A p-érték védelmében (?)

Ferenci Tamás
tamas.ferenci@medstat.hu

2016. szeptember 30.

Tartalom

Tartalomjegyzék

1. Felvezető gondolatok **1**

2. Történeti megjegyzések **2**

3. Egy megoldási lehetőség és problémái: a Bayes-faktor **4**

1. Felvezető gondolatok

Amit biztosan nem kívánok megvédeni...

- p-érték $\neq$ hibavalószínűség (a felcserélt feltétel igazából ugyanez)
- Folyománya: a kísérlet nem önmagában ad ismeretet, hanem módosítja a meglevő ismereteinket
- Folyomány folyománya: bayesi keret
- Klinikai és matematikai szignifikancia fogalmának keveredése (a nagy minta valójában ennek a speciális esete! a „méretnélküliség" valójában ugyanez!)
- De: a szignifikanciavadászat nem a p-érték partikuláris problémája
- De: a „ragaszkodás az 5%-os küszöbhöz" nem a p-érték partikuláris problémája, hanem a fisheri és neyman-pearsoni tradíció keveredése (pontosabban ez utóbbi helytelen alkalmazása)

Bárdits A, Németh R, Terplán Gy. Egy régi probléma újra elötérben: a nullhipotézis szignifikanciateszt téves gyakorlata. Statisztikai Szemle. 2016;97(1):52-75.

Megoldási lehetőségek az ASA álláspont szerint

"In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct." Wasserstein RL, Lazar NA. ASA Statement on Statistical Significance and P-Values. The American Statistician. 2016;70(2):129-133. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337-50.

Mindezek fontossága, a felelősségünk

„Why Most Published Research Findings Are False?" (John Ioannidis, 2005, PLoS Medicine)

„The case against [medical] science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, [medical] science has taken a turn towards darkness." (Richard Horton, 2015, Lancet)

2. Történeti megjegyzések

Két iskola ütközése: R. A. Fisher

- p-érték: „measure of evidence"
- Mégpedig *egyetlen* hipotézisre vonatkozóan (tehát nincs ellenhipotézis)
- *Egyetlen* kísérletből akarunk dolgozni (nincs „képzeletbeli ismételt mintavétel")
- *Nem* akarunk külső információt felhasználni (nincs prior valószínűség)
- *Nem* hibavalószínűség
- Hogyan is lehetne valószínűség → ahhoz több kísérletet kellene, legalább képzeletben, tekintenünk (ahogy annak a kijelentésnek, hogy „ez az érme 40% valószínűséggel ad fejet" is csak úgy van értelme, ha többször feldobhatjuk – ez a valószínűség frekvencionista értelmezése)
- Inkább informális index, esetleg az a szabály, de ez sem mechanikusan, hogy adott p alatt elutasítunk: „significance testing" (*nem* „hypothesis testing"!)

Két iskola ütközése: J. Neyman és E. Pearson

- Döntés: „hibavalószínűség"
- *Két* hipotézis között döntünk
- Kísérletek sorát tételezzük fel...
- ...és a célunk, hogy *hosszú távon* biztosítsunk adott első- és másodfajú hibaarányt

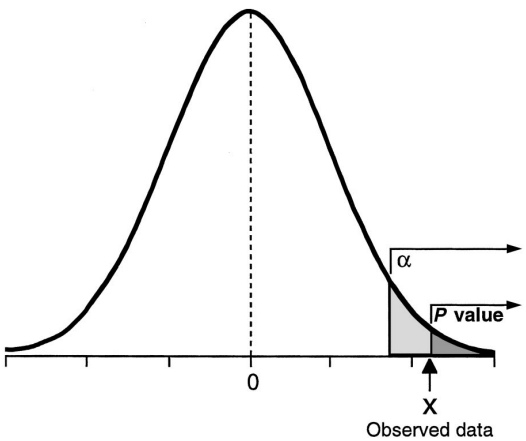
- Egy adott, konkrét hipotézisről úgysem dönthető el, hogy mennyire igaz, a kérdés csak hosszú távon értelmezhető
- Informális, szubjektív index helyett objektív, döntési szempontú („viselkedés, nem következtetés") megközelítés
- Elfogadási és elutasítási tartomány, *mindegy* hogy azon belül hová esik az empirikus tesztstatisztika (nincs bizonyíték-erősség)
- A határt eredetileg még úgy gondolták, hogy a kétféle hiba súlya alapján határozzák meg a kutatók
- „Hypothesis testing" (*nem* „significance testing")

A két iskola inkonzisztenciája

- Döntési határ van (és mindegy, hogy hová esik az empirikus tesztstatisztika) vagy nincs határ, és egy folytonos mértékünk van a bizonyíték erősségére?
- „Az érdekel minket, hogy az adott gyanúsított ártatlan-e, vagy az, hogy hosszú távon csak kevés ártatlant küldünk börtönbe?"
- A legfontosabb baj: nem lehet *egyszerre* beállítani a hosszútávú hibavalószínűséget *és* egyúttal az egyedi eredmény bizonyítőerejéről is nyilatkozni!
- A jelenlegi orvosi statisztikai gyakorlat azért keveri össze – teljesen inkonzisztensen – a kettőt, hogy megteremtse ennek az *illúzióját*

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol. 1993 Mar 1;137(5):485-96; discussion 497-501. Gigerenzer G. The superego, the ego, and the id in statistical reasoning. In: A handbook for data analysis in the behavioral sciences: Methodological issues. 1993:311-339. Lehmann EL. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?. In: Selected Works of EL Lehmann. 2012:201-208.

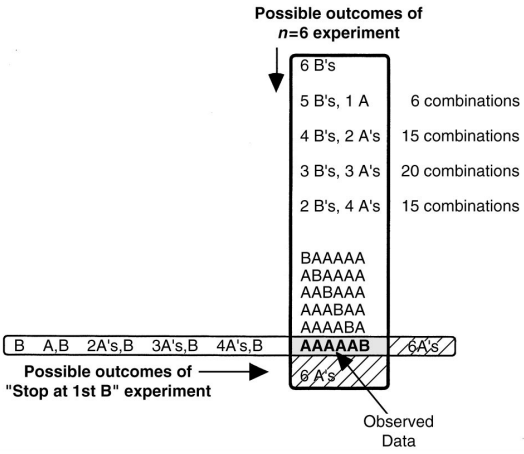
A kompromisszum elkerülésének illúziója



Úgy tűnhet, hogy a p-érték is egyfajta hibavalószínűség

„15. legjobb vagyok az évfolyamban" Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004.

A kompromisszum elkerülésének illúziója



Hosszútáv vs. rövidtáv Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004.

3. Egy megoldási lehetőség és problémái: a Bayes-faktor

A Bayes-faktor definíciója

A p-érték nem hibavalószínűség, az ugyanis valójában:

$$\mathbb{P}\left(H_0|\mathcal{S}\right) = \frac{\mathbb{P}\left(\mathcal{S}|H_0\right) \cdot \mathbb{P}\left(H_0\right)}{\mathbb{P}\left(\mathcal{S}\right)} \tag{1}$$

Elosztjuk egymással a null- és ellenhipotézisét, szerencsére $\mathbb{P}\left(\mathcal{S}\right)$ kiesik:

$$\frac{\mathbb{P}\left(H_0|\mathcal{S}\right)}{\mathbb{P}\left(H_1|\mathcal{S}\right)} = \frac{\mathbb{P}\left(\mathcal{S}|H_0\right)}{\mathbb{P}\left(\mathcal{S}|H_1\right)} \cdot \frac{\mathbb{P}\left(H_0\right)}{\mathbb{P}\left(H_1\right)} \tag{2}$$

Azaz:

$$\frac{\mathbb{P}\left(H_0|\mathcal{S}\right)}{1-\mathbb{P}\left(H_0|\mathcal{S}\right)} = \frac{\mathbb{P}\left(\mathcal{S}|H_0\right)}{\mathbb{P}\left(\mathcal{S}|H_1\right)} \cdot \frac{\mathbb{P}\left(H_0\right)}{1-\mathbb{P}\left(H_0\right)} \tag{3}$$

A Bayes-faktor definíciója

De hát ez nem más, mint az odds:

$$\text{odds}\left(H_0|\mathcal{S}\right) = \frac{\mathbb{P}\left(\mathcal{S}|H_0\right)}{\mathbb{P}\left(\mathcal{S}|H_1\right)} \cdot \text{odds}\left(H_0\right). \tag{4}$$

A jobb oldalon lévő tényező a Bayes-faktor (Bayes Factor, BF):

$$BF_{01} = \frac{\mathbb{P}\left(\mathcal{S}|H_0\right)}{\mathbb{P}\left(\mathcal{S}|H_1\right)}. \tag{5}$$

Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999 Jun 15;130(12):1005-13. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90(430):773-795. Good IJ. Weight of Evidence: A Brief Survey. In: Bayesian Statistics 2. 1985;249-270. Good IJ. Statistical Evidence. In: Encyclopedia of Statistical Sciences Vol 12. 2006.

A Bayes-faktor számítása

Szokásos θ-ra irányuló próba, $H_0 : \theta \in \theta_0$ és $H_1 : \theta \in \theta_1$ ($\theta_0 \cap \theta_1 = \emptyset$) mellett:

$$BF_{01} = \frac{\int_{\vartheta \in \theta_0} L\left(\mathcal{S}|\vartheta\right) \pi\left(\vartheta\right) \mathrm{d}\vartheta}{\int_{\vartheta \in \theta_1} L\left(\mathcal{S}|\vartheta\right) \pi\left(\vartheta\right) \mathrm{d}\vartheta}, \tag{6}$$

Mint az LR, csak szuprémum helyett végigintegrálunk

A Bayes-faktorok alkalmazása: általános megfontolások

- Esélyre hat, nem valószínűségre (gondoljunk a kettő között átszámolás jellegzetességeire)
- Prior kell
- Vagy: „minimum Bayes faktor”
- Csak módosítja a korábbi eredményeket (de ez igazából előny!)
- Mert: út a bayes-i logika felé
  - A legtöbb orvosi cikkben kevés szerepet kapnak a korábbi eredmények, biológiai megfontolások stb. (maximum kvalitatíve a diszkusszióban)
  - A filozófiai problémákból adódó illúzió: tudunk döntést hozni *csak* az adott mintára támaszkodva...?
  - Tipikus helyzet: a konklúziót már a diszkusszió legelső mondatában levonják...
  - ...és az igazából nem más, mint egy p-érték „nyelvi kifejtése”

Berger JO, Sellke T. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. Journal of the American Statistical Association. 1987;82(397):112-122. Goodman SN. Of P-values and Bayes: a modest proposal. Epidemiology. 2001 May;12(3):295-7.

A BF nagyságának jellemzése (Jeffreys szerint)

BF	dHart	bit	Interpretáció (Jeffreys)
$< 10^0$	< 0	< 0	Negatív
10^0-$10^{1/2}$	0-5	0-1,6	Említésre is alig méltó
$10^{1/2}$-10^1	5-10	1,6-3,3	Jelentős
10^1-$10^{3/2}$	10-15	3,3-5,0	Erős
$10^{3/2}$-10^2	15-20	5,0-6,6	Nagyon erős
$> 10^2$	>20	>6,6	Döntő

(A hartley – korábbi, de Bayes-faktorok kapcsán máig gyakran alkalmazott nevén ban – olyan mint a bit, csak nem kettes, hanem tízes alapú logaritmussal; a decihartley, ill. deciban a tizedrésze.) Jeffreys H. The theory of probability. Oxford University Press, 1961. Good IJ. Probability and the weighing of evidence. Charles Griffin & Company, 1950.

A Bayes-faktor története

- Harold Jeffreys már az – először 1935-ben megjelent – legendás művében, a Theory of Probability-ben bevezette
- De már előtte, a '10-es, '20-as években kidolgozta (Dorothy Wrinch-csel együtt)
- A logaritmikus skálát – és a deciban használatát – Alan Turing javasolta (ezzel ráadásul additívvá válik a bizonyíték mérőskálája!), és IJ Good proponálta erősen a XX. század második felében

Etz A, Wagenmakers EJ. J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test. Preprint, https://arxiv.org/abs/1511.08180. Good IJ. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II. Biometrika. 1979;66(2):393-396.

A Bayes-faktorok alkalmazása: a regresszió példája

Adott helyzetre konkretizálni kell, például regressziónál a szokásos

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \tag{7}$$

($\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ fae) mellett, ha vesszük a $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ kérdést, (6) formája:

$$BF_{01} = \frac{\prod_{i=1}^{n} \phi\left(\frac{y_i - \left(\alpha + \beta_1 x_{i,1} \ldots + \beta_{j-1} x_{j-1,i} + \beta_{j-1} x_{j+1,i} + \ldots + \beta_p x_{i,p}\right)}{\sigma}\right)}{\int_b \prod_{i=1}^{n} \phi\left(\frac{y_i - \left(\alpha + \beta_1 x_{i,1} \ldots + \beta_{j-1} x_{j-1,i} + b x_{j,i} + \beta_{j-1} x_{j+1,i} + \ldots + \beta_p x_{i,p}\right)}{\sigma}\right) \pi\left(b\right) \mathrm{d}b}, \tag{8}$$

ahol $\pi\left(b\right)$ a prior (a regressziós koefficiensre) Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g Priors for Bayesian Variable Selection. Journal of the American Statistical Association. 2008;103(481):410-423. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology. 2012;56(5):356-374. Rouder JN, Morey RD. Default Bayes Factors for Model Selection in Regression. Multivariate Behavioral Research. 2012;47(6):877-903.

A Bayes-faktorok alkalmazása: a regresszió példája

Leggyakoribb választás a Cauchy, ekvivalens a hierarchikus normális/inverz gamma modellel (de ez utóbbi direktebben átvihető többváltozós esetre), azaz:

$$\boldsymbol{\beta} | g \sim \mathcal{N}\left(\mathbf{0}, g\sigma^2 \left(\mathbf{X}^T \mathbf{X}/n\right)^{-1}\right) \tag{9}$$

$$g \sim \mathrm{InvGamma}\left(1/2, s^2/2\right), \tag{10}$$

ahol $\boldsymbol{\beta} = \left[\beta_i\right]_{i=1}^{p}$, $\mathbf{X} = \left[x_{i,j}\right]_{i=1,j=1}^{n,p}$ és s az új (hiper)paraméterek (JZS prior)

Szerencsére az ilyen – és ehhez hasonló – megoldások elég átfogóan implementálva vannak a `BayesFactor` R-csomagban

Egy gyakorlati vizsgálat

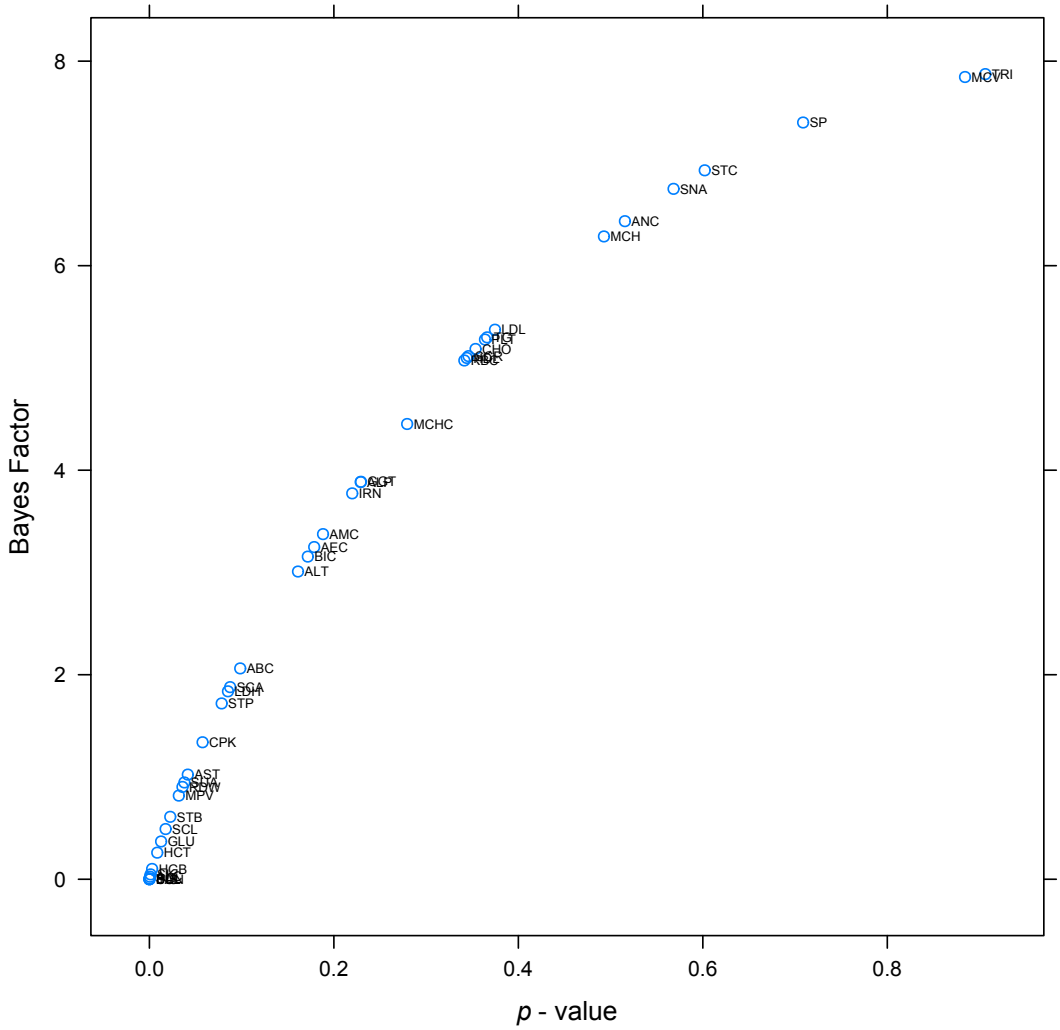
- A BF-eket igen kevéssé alkalmazzák (az is jórészt demonstratív)
- Próbáljuk ki a gyakorlatban
- Nem elméleti magyarázatra törekszünk

- Valós – nem szimulált – adatokon fogunk dolgozni
- 43 laborváltozó az NHANES legfrissebb (2013/14) ciklusából; a legkülönfélébb eloszlásokat követik, a legkülönfélébb korrelációs struktúrákkal
- A homogenitás kedvéért felnőtt férfiakra szűrve (a hiányzó adatúakat kihagytuk, egyszerűség kedvéért a súlyozással nem törődünk)
- $n = 1190$ (elég nagy ahhoz, hogy kisebb részhalmazok is értelmesen vizsgálhatóak legyenek)

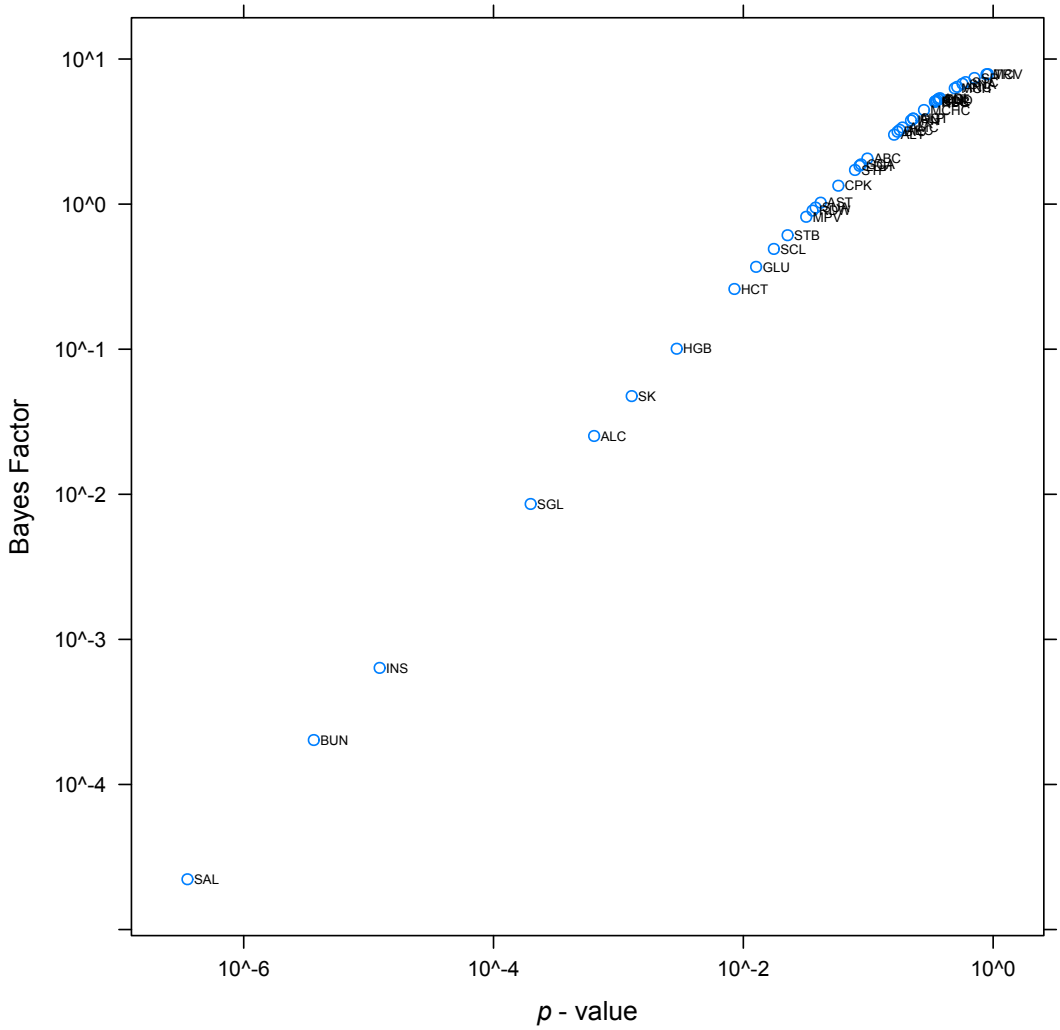
Egy gyakorlati vizsgálat

- Egy laborváltozót regresszálunk ki a többivel
- Kérdéseink:
  - Hogyan viszonyulnak a BF-ek a p-értékekhez?
  - Hogyan befolyásolják ezt az összefüggést a paraméterek, különösen az alkalmazott prior (s) és a mintanagyság?

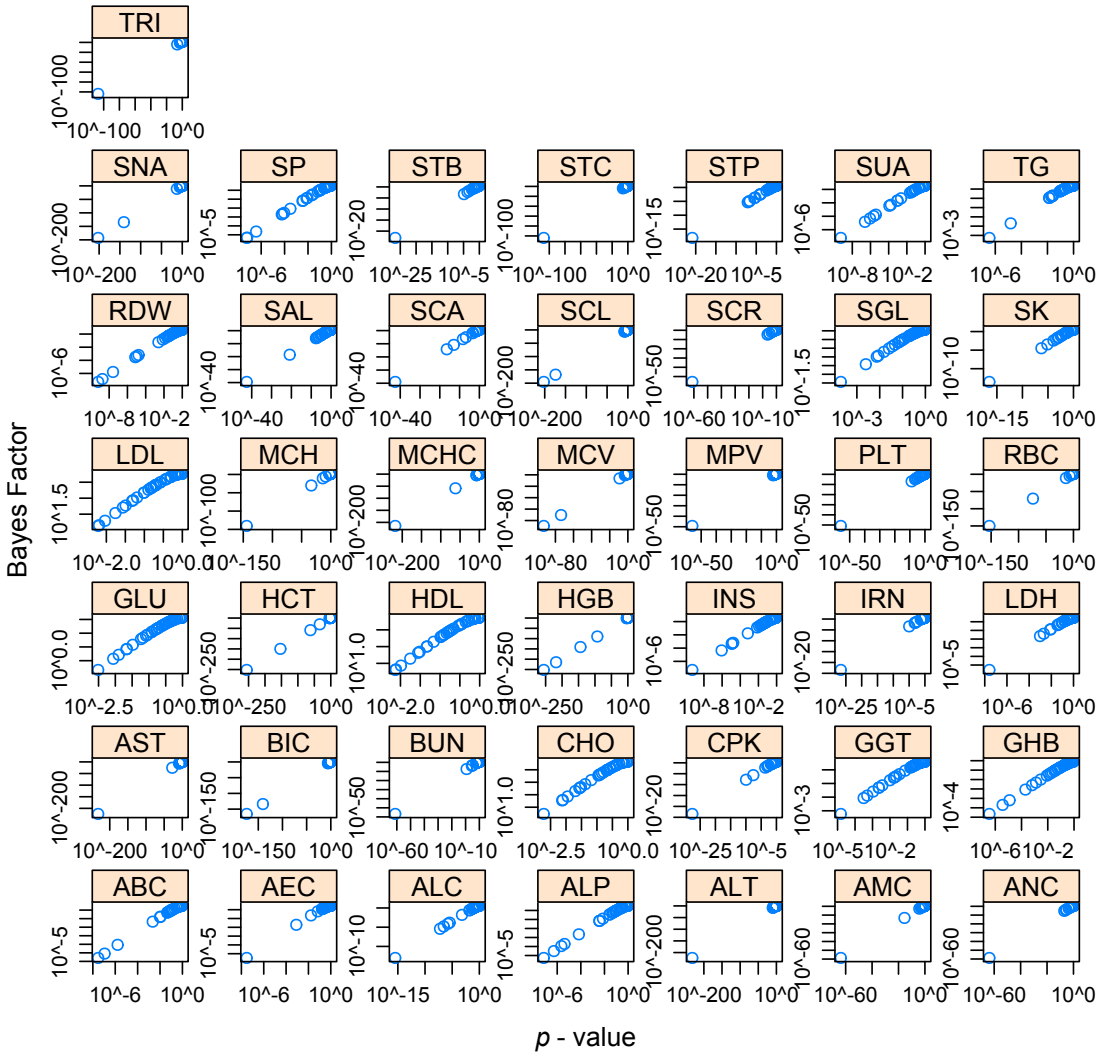
Összefüggés a glycohaemoglobin példáján



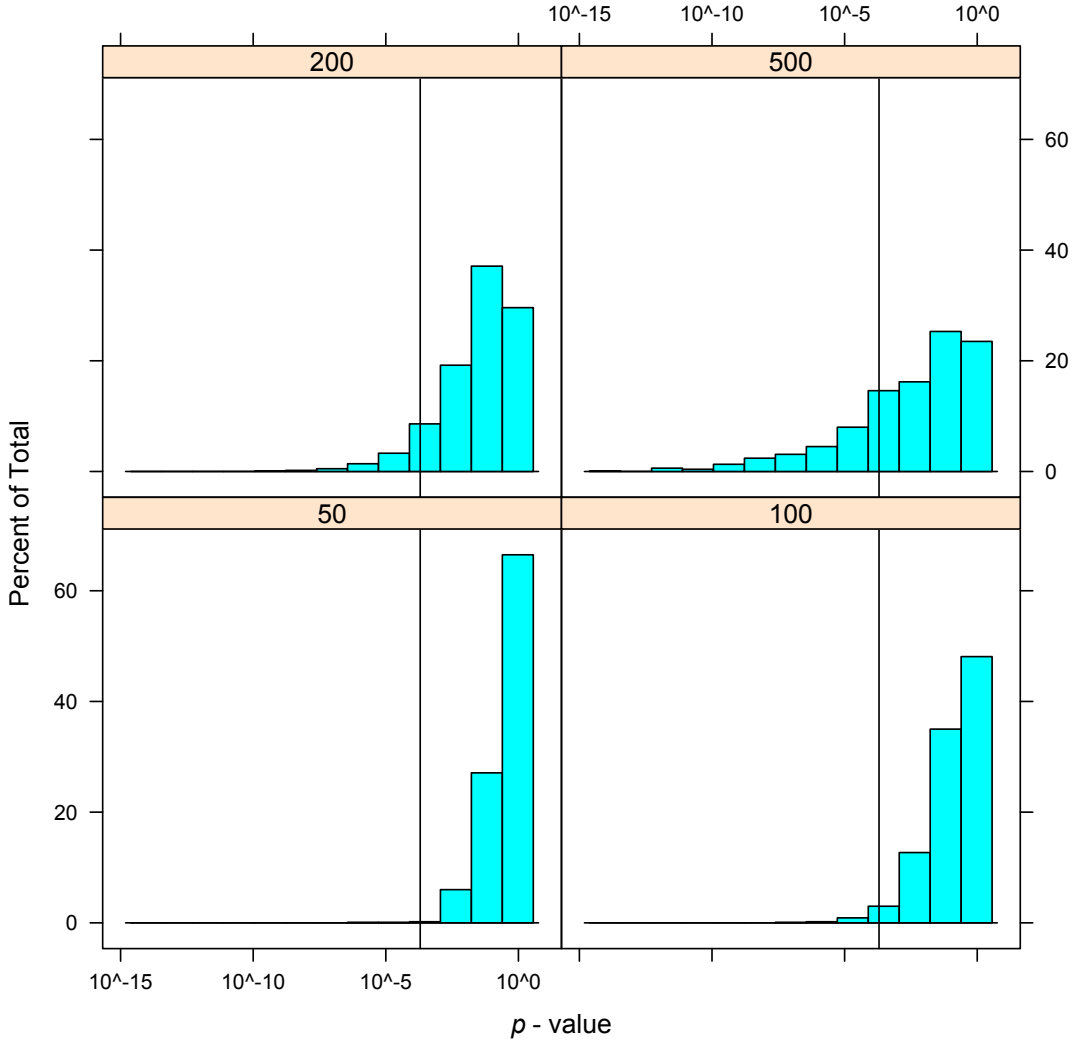
Összefüggés a glycohaemoglobin példáján



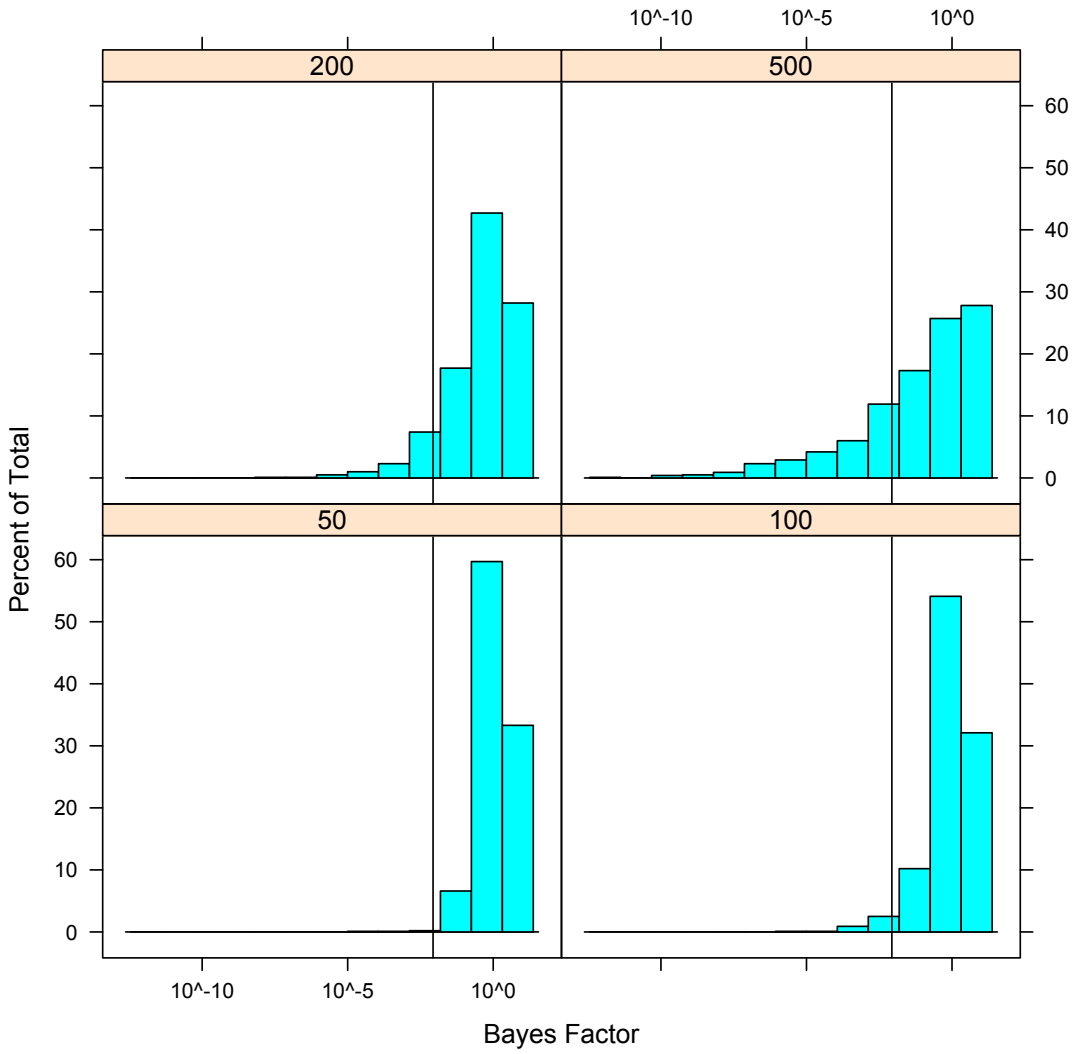
Az összes változót tekintve



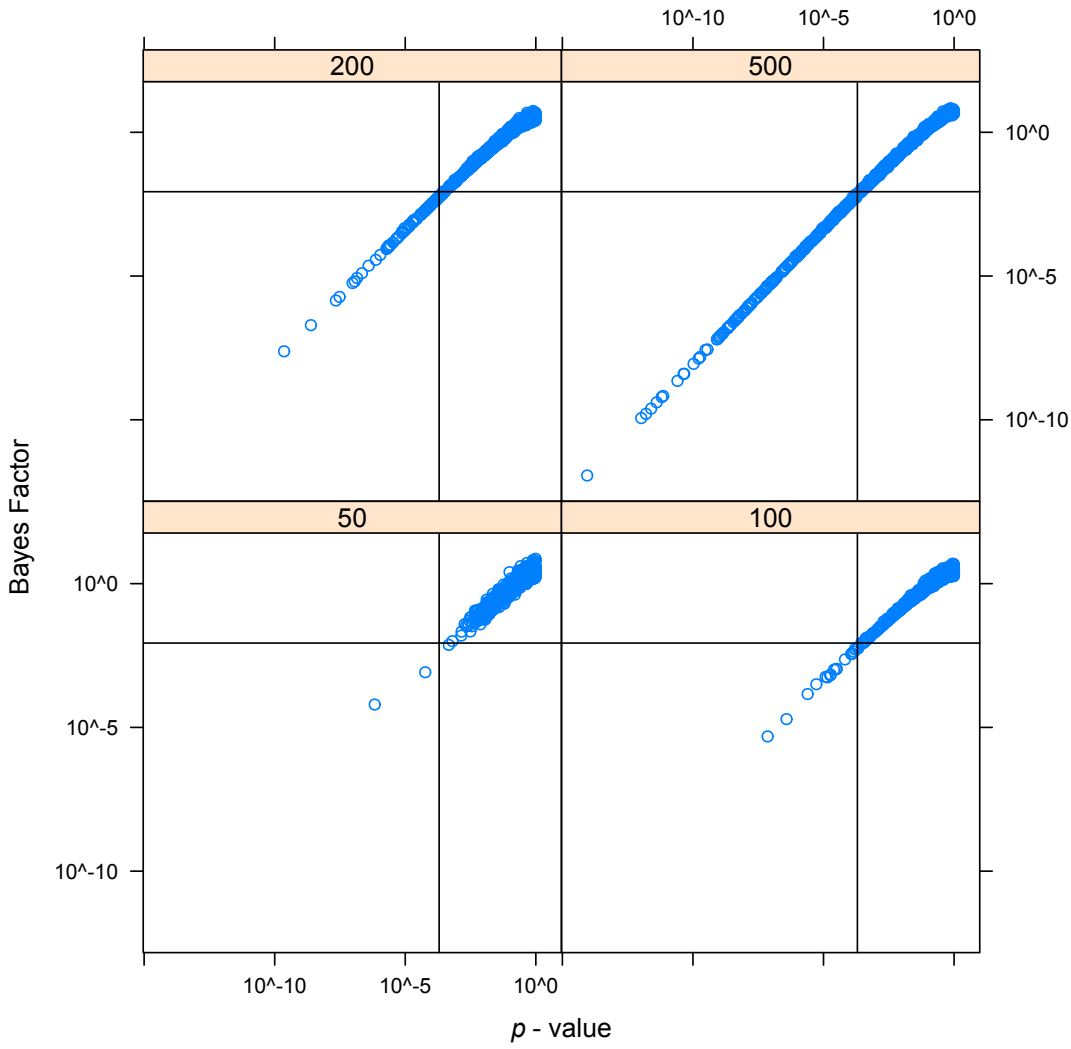
A mintanagyság (és mintavételi ingadozás) hatása: p-érték



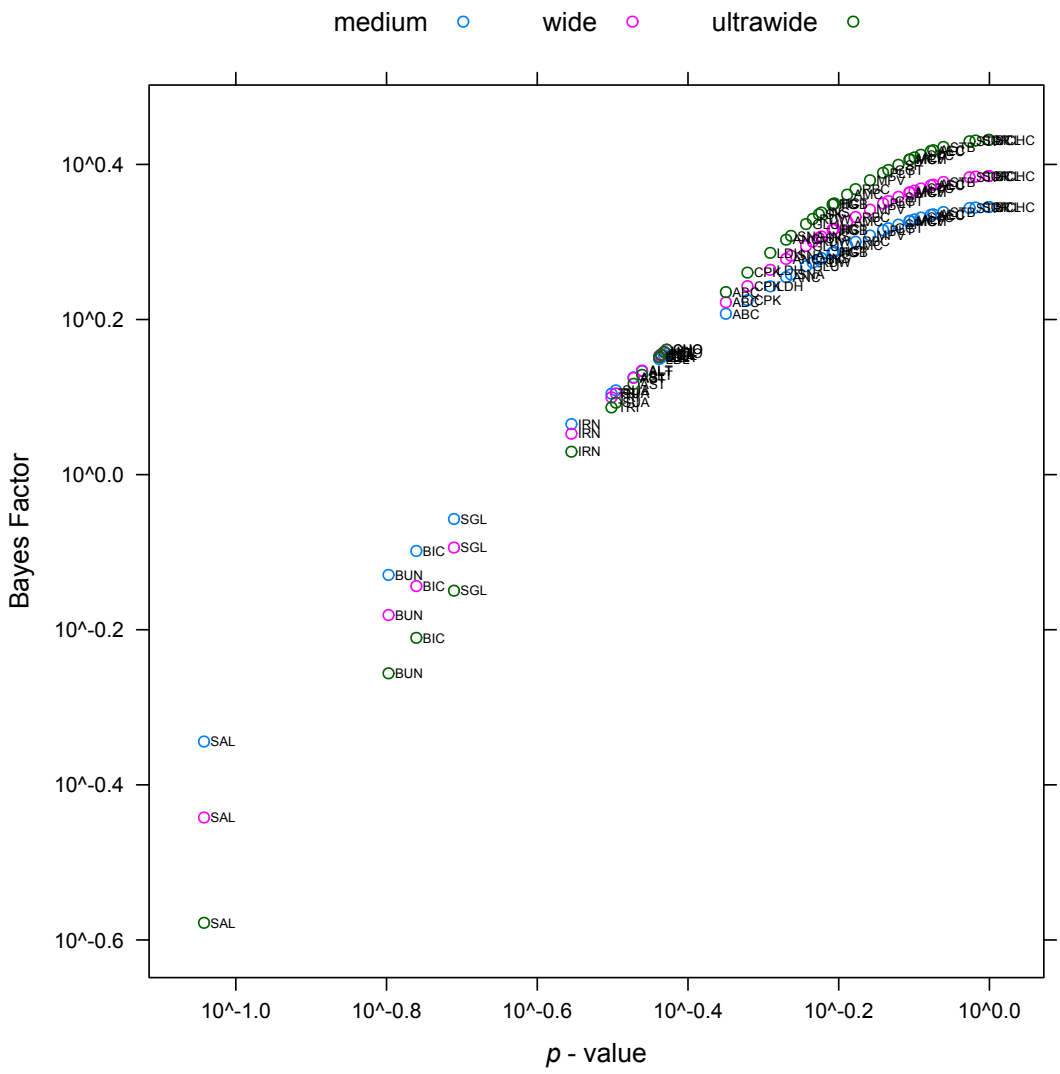
A mintanagyság (és mintavételi ingadozás) hatása: BF



A mintanagyság (és mintavételi ingadozás) hatása: együttes eloszlás



A prior hatása



Tanulságok

- A p-érték és a BF szoros kapcsolatban van (a kapcsolat ténye nem meglepő, a szorossága sokkal inkább)
- Minél nagyobb a mintaméret annál inkább
- Magyarázat: a mintanagyság növekedtével a BF tart az LR-hez (a likelihood függvény egyre pontszerűbb), ami viszont egy-egy értelműen megfeleltethető a p-értéknek
- A prior hatását illetően a sejtés: minél semmitmondóbb, annál hasonlóbbak lesznek a BF-ek (persze soha nem ugyanazok, hiszen az adatok is számítanak), ez annál kifejezettebb, minél kisebb a minta
- De ne feledjük, hogy az üzenete akkor is jobb a BF-nek!

Wetzels R, Grasman RPPP, Wagenmakers EJ. A Default Bayesian Hypothesis Test for ANOVA Designs. The American Statistician. 2012;66(2):104-111. Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Statistical Evidence in Experimental Psychology. Perspectives on Psychological Science: An empirical comparison using 855 t tests. 2011;6(3):291-298. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology. 2012;56(5):356-374. Marsman M, Wagenmakers EJ. Three Insights from a Bayesian Interpretation of the One-Sided P Value. Preprint. http://www.ejwagenmakers.com/inpress/MarsmanWagenmakersOneSidedPValue.pdf.

Köszönöm a figyelmet!

THINGS GOT REALLY INTERESTING WHEN THE STATISTICIAN STARTED DOING WARD ROUNDS.