vaikutusten (*effects*) ja niiden seurausten todellinen merkittävyys, ei testisuureen tilastollinen merkitsevyys
- Havaittavissa olevien suureiden ennustaminen tilastollisten mallien avulla
- "... should no longer be subordinate to $p < 0.05$. These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important."

P-arvojen vaihtoehdot

- P-arvot pitäisi laskea sekä nollahypoteesille että pienimmän merkityksellisen eron mukaiselle hypoteesille; otoskoon huomioon ottaminen
- P-arvot pitää ilmoittaa absoluuttisina, ei esim. "< 0.01" tai "< 0.05"
- s-arvot: $s = -\log_2(p)$
- "second generation p-values"
- analysis of credibility
- p-values with false-positive risks
- p-values with Bayes-factor bounds
- tukeutuminen päätösteorian mukaisiin menettelyihin (koesuunnittelu)

Milloin dikotominen tulkinta on ok?

- Teollinen laaduntarkkailu
- Mallinnusta koskevissa automatisoiduissa valinnoissa
- Signaalin seulomisessa
- Päätöksenteko-ongelmissa, konfirmatoriset kokeet

Tutkimuksen teon ja julkaisupolitiikan muutos

- “significant sameness vs. significant differences”
- Tutkimuksen suunnittelun merkitys korostuu entuudestaan, samoin suunnittelu, analyysin ja raportoinnin avoimuus
- Ei raportoida valikoituja tuloksia vaan kaikki
- ‘Manuscripts should be assessed for suitability for publication based on the substantive importance of the research without regard to their reported results.”
- “Everything should be published in some form if whatever we measured made sense before we obtained the data because it was connected in a potentially useful way to some research question. (Amrheim ym.) regard to their reported results” (Locascio)

Lopuksi

Vapaa sana ja keskustelua

Jatkosuunnitelma?

Viitteitä

- Altman: Practical Statistics for Medical Research. Chapman and Hall 1991.
- Clayton ja Hills: Statistical Models in Epidemiology. Oxford University Press, 2013.
- Greenland ym.: Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016.
- Nummenmaa ym. Tilastollisten menetelmien perusteet.
- Rothman ja Greenland: Modern Epidemiology. Lippincott–Raven 1998.
- Royall: The effect of sample size on the meaning of significance tests. The American Statistician 1986.
- Sterne ja Smith: Sifting the evidencewhat's wrong with significance tests? BMJ 2001
- Uhari ja Nieminen: Epidmiologia ja biostatistiikka. Duodecim.
- Wasserstein ym. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician 2016.
- Wasserstein ym. Moving to a World Beyond $p < 0.05$. The American Statistician 2019.
- Lisäksiä lainauksia The American Statistician -artikkeleista ja niiden viitteistä