日本社会心理学会 春の方法論セミナー
あなたの実験結果、再現できますか？
false-positive psychologyの最前線
2014/3/17

仮説検定における再現性の問題と新たな方法論

専修大学　岡田謙介

- 実験とは再現可能なものだ
  ―何度やっても同じように、失敗する

"Lab Rules"
http://www.cchem.berkeley.edu/cjrgrp/secret/secret.htm

2

- 再現性は科学の根幹
- 事前登録、追試、Mat&Meth,「研究者の自由度」、…
  - 重要なファクターは数多くある
    （cf. Simmons et al., 2011, *Psych Sci*）
- 今日は統計的な側面に絞ってお話しさせていただきます

3

復習: Neyman-Pearsonの帰無仮説検定

	真実	
判断	H_0（ない）	H_1（ある）
H_0（ない）	正しい判断	Type II Error false-negative 確率β
H_1（ある）	Type I Error false-positive 確率α	正しい判断

4

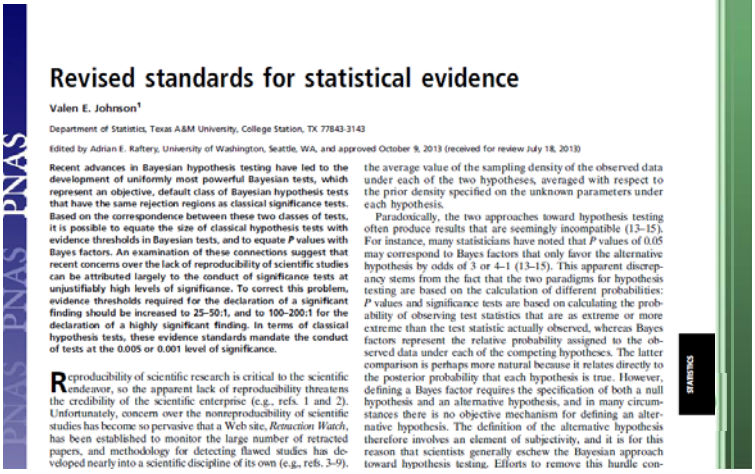
What if ....

$p < .05$ → $p < .005$ NEW

$(\alpha = .05)$　　$(\alpha = .005)$

5

これは最近のPNAS論文の主張



PNAS PNAS PNAS

Revised standards for statistical evidence

Valen E. Johnson[1]

Department of Statistics, Texas A&M University, College Station, TX 77843-3143

Edited by Adrian E. Raftery, University of Washington, Seattle, WA, and approved October 9, 2013 (received for review July 18, 2013)

Recent advances in Bayesian hypothesis testing have led to the development of uniformly most powerful Bayesian tests, which represent an objective, default class of Bayesian hypothesis tests that have the same rejection regions as classical significance tests. Based on the correspondence between these two classes of tests, it is possible to equate the size of classical hypothesis tests with evidence thresholds in Bayesian tests, and to equate *P* values with Bayes factors. An examination of these connections suggest that recent concerns over the lack of reproducibility of scientific studies can be attributed largely to the conduct of significance tests at unjustifiably high levels of significance. To correct this problem, evidence thresholds required for the declaration of a significant finding should be increased to 25–50:1, and to 100–200:1 for the declaration of a highly significant finding. In terms of classical hypothesis tests, these evidence standards mandate the conduct of tests at the 0.005 or 0.001 level of significance.

Reproducibility of scientific research is critical to the scientific endeavor, so the apparent lack of reproducibility threatens the credibility of the scientific enterprise (e.g., refs. 1 and 2). Unfortunately, concern over the nonreproducibility of scientific studies has become so pervasive that a Web site, *Retraction Watch*, has been established to monitor the large number of retracted papers, and methodology for detecting flawed studies has developed nearly into a scientific discipline of its own (e.g., refs. 3–9).

the average value of the sampling density of the observed data under each of the two hypotheses, averaged with respect to the prior density specified on the unknown parameters under each hypothesis.

Paradoxically, the two approaches toward hypothesis testing often produce results that are seemingly incompatible (13–15). For instance, many statisticians have noted that *P* values of 0.05 may correspond to Bayes factors that only favor the alternative hypothesis by odds of 3 or 4–1 (13–15). This apparent discrepancy stems from the fact that the two paradigms for hypothesis testing are based on the calculation of different probabilities: *P* values and significance tests are based on calculating the probability of observing test statistics that are as extreme or more extreme than the test statistic actually observed, whereas Bayes factors represent the relative probability assigned to the observed data under each of the competing hypotheses. The latter comparison is perhaps more natural because it relates directly to the posterior probability that each hypothesis is true. However, defining a Bayes factor requires the specification of both a null hypothesis and an alternative hypothesis, and in many circumstances there is no objective mechanism for defining an alternative hypothesis. The definition of the alternative hypothesis therefore involves an element of subjectivity, and it is for this reason that scientists generally eschew the Bayesian approach toward hypothesis testing. Efforts to remove this hurdle con-

STATISTICS

ベイズファクター（Bayes Factor, BF）

- 2つの仮説（モデル）の、事後オッズと事前オッズの比

$$BF_{10} = \frac{p(H_1|X) \Big/ p(H_0|X)}{p(H_1) \Big/ p(H_0)}$$

- データによって与えられた、仮説H_0に比して仮説H_1を支持する程度（オッズ）の変化を表す

（Bernardo & Smith, 1994; Lavine & Schervish, 1999, *JASA*）

7

ベイズファクターのrules of thumb

Jeffreys (1961)

BF_{10}	解釈
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

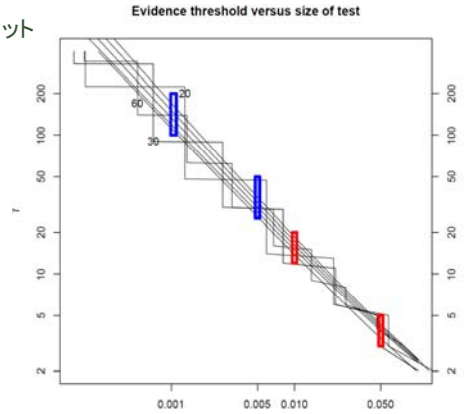
Kass & Raftery (1995, *JASA*)

BF_{10}	解釈
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
>150	Very strong

※BFはJohnson (2013 *Annals Stat*)の方法）

p値とBFの対応：理論 (Johnson, 2013, *PNAS* Fig 1)

BFのカットオフ値

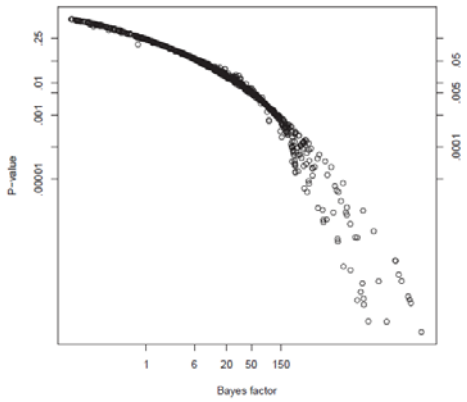


- p値のカットオフ値 ＝ 有意水準α

9

p値とBFの対応：実データ(Johnson, 2013, *PNAS* Fig 2)

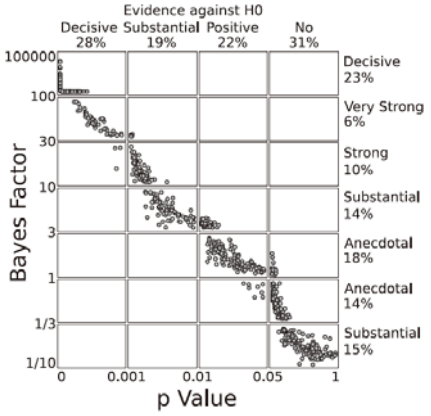
- Wetzels et al. (2011, *Persp Psych Sci*) の収集した855のt検定



10

p値とBFの対応：実データ(Wetzels et al. 2011, Fig 3)

- 同じデータ、Rouder et al. (2009, *Psych Bull Rev*)のBF



11

Johnson (2013, *PNAS*)

- Johnson (2013, *Ann Stat*)の「一様最強力ベイズ検定」を介して、p値とベイズファクター(BF)のカットオフ値を対応づける
- すると、$p = .05$は$BF = 3 \sim 5$に対応する. これは、BFの標準的な解釈としては強い証拠とは言えない.
- BFの標準的な解釈で強い証拠とされる$BF = 20 \sim 50$に対応するのは、$p = .005$である
- したがって、

NEW

$p < .05$ ⇒ $p < .005$

- 「高すぎる有意水準が、再現性の問題の原因」

12

「p<.05」は甘すぎる基準か？

- そうかもしれない
- Bemの「超能力」結果もBFで見ると効果は小さい
  (Rouder & Morey, 2011, *Psychon Bull Rev*)
- 同種の議論は昔からある(e.g., Berger & Selke, 1987, *JASA*)

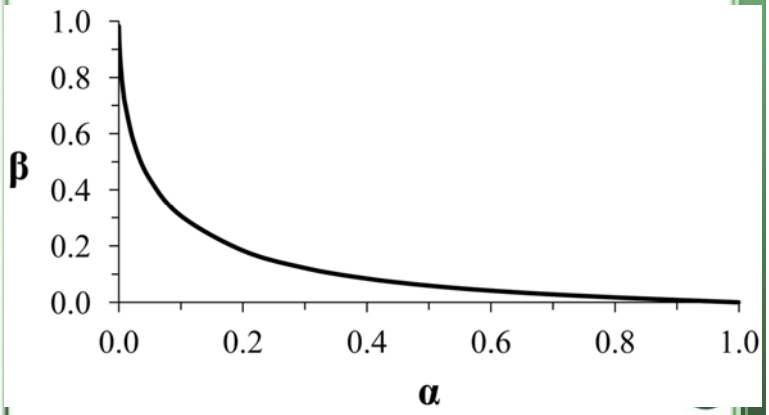
Table 7. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_{NOR})$ When $\pi_0 = \frac{1}{2}$

P Value (p)	t	$\underline{Pr}(H_0 \mid x, G_{NOR})$	$\underline{Pr}(H_0 \mid x, G_{NOR})/(pt^2)$
.10	1.645	.412	1.52
.05	1.960	.321	1.67
.01	2.576	.133	2.01
.001	3.291	.0235	2.18

- α=.05の根拠はそもそも大きくない
- ただし、αを下げることは、βを(ときに激しく)上げることでもある

13

αとβの関係



Mudge et al. (2012, *Plos One*) 独立な2群のt検定, $n_1 = n_2 = 10, \delta = 1.0$

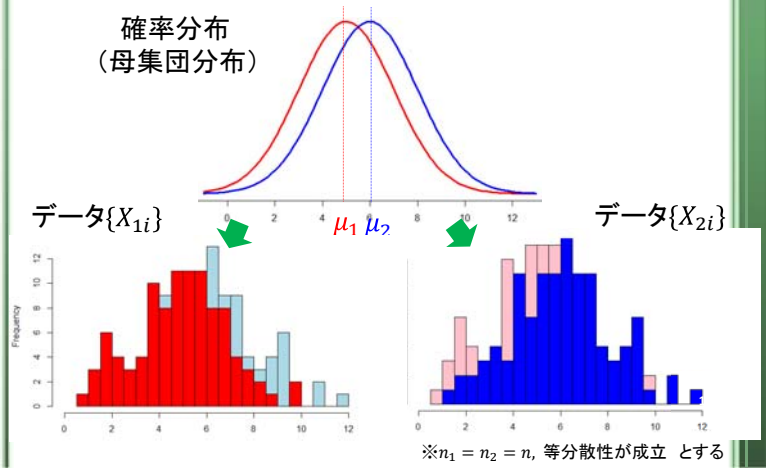
翻って、p値とは何か

- p値はprobabilityのpだときいたし、何かの確率だろう。えっと…

- 「帰無仮説が正しい確率」
- 「研究者の仮説が間違っている確率」

15

仮説検定のロジック(例:t検定)



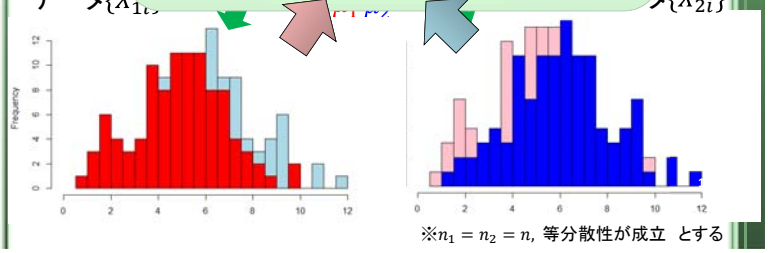
※$n_1 = n_2 = n$, 等分散性が成立 とする

仮説検定のロジック(例:t検定)

検定統計量

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

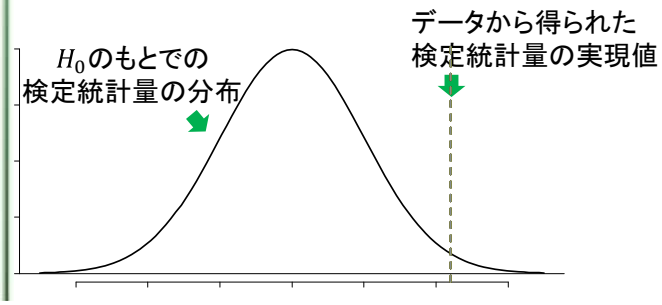
ここで $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$



※$n_1 = n_2 = n$, 等分散性が成立 とする

p値

- $H_0: \mu_1 = \mu_2$が真のときの検定統計量 t の分布は既知

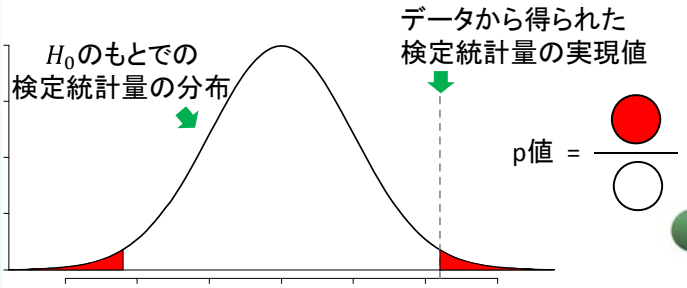


18

p値

$p = P(|t_{rep}| > |t| \,|\, H_0)$

- $H_0: \mu_1 = \mu_2$が真のときの検定統計量 t の分布は既知
- H_0が真で、今回と同じ標本サイズのデータを取得することを繰り返したとき、今回得られたよりも極端な検定統計量の値が得られる確率がp値

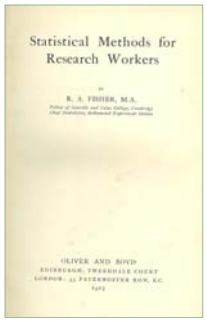


19

検定の生まれた時代：

http://psychclassics.yorku.ca/

- R. A. Fisherの世界的ベストセラー



14版まで

『研究者**のための**統計的方法』(1925)

9版まで

『実験計画法』(1935)

20

検定の生まれた時代：1920-30s

- 実験データを評価する「科学的な」方法を多くの研究者が求めていた
- 農事試験での実用性が示された
  - 試験の解釈をめぐる、専門家と非専門家とのコミュニケーション規則としての役割も(柴村, 2004)
- 計算機はなく、柔軟に「統計モデルをデータに当てはめる」ことはほぼ不可能だった
- 必要な検定統計量(*t*, *F*, …)の分布表が提供された
  - Fisherの「計算機」calculators

21

e.g. Lee & Pearson (1925) *Biometrika*

Table of the First Twenty Tetrachoric Functions to Seven Decimal Places

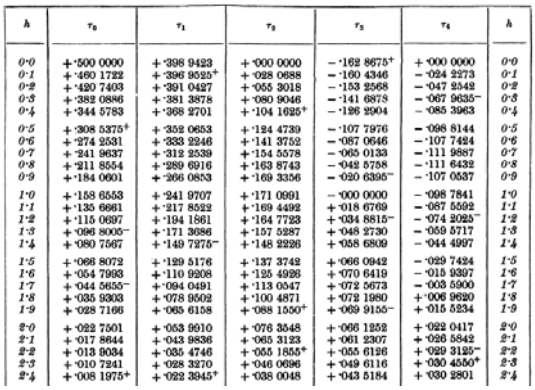


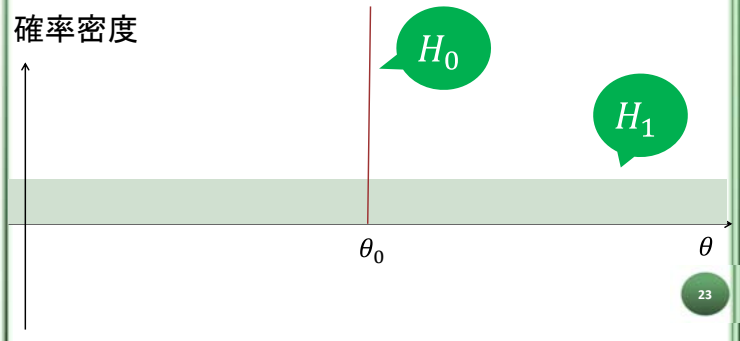
Table of the First Twenty Tetrachoric Functions.

h	r_0	r_1	r_2	r_3	r_4	h
0·0	+·500 0000	+·398 9423	+·000 0000	−·162 8675⁺	+·000 0000	0·0
0·1	+·460 1722	+·396 9525⁺	+·028 0688	−·160 4346	−·024 2273	0·1
0·2	+·420 7403	+·391 0427	+·055 3018	−·153 2568	−·047 2342	0·2
0·3	+·382 0886	+·381 3878	+·080 9046	−·141 6878	−·067 9635⁻	0·3
0·4	+·344 5783	+·368 2701	+·104 1625⁺	−·126 2904	−·085 3963	0·4
0·5	+·308 5375⁺	+·352 0653	+·124 4739	−·107 7976	−·098 8144	0·5
0·6	+·274 2531	+·333 2246	+·141 3752	−·087 0646	−·107 7424	0·6
0·7	+·241 9637	+·312 2539	+·154 5578	−·065 0133	−·111 9887	0·7
0·8	+·211 8554	+·289 6916	+·163 8743	−·042 5758	−·111 6432	0·8
0·9	+·184 0601	+·266 0853	+·169 3356	−·020 6395⁻	−·107 0537	0·9
1·0	+·158 6553	+·241 9707	+·171 0991	−·000 0000	−·098 7841	1·0
1·1	+·135 6861	+·217 8522	+·169 4492	+·018 6769	−·087 5592	1·1
1·2	+·115 0697	+·194 1861	+·164 7723	+·034 8815⁻	−·074 2025⁻	1·2
1·3	+·096 8005⁻	+·171 3686	+·157 5287	+·048 2730	−·059 5717	1·3
1·4	+·080 7567	+·149 7275⁻	+·148 2226	+·058 6809	−·044 4997	1·4
1·5	+·066 8072	+·129 5176	+·137 3742	+·066 0942	−·029 7424	1·5
1·6	+·054 7993	+·110 9208	+·125 4926	+·070 6419	−·015 9397	1·6
1·7	+·044 5655⁻	+·094 0491	+·113 0347	+·072 5673	−·003 5900	1·7
1·8	+·035 9303	+·078 9502	+·100 4871	+·072 1980	+·006 9620	1·8
1·9	+·028 7166	+·065 6158	+·088 1550⁺	+·069 9155⁻	+·015 5234	1·9
2·0	+·022 7501	+·053 9910	+·076 3548	+·066 1252	+·022 0417	2·0
2·1	+·017 8644	+·043 9836	+·065 3123	+·061 2307	+·026 5842	2·1
2·2	+·013 9034	+·035 4746	+·055 1855⁺	+·055 6126	+·029 3125⁻	2·2
2·3	+·010 7241	+·028 3270	+·046 0696	+·049 6116	+·030 4550⁺	2·3
2·4	+·008 1975⁺	+·022 3945⁺	+·038 0048	+·043 5184	+·030 2801	2·4

22

検定のそもそもの問題点

- 点仮説のH_0は、1点をのぞいて確率ゼロである



23

仮説検定の枠組みの問題点

$H_0: \mu_1 = \mu_2$

- 帰無仮説H_0は**常に間違っている**

(Loftus, 1996, *Curr Dir Psych Sci*)

$H_1: \mu_1 \neq \mu_2$

- 対立仮説H_1は**なにも主張していない**

- 仮説検定とp値に依存するのは危険

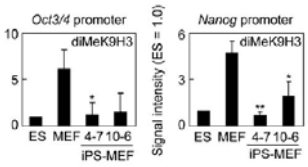
24

False-positiveについて

- ないものをあると言ってしまうこと
- 差や影響がない、0であるという前提が「常に間違っている」のならば、false-positiveの議論はそもそもおかしな感じ

- 「ない」帰無仮説H_0 の棄却によって言いたいことを主張する、という枠組みから離れてみては？

それから100年近く…



s, MEFs, and iPS cells (MEF4-7 and MEF10-6). Data were quantified by real-tir
values compared to ES cells (n = 3). *p < 0.05; **p < 0.01 compared to MEFs.

(Takahashi & Yamanaka, 2006, *Cell*)

The Standard Model Higgs boson is excluded at 95% CL in the mass range 111–559 GeV, except for the narrow region 122–131 GeV. In this region, an excess of events with significance 5.9σ, corresponding to $p_0 = 1.7 \times 10^{-9}$, is observed. The excess is driven by the two channels with the highest mass resolution, $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$, and the equally sensitive but low-resolution $H \to WW^{(*)} \to \ell\nu\ell\nu$ channel. Taking into account the entire mass range of the search, 110–600 GeV, the global significance of the excess is 5.1σ, which corresponds to $p_0 = 1.7 \times 10^{-7}$.

(ATLAS Collaboration, 2012, *Phys Lett B*)

26

心理学における統計改革（statistical reform）

2009 APA Manual第6版
具体的な指示・記載へ

1994 Cohen
『地球は丸い(p<.05)』

Finch et al. (2001)など
実効力のある改革へ

1996 APA 推測統計に
関する専門委員会設置

Kline (2004)
『有意性検定を超えて』APA
多くの論文

Wilkinson & APA Task Force (1999)
『心理学の論文誌における統計的方法』

2001 APA Manual第5版
効果量をより推奨

(Fidler, 2010, ICOTS8)

27

既存の「統計改革」の推奨

- 効果量　…単純
- 信頼区間　…仮説検定と裏表の関係
- 検定力分析　…仮説検定の枠組み内

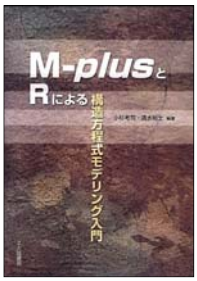
- もちろんどれも大事ですが、

➡ もう一歩進みたい

28

現代 with PC

- 複雑な統計モデルでも、汎用ソフトウェアで柔軟に構築・推定できる
  - 検定の作られた時代とは決定的に違う



Mplus
(Muthen)

BUGS
(Spiegelhalter)

Stan
(Gelman)

29

統計学からの提言

型にはまった
検定
と付随する枠組み

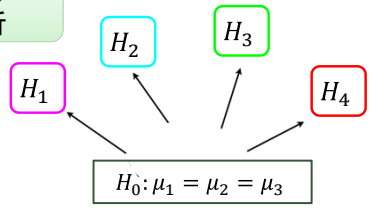
➡

オーダーメイドの
仮説・モデルの積極的利用

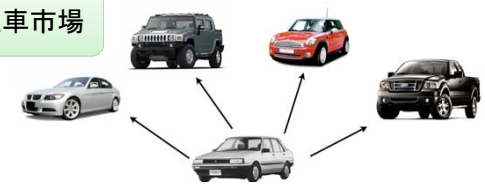
30

画一的分析から、現象のモデル構築・評価へ

成熟した統計分析



成熟した車市場



31

統計学からの提言（イイカエ）

(cf. Gelman, 2000, *Comp Stat*)

Sign(符号)

Type I error
Type II error
を気にするよりも

Type S error
Type M error
を気にしよう！

Magnitude (大きさ)

32

- …と盛り上げておいてなんですが
- 閑話休題
- p値とsampling intention、停止規則

33

頻度論とベイズの違い

- ベイズ統計学は、母数を確率変数と考える統計学

	頻度論	ベイズ
母数 θ	定数	確率変数
データ x	確率変数	定数

34

p値と停止規則のもう1つの関係

- p値は、サンプリングの停止規則に依存する
- 例：コインを12回投げて3枚表が出た。このコインはフェアなコインか？

  $H_0: \pi = \frac{1}{2}$のもとでのp値を求めるとき
- 【状況1】「12回投げる」ことが事前に決まっていたとき、二項検定. $p = 0.073$
- 【状況2】「3枚表が出るまで投げる」ことが事前に決まっていたとき、負の二項検定. $p = 0.033$
- 同じデータでもp < .05か否かが変わる

(e.g., Little, 2005, *Am Stat*; ここでは対立仮説を$H_1: \pi < \frac{1}{2}$としているが、両側検定でも同様)

35

もちろん、t検定でも

(cf. Kruschke, 2013, *JEP: General*)

- [状況1] $N = 8$のデータを収集することを計画した。実際に$N = 8$を得た。
- [状況2] Nは決めずに4時間データを収集することを決めていた。集まったデータは$N = 8$だった。
- [状況3] $N = 4$ のデータを収集することを計画した。集めて分析したところ有意でなかったので、さらにNを足して$N = 8$を得た。
  （もし$N = 4$で有意になったら止めていた）

36

[状況1]

- 将来の繰り返しでも、$N = 8$の収集が繰り返される
  したがってH_0のもとでのtの分布は$t(7)$

[状況2]

- 将来の繰り返しでのNは、確率的に変動する。$N = 6,7,8,9,10$である確率がそれぞれ20%ずつとすると、H_0のもとでの検定統計量tの分布は

$$f(t) = 0.2t(5) + 0.2t(6) + 0.2t(7) + 0.2t(8) + 0.2t(9)$$

p値はサンプリングの停止規則に依存して変わる
既存のp-hacking研究では考慮されていない（と思う）

統計学からの提言

型にはまった
検定
と付随する枠組み

→ オーダーメイドの
仮説・モデルの積極的利用

38

実験と調査・観察

- 違いは条件への**ランダム割り当て**の有無
- 実験では、関心のある要因の各水準（条件）へ個体をランダムに割り当てることにより、それ以外の従属変数に影響を与えうる要因の影響を平均的に除くことができる
  - 説明変数が少なくて済む
- 調査・観察では、関心のある要因以外にも、従属変数に影響を与えうる要因が（多く）ある
  - 説明変数の候補、および従属変数への影響の与え方が複雑になる➡適切なモデリングが必要

39

提案

- **情報仮説の評価** ← Type S Error

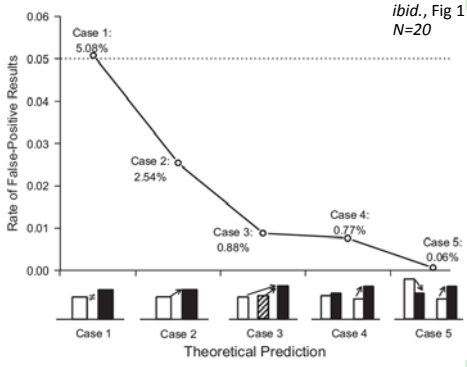
「十分に複雑」な

- **統計モデルの構築・評価**
  - **感度分析** ← Type M Error
  - **事後予測チェック**

40

Murayama et al. (in press)の提案(1)

- 事前に情報仮説を持っておく
  → 検定でのfalse-positiveを減らせる
- ただし検定では情報仮説のよさを直接評価できない



例：Fonken et al. (2012, *PNAS*).

Light at night increases body mass by shifting the time of food intake

Laura K. Fonken[a,1], Joanna L. Workman[b], James C. Walton[a], Zachary M. Weil[a], John S. Morris[b], Abraham Haim[c], and Randy J. Nelson[a,b]

Departments of [a]Neuroscience and [b]Psychology, Ohio State University, Columbus, OH 43210; and [c]Israeli Center for Interdisciplinary Research in Chronobiology, University of Haifa, Haifa 31905, Israel

Edited* by David L. Denlinger, Ohio State University, Columbus, OH, and approved September 3, 2010 (received for review June 24, 2010)

The global increase in the prevalence of obesity and metabolic disorders coincides with the increase of exposure to light at night (LAN) and shift work. Circadian regulation of energy homeostasis is controlled by an endogenous biological clock that is synchronized by light information. To promote optimal adaptive functioning, the circadian clock prepares individuals for predictable events such as food availability and sleep, and disruption of clock function causes circadian and metabolic disturbances. To determine whether a causal relationship exists between nighttime light exposure and obesity, we examined the effects of LAN on body mass in male mice. Mice housed in either bright (LL) or dim (DM) LAN have significantly increased body mass and reduced glucose tolerance compared with mice in a standard (LD) light/dark cycle, despite equivalent levels of caloric intake and total daily activity output. Furthermore, the timing of food consumption by DM and LL mice differs from that in LD mice. Nocturnal rodents typically eat substantially more food at night; however, DM mice consume 55.5% of their food during the light phase, as compared with 36.5% in LD mice. Restricting food consumption to the active phase in DM mice prevents body mass gain. These results suggest that low levels of light at night disrupt the timing of food intake and other metabolic signals, leading to excess weight gain. These data are relevant to the coincidence between increasing use of light at night and obesity in humans.

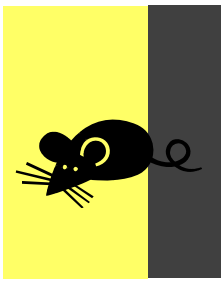
Multiple studies suggest a link between the molecular circadian clock and metabolism (reviewed in ref. 9). Mice harboring a mutation in their clock genes are susceptible to obesity and metabolic syndrome (10). *Clock* mutants show profound changes in circadian rhythmicity as well as disrupted diurnal food intake and increased body mass. Serum leptin, glucose, cholesterol, and triglyceride levels also are increased in *Clock* mutants compared with wild-type mice. Mice lacking the vasoactive intestinal peptide receptor 2 pathway, which plays an important role in SCN communication (11), have metabolic abnormalities similar to those in *Clock*-mutant mice (12). Furthermore, consumption of a high-fat diet alters circadian rhythmicity and the cycling of circadian clock genes in mice (13).

Because multiple studies have linked disruption of the mo-

in constant lighting have no temporal cue to distinguish time of

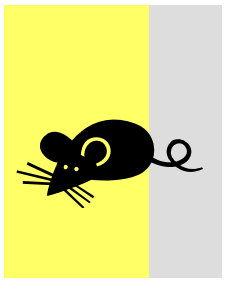
'Lock5Data' package
in R/CRAN
(Lock et al., 2012, Wiley)

Light/Dark (LD)群



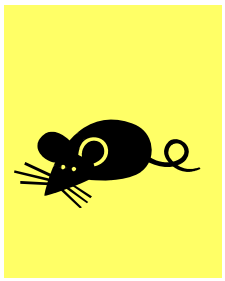
$\sim N(\mu_{LD}, \sigma^2)$

Light/Dim Light (DM)群



$\sim N(\mu_{DM}, \sigma^2)$

Continuous Light (LL)群



$\sim N(\mu_{LL}, \sigma^2)$

- 従属変数は体重増分[g]

43

研究仮説

- $H_1: \mu_{LD} < \mu_{DM} < \mu_{LL}$
  - 夜が明るいほど体重は増加する
- $H_2: \mu_{LD} < \{\mu_{DM}, \mu_{LL}\}$
  - 夜が暗くないと、体重は増加する

→ 情報仮説（informative hypothesis）

- $H_a: \mu_{LD}, \mu_{DM}, \mu_{LL}$
  - とくに一貫した関係はない

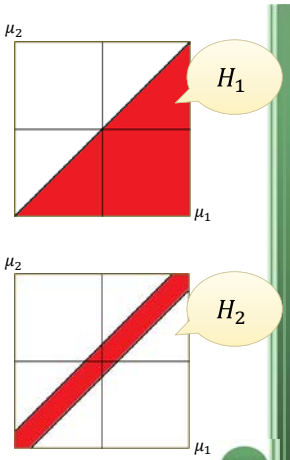
→ 無制約仮説

統計的データ解析における仮説とは、
パラメータθ に関する仮説$H(\theta)$である。

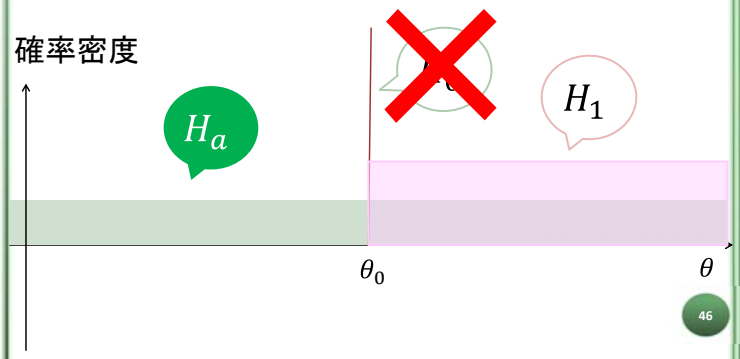
44

2群の平均値の比較

- 考えられる仮説
  - $H_0: \mu_1 = \mu_2$
  - $H_1: \mu_1 > \mu_2$
  - $H_2: |\mu_1 - \mu_2| < 1$
  - $H_a: \mu_1, \mu_2$
- H_1やH_2のような、研究者の仮説を反映して、パラメータに不等式制約を入れた仮説を情報仮説（informative hypothesis）という

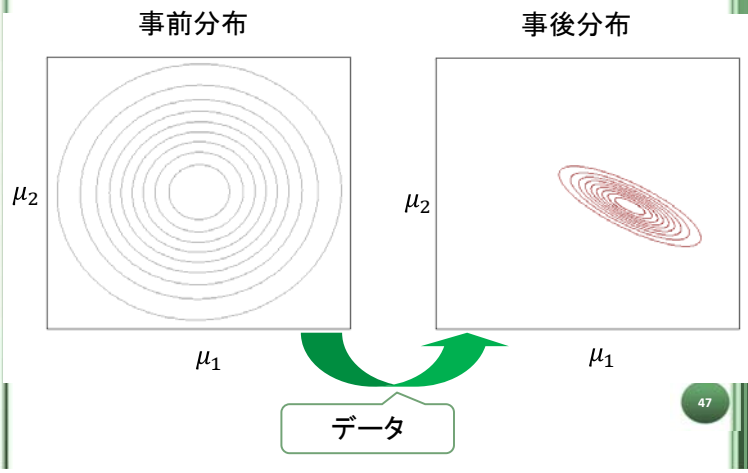


45



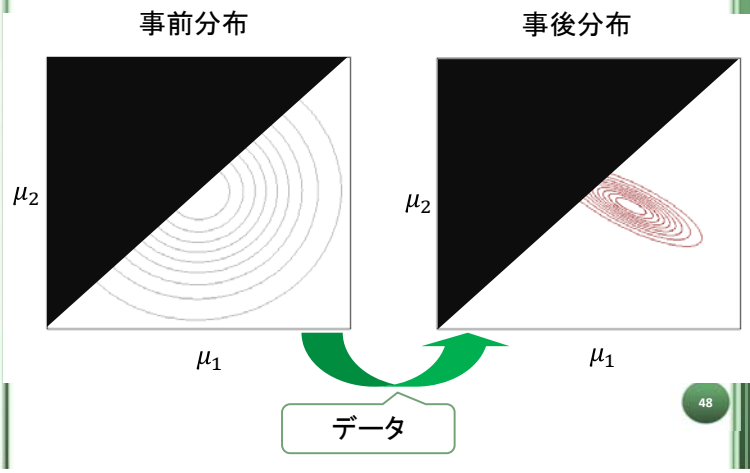
46

「$H_a: \mu_1, \mu_2$」の下での事前分布と事後分布



47

「$H_1: \mu_1 > \mu_2$」の下での事前分布と事後分布



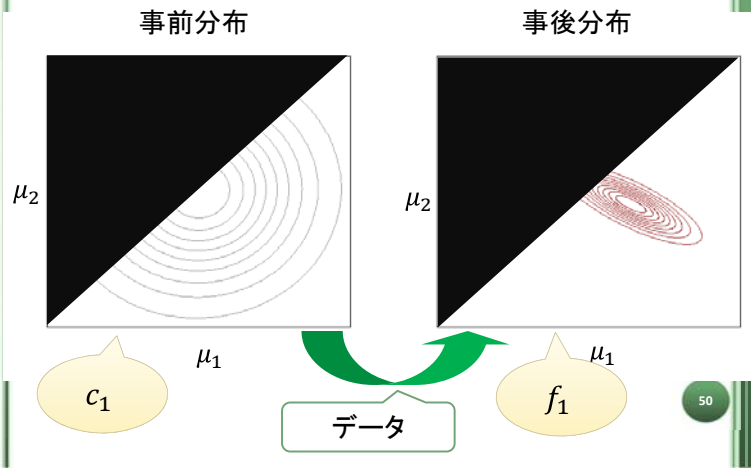
48

情報仮説のベイズファクター

- c_i：無制約仮説H_aの事前分布のうち、情報仮説H_iと一致する割合（モデルの複雑さ, complexity）
- f_i：無制約仮説H_aの事後分布のうち、情報仮説H_iと一致する割合（モデルの当てはまり, fit）

49

「H_1: $\mu_1 > \mu_2$」の下での事前分布と事後分布



50

情報仮説のベイズファクター

- c_i：無制約仮説H_aの事前分布のうち、情報仮説H_iと一致する割合（モデルの複雑さ, complexity）
- f_i：無制約仮説H_aの事後分布のうち、情報仮説H_iと一致する割合（モデルの当てはまり, fit）
- H_iとH_aを比較するベイズファクターは

$$BF_{ia} = \frac{f_i}{c_i}$$

51

（Klugkist, Laudy, and Hoijtink, 2005, *Psych Meth*）

結果

- H_1: $\mu_{LD} < \mu_{DM} < \mu_{LL}$　夜が明るいほど体重は増加する
- H_2: $\mu_{LD} < \{\mu_{DM}, \mu_{LL}\}$　夜が暗くないと、体重は増加する
- H_a: $\mu_{LD}, \mu_{DM}, \mu_{LL}$　とくに一貫した関係はない

	H_1	H_2	H_u
c_i (complexity)	0.1667	0.3333	—
f_i (fit)	0.9277	0.9349	—
BF_{iu} (Bayes factor)	5.57	2.80	1.00
PMP_i (Posterior model probability)	0.59	0.30	0.11

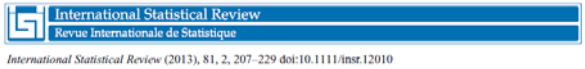
詳細・プログラム → 岡田（印刷中）基礎心研 (上はTab 3)
Hoijtink (2013, Chapman&Hall/CRC; 2011, Springer)

52

事前分布の影響

- でもベイズ推定って事前分布をどうするの？

➡あるクラスの情報仮説（同等集合equivalent setに属するもの）では、無情報事前分布を利用すれば結果に事前分布が影響しない　(Hoijtink, 2013, *Int Stat Rev*)



International Statistical Review (2013), 81, 2, 207–229 doi:10.1111/insr.12010

Objective Bayes Factors for Inequality Constrained Hypotheses

Herbert Hoijtink[1,2]

53

提案

Type S Error

- 情報仮説の評価

「十分に複雑」な

- 統計モデルの構築・評価
  - 感度分析
  - 事後予測チェック

Type M Error

54

統計モデルとは

- 確率的現象としてのデータを生み出す真のメカニズムを、確率分布を用いて表現(近似)したもの

"All models are wrong, but some are useful"
— George E. P. Box

- H_0やH_1ではなく、**役に立つ**モデルを構築・評価したい

55

(pic: wikipedia)

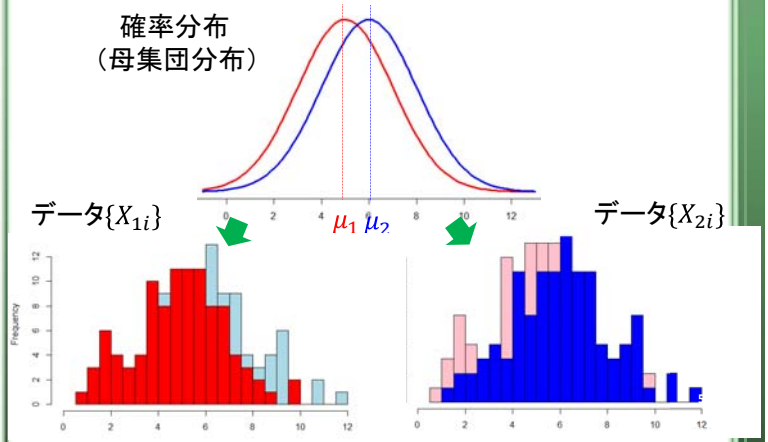
統計モデルとは

- 確率的現象としてのデータを生み出す真のメカニズムを、確率分布を用いて表現(近似)したもの
- 例:独立な2群のt検定のモデル

$$X_1 \sim N(\mu_1, \sigma^2)$$
$$X_2 \sim N(\mu_2, \sigma^2)$$

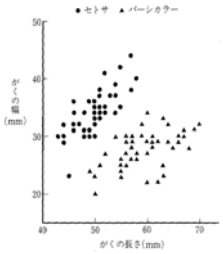
56

t検定のモデル(図示)



KISS: keep it simple and stupid....? (Robert Axelrod, 1997)

- 単純なモデルは、仮定が少ないぶん、頑健と言われる
- しかし、適切に情報を利用すること、頭を使うことの重要性は変わらない. 単純すぎるモデルは、複雑すぎるモデルと同様に、誤りのもとである.

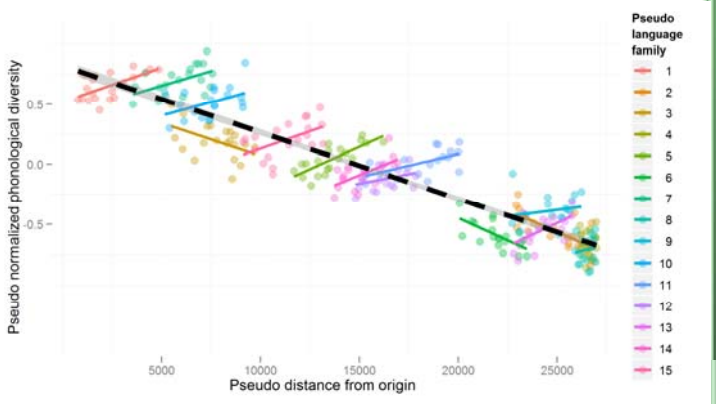


Fisherのアヤメデータ

とくに調査・観察データでは重要
例:層別相関

58

シンプソンのパラドックス



Jaeger et al. (2011, *Linguist Typol*) Fig 2.

除外変数バイアス (omitted variable bias)

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

が真のモデルなのに、説明変数x_2を含めずに

$$y = \beta_1 x_1 + \varepsilon$$

を使ってしまった場合

- $\beta_2 = 0$ or x_1とx_2が無相関 → バイアスはない
- そうでなければ**β_1の推定量にバイアスがある**

β_1	$Cor(x_1, x_2) > 0$	$Cor(x_1, x_2) < 0$
$\beta_2 > 0$	正のバイアス	負のバイアス
$\beta_2 < 0$	負のバイアス	正のバイアス

過剰変数の場合

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

が真のモデルなのに、説明変数x_3を含めて

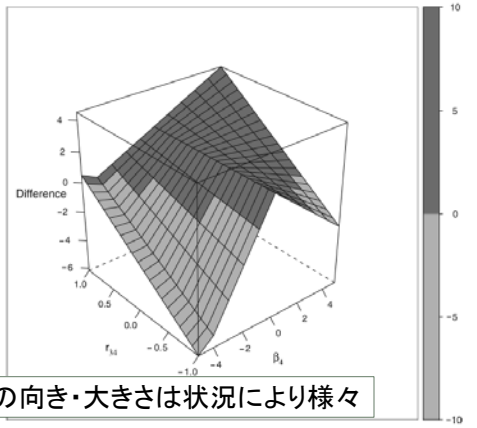
$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

を使ってしまった場合

- 推定にバイアスはない
  （が、推定量の分散は大きくなる＝効率的でなくなる）
- 説明変数の不足は、説明変数の過剰よりも深刻

61

除外変数バイアス (Clarke, 2005, *CMPS*, Fig 1)



バイアスの向き・大きさは状況により様々

FIGURE 1 The effect of β_4 and r_{34} on the difference in the absolute values of the two biases.

62

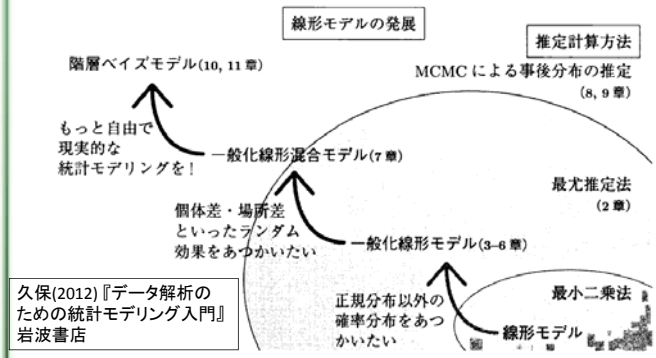
統計モデルの高度化

- 複雑な現象をモデリング・予測するためには、適切な統計モデルを用いる必要がある
- 統計モデルの一般化・包括化が進んでいる
  - **GLLAMM**
    ＝一般化線形モデル
    ＋潜在変数モデル
  - **セミパラメトリックモデル**
    関心のある部分はパラメトリック
    そうでない部分は
    ノンパラメトリック

cf. 星野 (2009)
『調査観察データの統計科学』
岩波書店

63

モデルの複雑化とMCMC法



久保(2012)『データ解析のための統計モデリング入門』岩波書店

図 1.2 線形モデルを発展させる説明のプラン．まずポアソン分布や二項分布を使った一般化線形モデル(GLM)を導入し，それを現実的なデータ解析に使えるように階層ベイズモデル化する．

64

予測の視点

- 手元のデータに(だけ)完全に当てはまるモデルは、いくらでも作れてしまう
  - 帯域幅と忠実度のジレンマ(Cronbach & Gleser, 1965)
  - 汎用性のあるモデルをどう選ぶか？

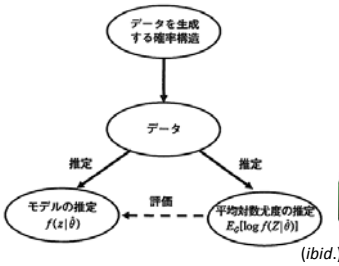
➡アイディア：統計的モデリングの真の目的は、現在のデータの忠実な記述や、真の分布の推定ではなく、将来得られるデータをできるだけ正確に予測すること
  - Akaike (1974, *IEEE TAC*), 赤池(1995, 朝倉書店)

65

AICは予測の指標

$$AIC = -2\log \mathrm{L}(\theta|X) + 2k$$

- AICは、最尤法によって推定したモデルを予測の観点から評価したことで、適用範囲の広い柔軟な指標となった(小西・北川, 2004, 朝倉書店)
  - ただし漸近的な指標
    (N→∞)



(*ibid.*)

統計モデルの高度化

- よい統計モデルはどのように選択できるか？
  - 統計モデルに関しても「研究者の自由度」が存在することになる
- モデル評価指標（情報量規準、適合度指標、ベイズファクター、etc）
  - 便利だが、意味と限界を意識して使うべき
- より簡便で汎用的な方法
  - 感度分析（sensitivity analysis）
  - 事後予測チェック（posterior predictive check）

67

感度分析（sensitivity analysis）

- 分析モデルを、ほかの合理的なモデルに変えたときに、結論はどれだけ変わってしまうのか？

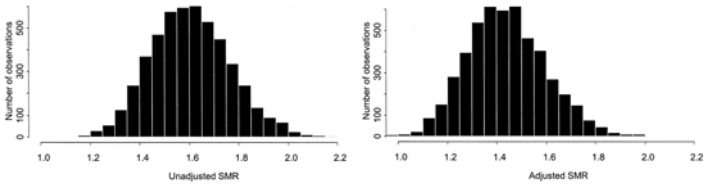
(Gelman et al, 2013, *CRC*)

- データが少し変わったときに、結論はどれだけ変わってしまうのか？
  - 交差検証法、leave-one-outなど

68

感度分析の例1（Steenland & Greenland, 2004, *Am J Epidemiol*）

- 4,624名の労働者のコホート研究において、モデルから推定した標準化死亡率（シリカ暴露群vs非暴露群）
- 左：喫煙の影響を考慮しない場合
- 右：喫煙の影響を考慮する場合

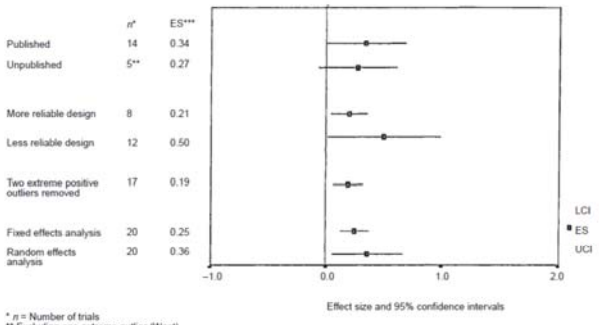


- いずれの場合でも標準化死亡率は暴露群で高く、その割合は点推定値で約50%増ほど

69

感度分析の例2（Sheard & Maguire, 1999, *Brit J Cancer*）

- 心理学的介入の、がん患者の抑うつに対する効果のメタ分析



* *n* = Number of trials
** Excluding one extreme outlier (West)
*** ES = Effect size (All random effects analysis with exception of fixed effects statistic) Vertical line is zero
Figure 6 Sensitivity analysis: depression

70

事後予測チェックの考え方

- よいモデルならば、そのモデルから生成された将来のデータは、観測データと似ているだろう
- 事後予測分布と観測データの整合性が十分であることを、モデルの必要条件としよう

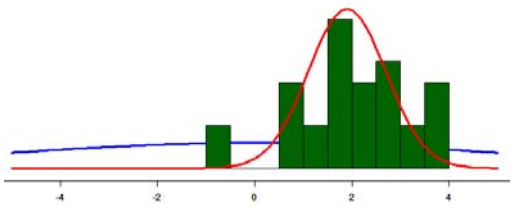
- アイディア：Guttman (1967, *JRSS-B*)
- 事後予測分布をモデルチェック・モデル評価に応用：Rubin (1981, *J Educ Stat*; 1984, *Ann Stat*)
- Gelman et al. (1996, *Stat Sinica*)：モデルチェックのための統計量の提供
- Bayarri & Berger (2000, *JASA*)：部分事後予測チェック（客観ベイズ）

71

(ベイズ統計学からみた)統計的推論

事後分布 posterior　　データ分布 尤度 likelihood　　事前分布 prior

$$p(\theta|X) \propto p(X|\theta) \cdot p(\theta)$$



72

(ベイズ統計学からみた)統計的推論

事前分布$p(\theta)$

データX

事後分布$p(\theta|X)$

将来のデータX_{rep}

事後予測分布 $p(X_{rep}|X)$

73

例1：スキージャンプの回帰予測

- スキージャンプ競技において、1回目の飛距離のデータから2回目の飛距離のデータを線形予測
- ソチオリンピック・男子ラージヒル競技における2回目も飛んだ30名のデータ（FISウェブサイトより）

http://data.fis-ski.com/dynamic/results.html?sector=JP&raceid=3854

74

データ

```
> jumpdata
                  name year nation jump1 jump2 points
1           STOCH Kamil 1987    POL 139.0 132.5  278.7
2         KASAI Noriaki 1972    JPN 139.0 133.5  277.4
3           PREVC Peter 1992    SLO 135.0 131.0  274.8
4        FREUND Severin 1988    GER 138.0 129.5  272.2
5      FANNEMEL Anders 1991    NOR 132.0 132.0  264.3
6         KRAUS Marinus 1991    GER 131.0 140.0  257.4
7  SCHLIERENZAUER Gregor 1990    AUT 132.5 130.5  255.2
8       HAYBOECK Michael 1991    AUT 134.0 125.5  254.7
9             ITO Daiki 1985    JPN 137.5 124.0  252.5
10       SHIMIZU Reruhi 1993    JPN 130.0 134.5  252.2
11      KOIVURANTA Anssi 1988    FIN 131.5 121.5  250.6
12          KOT Maciej 1991    POL 126.0 123.5  250.4
13        TAKEUCHI Taku 1987    JPN 132.5 122.5  249.3
14   DESCHWANDEN Gregor 1991    SUI 134.5 123.0  247.4
15          ZIOBRO Jan 1991    POL 128.5 129.5  246.6
```

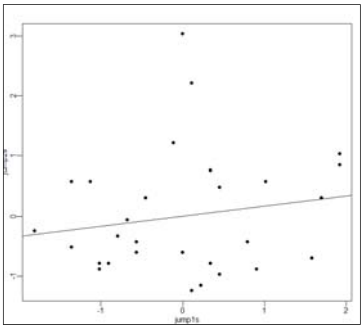
http://data.fis-ski.com/dynamic/results.html?sector=JP&raceid=3854

結果

- 標準化飛距離を使い、回帰モデル

$$y = \beta x + \alpha + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

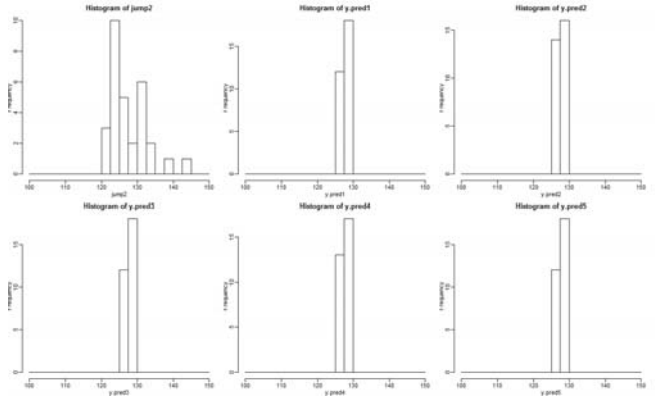
- によって1回目から2回目の飛距離を予測すると…



76

事後予測チェック（データ（左上）と5つの事後予測標本）

- 回帰モデルは、このデータの予測に適していない



77

例2: 死亡率への指数型モデル

（9つの事後予測標本）　（Gelman, Meng, & Stern, 1996, *Stat Sinica*）

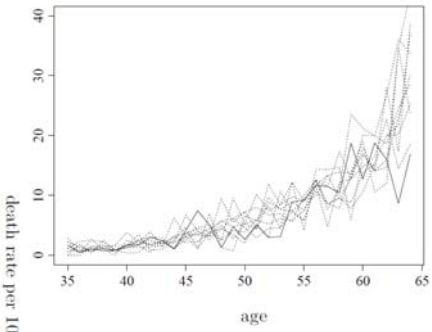


Figure 5. Nine draws from the posterior predictive distribution of mortality rates, corresponding to the nine draws of Figure 4, with the raw data (solid line) as a comparison.

78

Take-home messages

Sign(符号)

Type I error
Type II error
を気にするよりも

Type S error
Type M error
を気にしよう！

Magnitude
(大きさ)

79

Take-home messages

- 再現性の問題の一端は、非現実的なH_0を使う仮説検定への過度な依存にある
- 仮説検定の枠内で…停止規則、検定力分析など
- 仮説検定を離れて
  - パッケージ化された分析にデータを押し込むのではなく、状況にあった仮説・モデルでデータを分析する姿勢（研究デザインの重視 cf. 南風原, 2011, 東大出版）
  - 情報仮説の評価
  - モデル構築・評価（感度分析、事後予測チェック）

80

統計プログラムの論文誌

81

オープンデータの論文誌

82