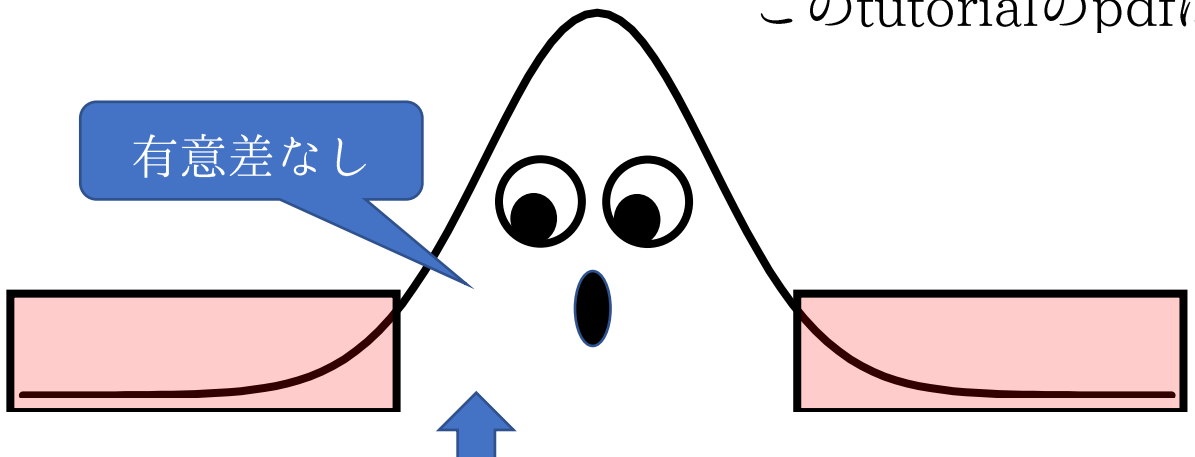
平均値の差の検定と効果量: 実験結果の適切な報告の仕方

(90分tutorial)

酒井　哲也（早稲田大学)

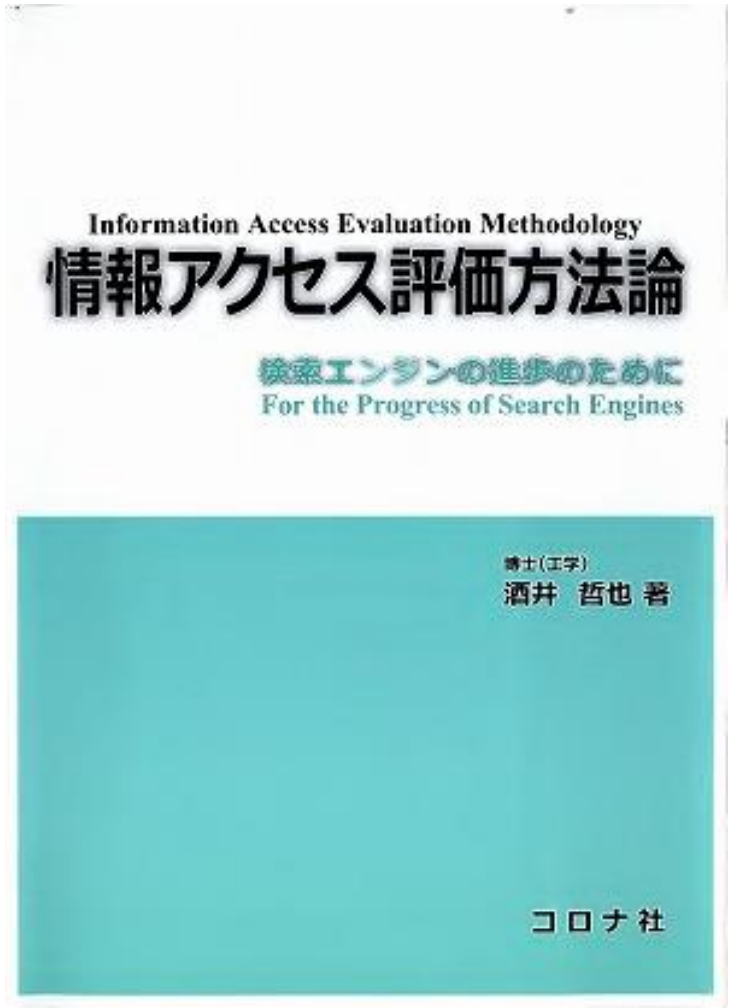
sakailab.com/tetsuya/

このtutorialのpdfは上記ページより入手可

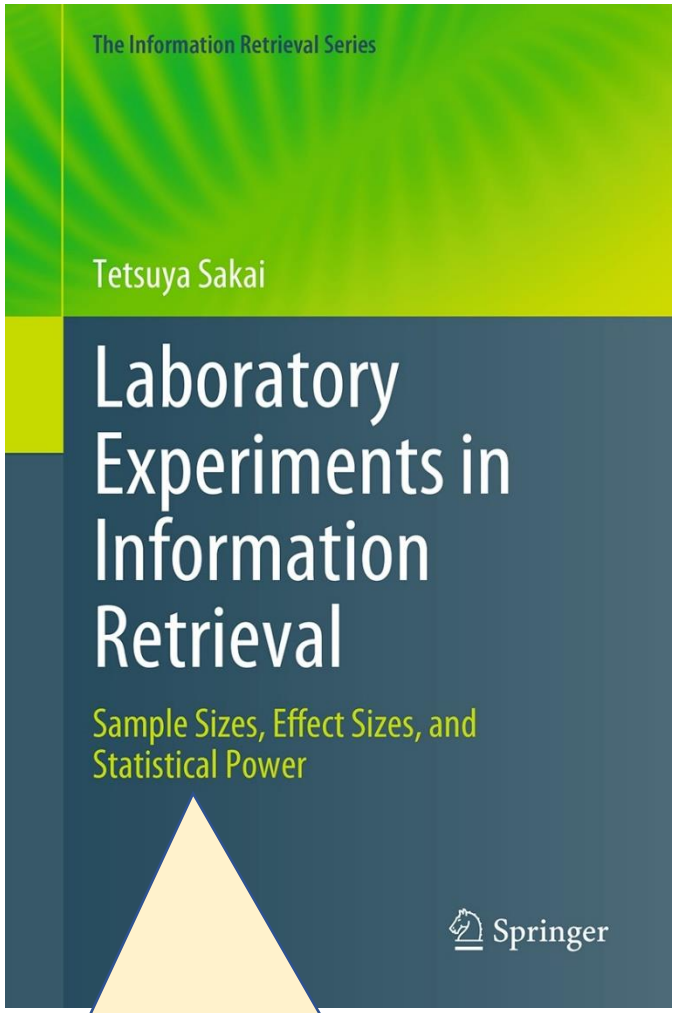


March 8, 2023@DEIM 2023, 岐阜

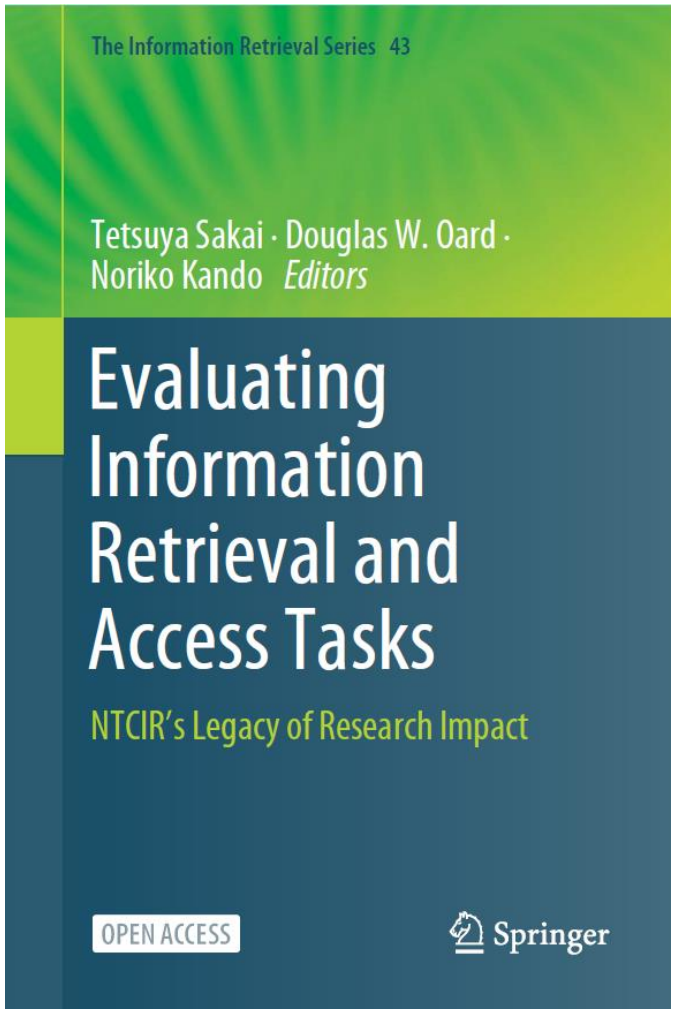
2015



2018



2020



このtutorialより詳しい内容はこの本で!
http://sakailab.com/leirbook/

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

William Sealy Gosset (1876-1937)

Ziliak and McCloskey: The Cult of Statistical Significance, p.3, 2008

"He worked at the Guinness Brewery in Dublin, where for most of his working life he was the head experimental brewer. He saw in 1905 the need for a small-sample test because he was testing varieties of hops and barley in field samples with N as small as four."

おいしいギネスビールの開発に取り組んでいた社員。
(サンプルサイズn=4くらいが彼の関心事だった!)

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Gossetは
Studentというペンネームで
t検定の基本的アイデアについて
発表

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form

ギネスビールを飲むときと
統計的検定をやるときは
Gossetのことを思い出してください

t検定の頑健性

- t検定の導出は母集団分布が正規分布であるという仮定からスタート。しかし!
- データが厳密に正規分布に従っていなくても基本的には適用可。

実際、正規性の仮定に依存しないランダム化検定（Part [IV]）と意外にも似た結論が得られる。(それでも心配ならランダム化検定か古典的ノンパラメトリック検定を行えばよい)

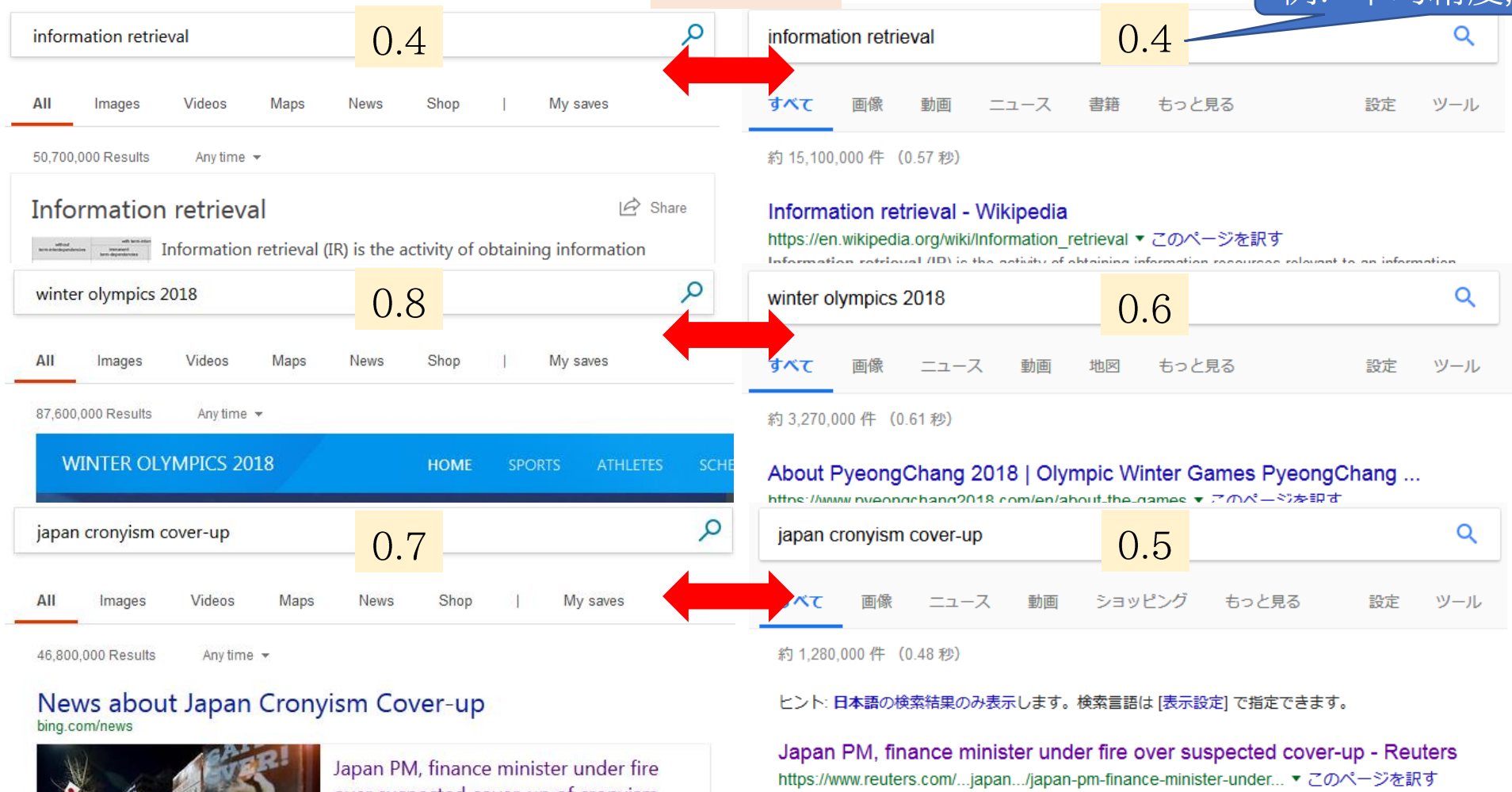
- とくにサンプルサイズがある程度大きくなると、中心極限定理によりサンプル平均はほぼ正規分布に従う。
- 「正規性の検定」⇒「t検定」という手順は適切か？Part III (ワインを飲みすぎた客の話) を聴いてから考えてみよう。

どちらの検索エンジンがよいか (対応のあるデータ)

基本的な検索評価指標について:
https://waseda.box.com/sakai14PROMISE
http://sakailab.com/iaembook/

n = 3

例：平均精度，nDCG



対応のあるデータのt検定（1）

x_{1j}：システム1，第jトピックのスコア

x_{2j}：システム2，第jトピックのスコア

各トピックのスコアが独立かつ

$$x_{1j} \sim N(\mu_1, \sigma_1^2), \quad x_{2j} \sim N(\mu_2, \sigma_2^2)$$

のとき、各スコアの差 $d_j = x_{1j} - x_{2j}$ について

正規分布の性質より $d_j \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

対応のあるデータのt検定（2）

$$d_j \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

(スコア差の)母分散

⇒

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{\sqrt{V_d / n}} \sim t(n-1)$$

自由度 ϕ =n-1

詳しくは http://sakailab.com/leirbook/ (Chapter 1)

これが母分散だったらtは標準正規分布に従う。
代わりにその不偏推定量を用いるためt分布に従う

ここで

サンプル平均

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j$$

不偏分散

$$V_d = \frac{\sum_{j=1}^{n}(d_j - \bar{d})^2}{n-1}$$

対応のあるデータのt検定（3）

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{\sqrt{V_d / n}} \sim t(n-1)$$

詳しくは http://sakailab.com/leirbook/ (Chapter 1)

コンピュータサイエンス系の実験では多くの場合両側検定がお奨め

両側検定:

$H_0 : \mu_1 = \mu_2$　　$H_1 : \mu_1 \neq \mu_2$

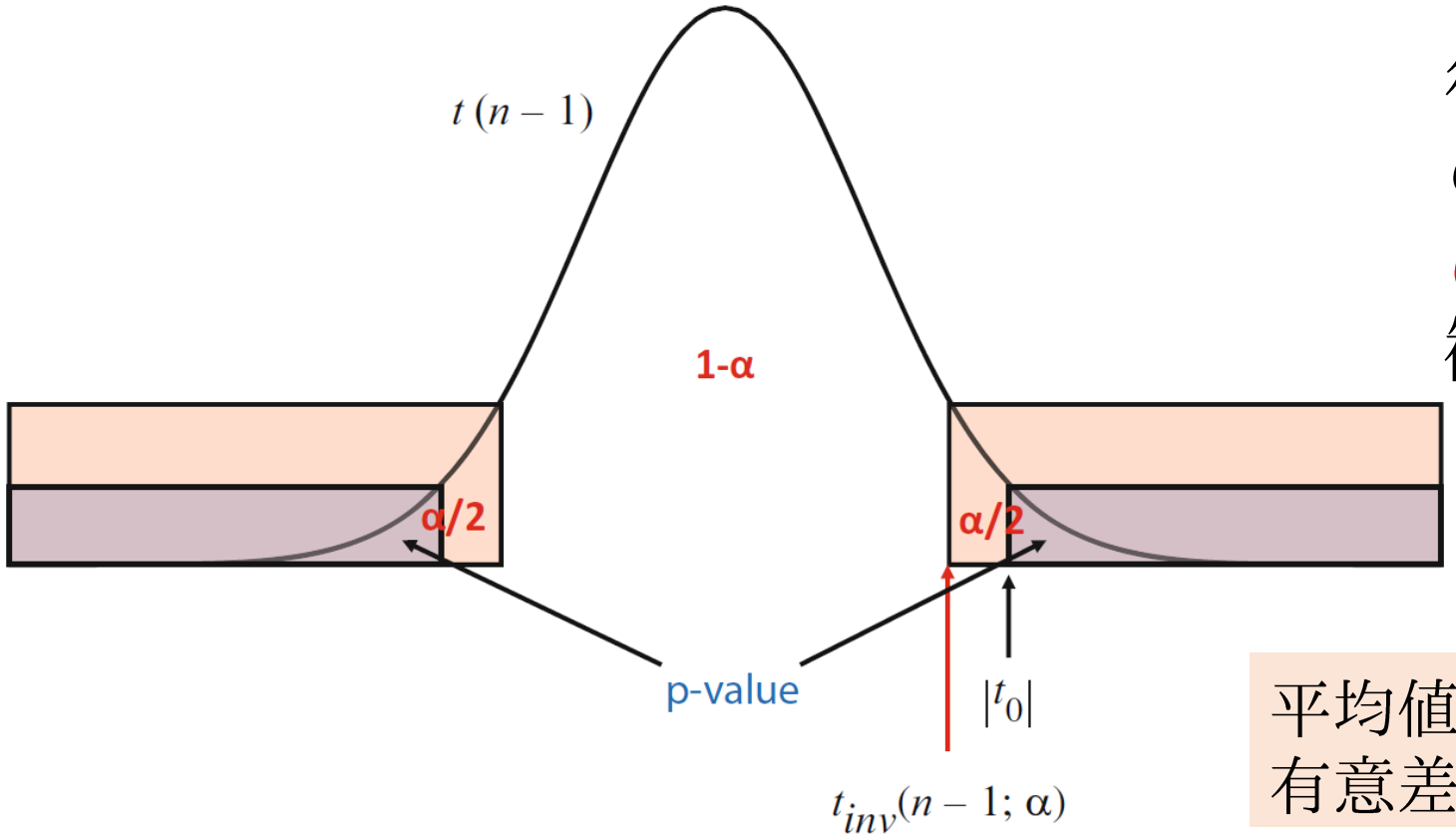
帰無仮説H_0 のもとでは以下が成り立つはず

検定統計量（t値）

$$t_0 = \frac{\bar{d}}{\sqrt{V_d / n}} \sim t(n-1)$$

対応のあるデータのt検定（4）

帰無仮説H0 のもとでは $t_0 = \dfrac{\bar{d}}{\sqrt{V_d/n}} \sim t(n-1)$



従って $|t_0| \geq t_{inv}(n-1;\alpha)$
のとき(H0のもとで
ほぼありえないことが
観測されたので)H0を棄却

「t分布に従っとらんのとちゃうか」

平均値の差は有意水準αにおいて
有意差あり（母平均はおそらく異なる）

http://www.f.waseda.jp/tetsuya/20topics3runs.mat.csv

m=3 systems

n=20 topics

	A	B	C
1	System1	System2	System3
2	0.4695	0.3732	0.3575
3	0.2813	0.3783	0.2435
4	0.3914	0.3868	0.3167
5	0.6884	0.5896	0.6024
6	0.6121	0.4725	0.4766
7	0.3266	0.233	0.2429
8	0.5605	0.4328	0.4066
9	0.5916	0.5073	0.4707
10	0.4385	0.3889	0.3384
11	0.5821	0.5551	0.4597
12	0.2871	0.3274	0.2769
13	0.5186	0.5066	0.4066
14	0.5188	0.5198	0.3859
15	0.5019	0.4981	0.4568
16	0.4702	0.3878	0.3437
17	0.329	0.4387	0.2649
18	0.4758	0.4946	0.4045
19	0.3028	0.34	0.3253
20	0.3752	0.4895	0.3205
21	0.2796	0.2335	0.224

20topics3runs.mat.csv をRで読み込む

```
> matURL <- "http://www.f.waseda.jp/tetsuya/20topics3runs.mat.csv"
> mat <- read.csv( file=matURL, header=TRUE )
> mean( mat$System1 )
[1] 0.45005
> mean( mat$System2 )
[1] 0.427675
```

サンプル平均は
システム1のほうが高い

Rで対応のあるデータのt検定を実行

```
> t.test(mat$System1, mat$System2, paired=TRUE)

        Paired t-test

data:  mat$System1 and mat$System2
t = 1.3101, df = 19, p-value = 0.2058
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01337109  0.05812109
sample estimates:
mean of the differences
                 0.022375
```

t値

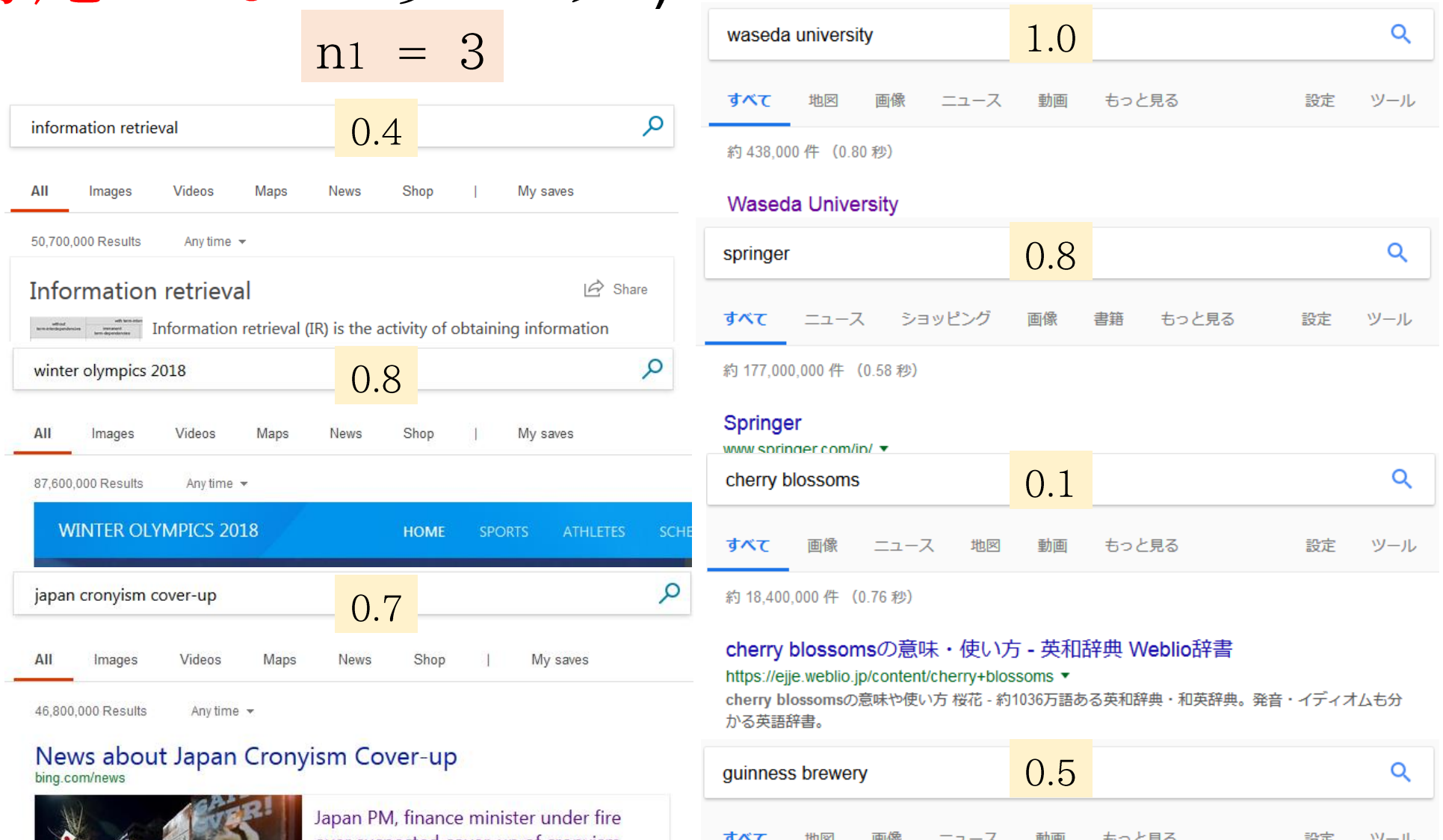
自由度φ

両側検定

どちらの検索エンジンがよいか
(対応のないデータ)

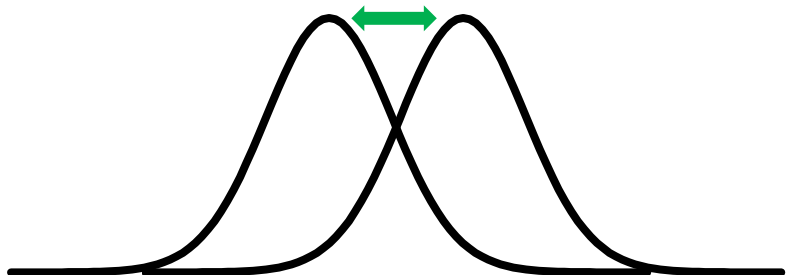
n1 = 3

n2 = 4



対応のないデータのt検定（1）

形は同じ、位置は違うかも



x1j: システム1，第jトピックのスコア
x2j: システム2，第jトピックのスコア
各トピックのスコアが独立かつ

$$x_{1j} \sim N(\mu_1, \sigma^2), \quad x_{2j} \sim N(\mu_2, \sigma^2)$$

以下、等分散性を仮定したStudentのt検定を議論。
等分散性を仮定しないWelchのt検定もあるが、
必ずしも後者のほうがよいというわけではない。
前者は母分散やサンプルサイズが極端に違わない限り頑健

詳しくは 永田靖: 入門 統計解析法，日科技連，1992
もしくは http://sakailab.com/leirbook/ (Chapter 2)

対応のないデータのt検定（2）

$$x_{1j} \sim N(\mu_1, \sigma^2), \quad x_{2j} \sim N(\mu_2, \sigma^2)$$

$$\Rightarrow \quad t = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet} - (\mu_1 - \mu_2)}{\sqrt{V_p(1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2)$$

詳しくは永田靖：入門　統計解析法，日科技連，1992
もしくは　http://sakailab.com/leirbook/　(Chapter　1)

サンプル平均

$$\bar{x}_{1\bullet} = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}, \qquad \bar{x}_{2\bullet} = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}$$

$$S_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_{1\bullet})^2, \qquad S_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_{2\bullet})^2,$$

併合分散

$$V_p = \frac{S_1 + S_2}{n_1 + n_2 - 2}$$

対応のないデータのt検定（3）

$$t = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet} - (\mu_1 - \mu_2)}{\sqrt{V_p(1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2)$$

H_0 ： $\mu_1 = \mu_2$　　H_1 ： $\mu_1 \neq \mu_2$

帰無仮説H_0 のもとでは以下が成り立つはず

検定統計量（t値）

$$t_0 = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet}}{\sqrt{V_p(1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2)$$

従って $|t_0| \geq t_{inv}(n_1+n_2-2; \alpha)$ のときH_0を棄却

平均値の差は有意水準αにおいて
有意差あり（母平均はおそらく異なる）

Rで対応のないデータのt検定を実行

```
> t.test(mat$System1, mat$System2, var.equal = TRUE)

        Two Sample t-test

data:  mat$System1 and mat$System2
t = 0.6338, df = 38, p-value = 0.53
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04909139  0.09384139
sample estimates:
mean of x mean of y
 0.450050  0.427675
```

データが対応しているという情報を
活用した場合のp値と比べてみよう

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

半世紀以上炎上している統計的検定(1)

“Little advancement in the teaching of statistics is possible […] until the literature and classroom be rid of terms so deadening to scientific enquiry as null hypothesis, population […]”

Deming, W.E.: On Probability as a Basic for Action, American Statistician, 29(4), 1975.

超訳: 学校で帰無仮説とか母集団とかアホらしいことを教えるのはやめちまえ

半世紀以上炎上している統計的検定(2)

“Despite the stranglehold that hypothesis testing has on experimental psychology, I find it difficult to imagine a less insightful means of transiting from data to conclusions.”

Loftus, G.R.: On the Tyranny of Hypothesis Testing in the Social Sciences, Contemporary Psychology, 36(2), 1991.

超訳: データから結論を得るためのこれよりアホな方法はちょっと考えらんない

半世紀以上炎上している統計的検定(3)

“And we, as teachers, consultants, authors, and otherwise perpetrators of quantitative methods, are responsible for the ritualization of null hypothesis significance testing (NHST; I resisted the temptation to call it statistical hypothesis inference testing) to the point of meaninglessness and beyond.”

Cohen, J.: The Earth is Round (p<.05), American Psychologist, 49(12), 1994.

超訳: 検定を無意味な儀式にしちゃったのはワシらのせい
NHSTというより○○○○

統計的検定の問題点1

- 多くの人が検定を正しく理解していない and/or 正しく利用していない
- American Statistical Associationの2016年の声明の一部:

"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."

Gossetの論文から100年以上たってもまだこんな注意が必要 orz

p-value = Pr(D+|H)
not Pr(H|D)!

p値を正しく説明できますか?
「帰無仮説のもとで、観測されたデータもしくはさらに極端なことが起こる確率」

研究者は「観測されたデータのもとで自分の仮説が正しい確率」を知りたかったりする

統計的検定の問題点2

- 二元論「有意差あり？ 有意差なし？　どっちなの？」

p=0.049 ⇒ 有意差あり、論文が書ける？
p=0.051 ⇒ 有意差なし、論文が書けない？

- より重要なのは

どれくらいの差があるのか
その差はわれわれにとってどのような意味があるのか

実用的有意差

論文に「p=0.049」と書いたほうが「p<0.05」と書くよりは情報量が多いが、それでもまだ不十分!

統計的検定の問題点3

差の大きさ、本当の関心事

P-value ＝ f(サンプルサイズ，効果量)

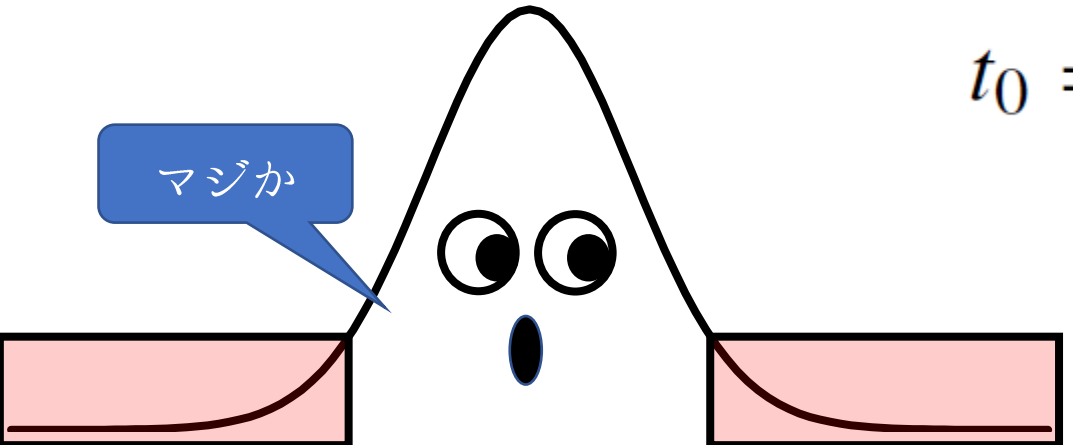
効果量大 ⇒ p-value小

サンプルサイズ大 ⇒ p-value小

効果量（差が標準偏差いくつぶんか）が大 ⇒ t0 が大 ⇒ p値が小
はいいとして…

例：対応のあるデータのt値

$$t_0 = \frac{\bar{d}}{\sqrt{V_d/n}} = \sqrt{n}\frac{\bar{d}}{\sqrt{V_d}}$$



サンプルサイズが大 ⇒ t0 が大 ⇒ p値が小
つまり十分大きなサンプルをとれば小さな差も有意になる

効果量(ここでは差が標準偏差いくつぶんか)の例

- 対応のあるデータのt検定より

$$d_{paired} = \frac{\bar{d}}{\sqrt{V_d}} \qquad \left(t_0 = \sqrt{n} d_{paired} \right)$$

- 対応のないデータのt検定より

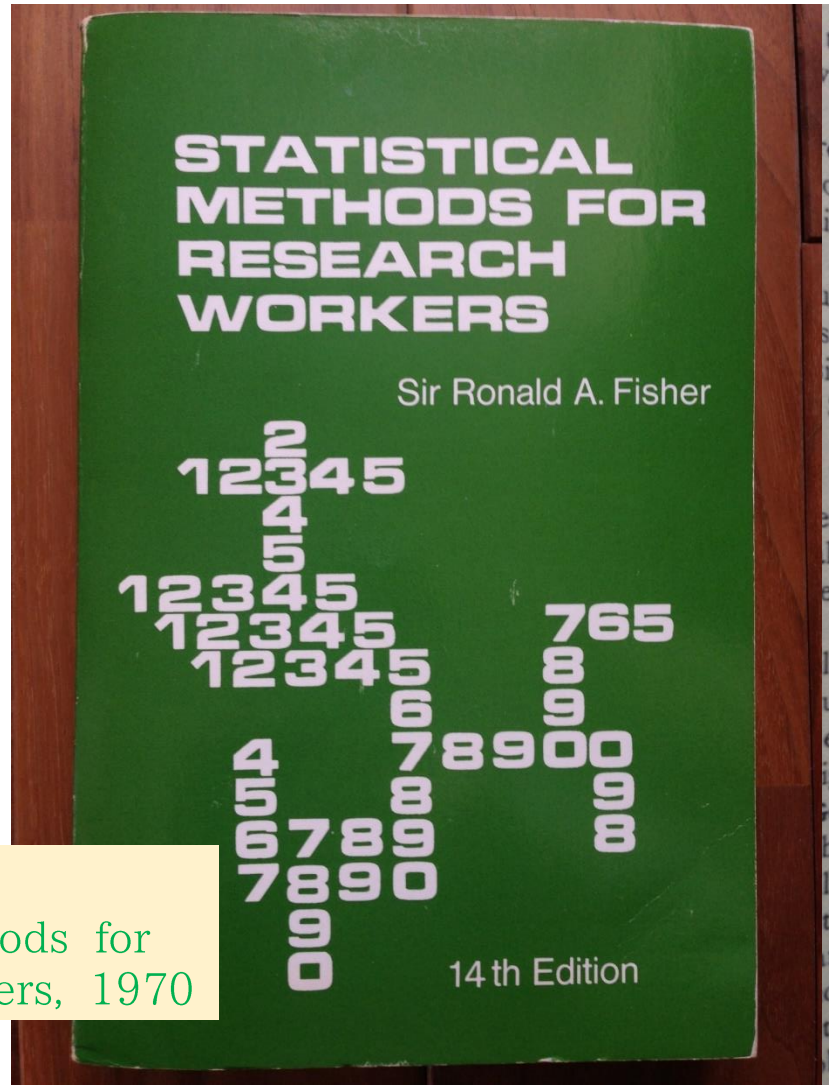
Hedge’s g

$$g = \frac{\bar{x}_{1\bullet} - \bar{x}_{2\bullet}}{\sqrt{V_p}} \qquad \left(t_0 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} g \right)$$

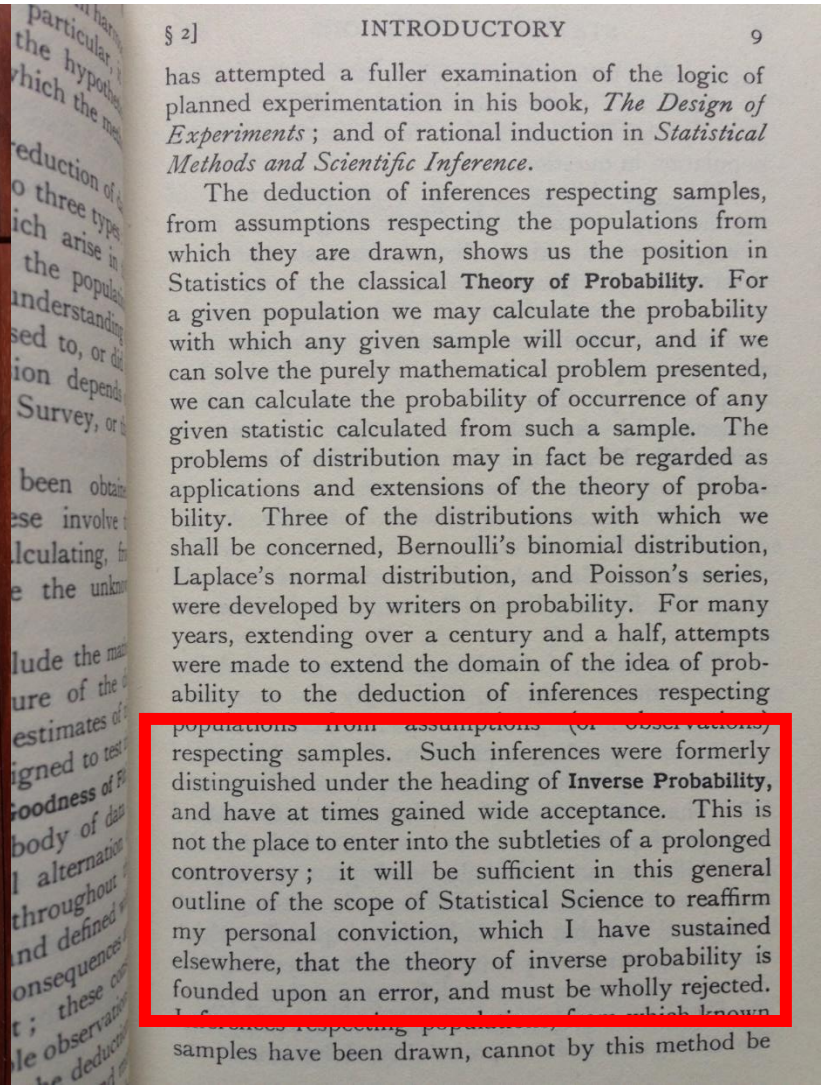
Glass’s Δもおすすめ: http://sakailab.com/leirbook/ (Chapter 5)

統計的検定に代わるアプローチ（の例）

Fisherがおそらく
大嫌いだった
Bayes統計（事後確率）



Fisher, R.A.:
Statistical Methods for
Research Workers, 1970



§ 2] INTRODUCTORY 9

has attempted a fuller examination of the logic of planned experimentation in his book, *The Design of Experiments*; and of rational induction in *Statistical Methods and Scientific Inference*.

The deduction of inferences respecting samples, from assumptions respecting the populations from which they are drawn, shows us the position in Statistics of the classical **Theory of Probability.** For a given population we may calculate the probability with which any given sample will occur, and if we can solve the purely mathematical problem presented, we can calculate the probability of occurrence of any given statistic calculated from such a sample. The problems of distribution may in fact be regarded as applications and extensions of the theory of probability. Three of the distributions with which we shall be concerned, Bernoulli's binomial distribution, Laplace's normal distribution, and Poisson's series, were developed by writers on probability. For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences were formerly distinguished under the heading of **Inverse Probability,** and have at times gained wide acceptance. This is not the place to enter into the subtleties of a prolonged controversy; it will be sufficient in this general outline of the scope of Statistical Science to reaffirm my personal conviction, which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected.

samples have been drawn, cannot by this method be

Bayes統計を一枚で説明（できない）

詳しくは 豊田秀樹: はじめての統計データ分析，朝倉書店，2016 など

θ：母数（例えば母平均。古典的検定では定数だが、Bayes統計では変数。）
x: データ

θ の事前確率
(誰にもわからないので「適当に」仮定)

尤度

θ の事後確率 $f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int_{-\infty}^{+\infty} f(x|\theta)f(\theta)d\theta}$

みたことあるはずの定理!

- 歴史的には事後確率の計算が困難だったが、
MCMC (Markov Chain Monte Carlo) 法の発展により今やPCでも計算可能
- 母数や任意の仮説Hに対するPr(H|D) (p.25)が推定できる (p値より直感的?)

（が少なくとも私の周辺の研究コミュニティではBayes統計が盛り上がる気配はない…）

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

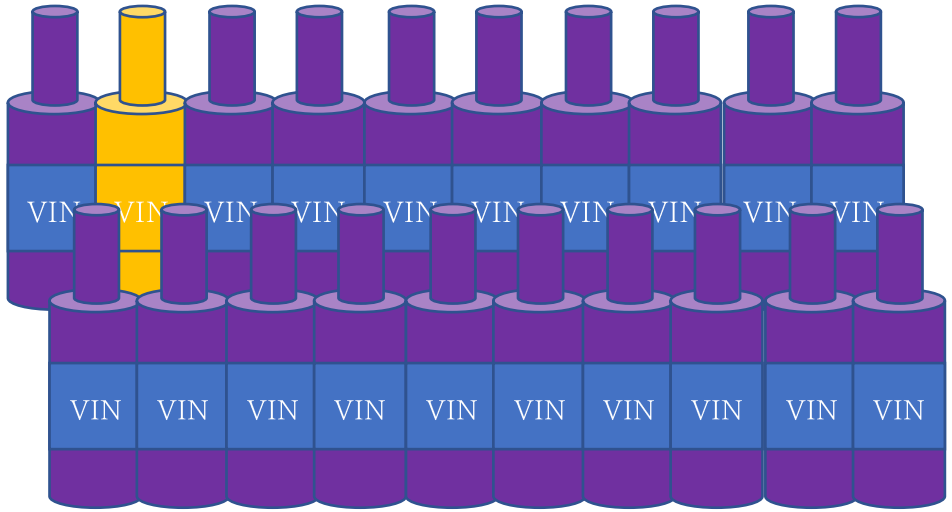
(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

今、m個のシステムの平均値を比較しており、
全k=m(m-1)/2 システム対の差に興味がある。
そこでさきほど復習したt検定をk回繰り返した。

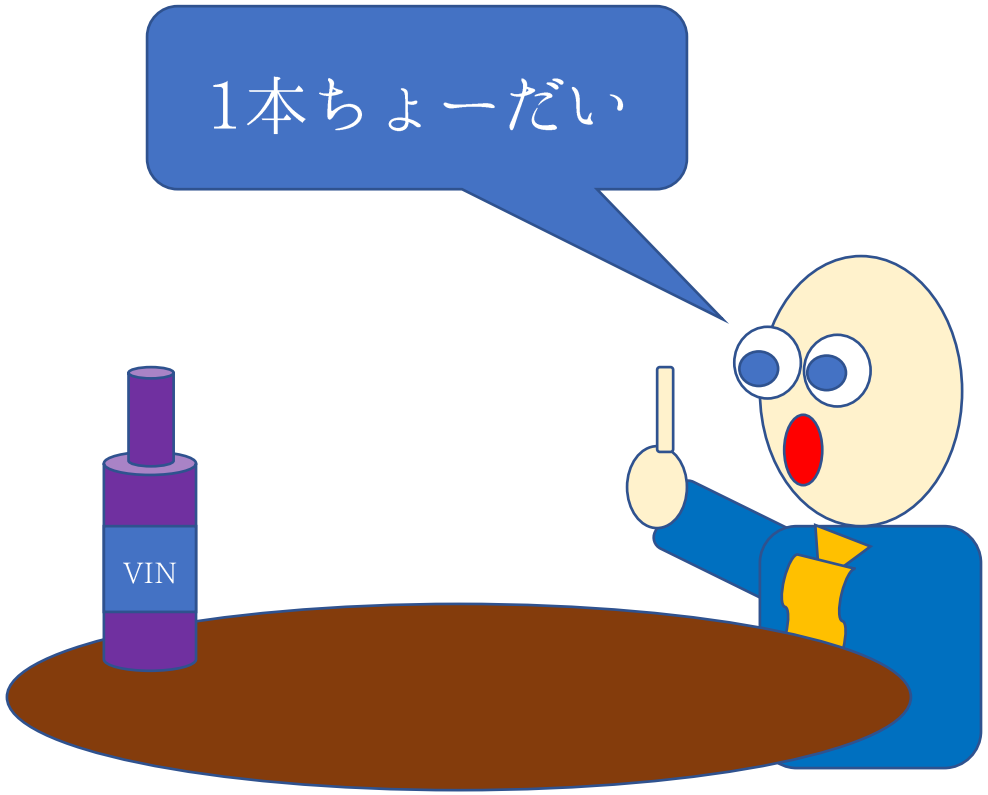
このやり方は正しいか？

レストランにワインセラーがある(ワインは無限にある)。ただし20本に1本は酸っぱいワインである。



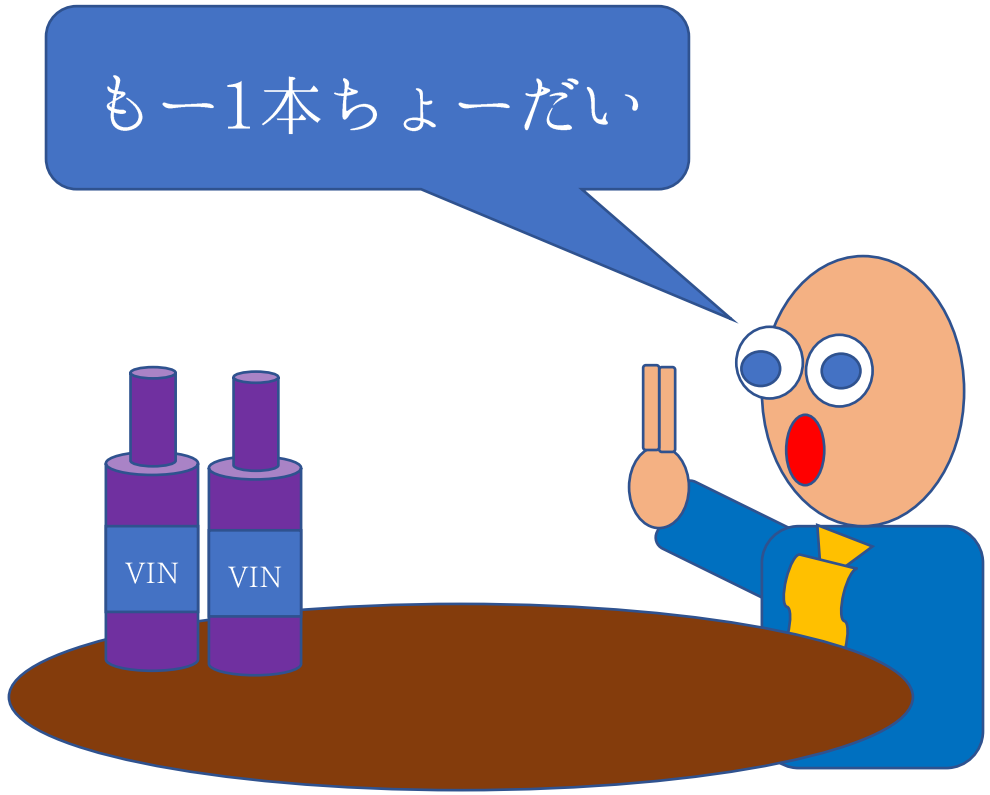
ワインセラーからランダムにワインを取り出す。
それが酸っぱい確率は
1/20 ＝ 0.05

お客さんがワインを1本注文した場合



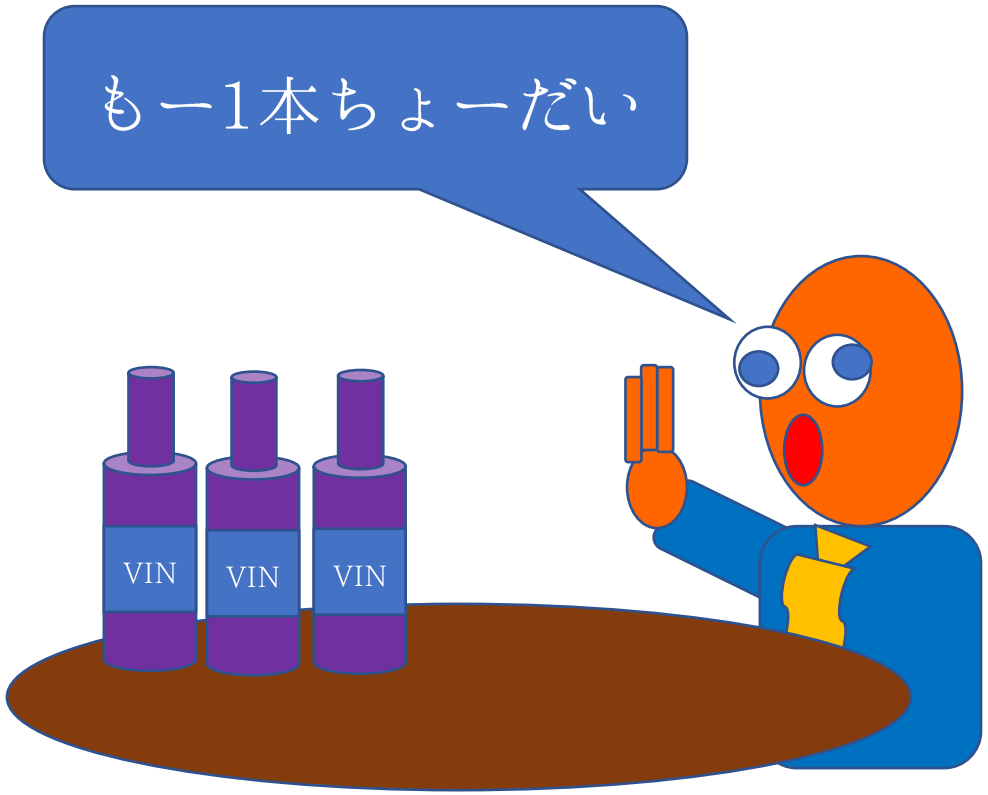
この人に酸っぱいワインを飲ませてしまう確率は
0.05

お客さんがワインを2本注文した場合



この人に酸っぱいワインを飲ませてしまう確率は
1-P(2本ともセーフ)
=1-0.95^2
=0.0975

お客さんがワインを3本注文した場合



この人に酸っぱいワインを飲ませてしまう確率は
1-P(3本ともセーフ)
=1-0.95^3
=0.1426

全体（ファミリー）としての第一種の誤り

レストランは個々のワインが酸っぱい確率ではなく、
k本のワインを出した場合にお客さんが酸っぱいワインを
体験してしまう確率を例えば5%以下に抑えたい。

われわれ研究者の場合

母平均に差がないのにあわてて差があると結論づけてしまう確率

第一種の誤り確率を α に抑えた上でt検定をk=m(m-1)/2回繰り返すと、ファミリーとしての第一種の誤り確率は $1-(1-\alpha)^k$ まで膨れ上がる可能性がある。

k対のうちひとつでもあわてて差があると結論づけてしまう確率

抑えたいのはむしろ後者!

例: $\alpha=0.05$, k=10

⇒ ファミリーとしての第一種の誤り確率≒40%!

第二種の誤り確率（母平均に差があるのにぼんやりして見逃してしまう確率）は β で表す。Part VI参照

t検定×k回 ＋ Bonferroni補正 はちょっと古い

Bonferroni補正: αをkで割っておく。もしくはp値にkをかけてからαと比較する。

“The old fashioned approach was to use Bonferroni’s correction: […] you divide your α value by the number of comparisons you have done. […] Bonferroni’s correction is very harsh and will often throw out the baby with the bathwater. […] The modern approach is […] to use the wonderfully named Tukey’s honestly significant differences”

HSD

Crawly, M.J.: Statistics: An Introduction Using R (Second Edition), Wiley, 2015.

Bathwater ＝ 母平均に差がないシステム対 (捨てたいもの)
Baby ＝ 母平均に差があるシステム対 (捨てたくないもの)

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

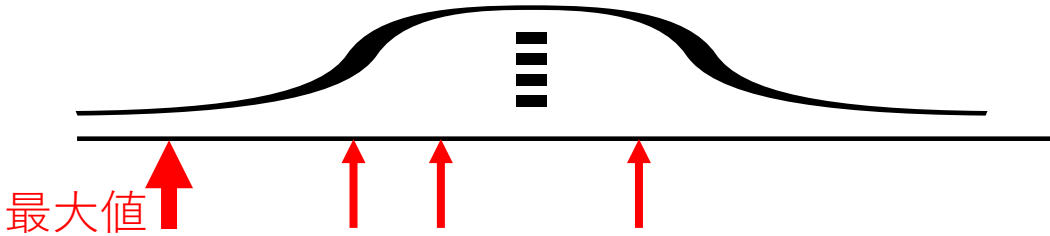
(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

Tukey HSD検定の原理



- t検定をk=m(m-1)/2回繰り返す代わりに、k個の検定統計量の最大値(最高・最低のシステムの差から計算)に着目。
- この最大値はt分布そのものではなくStudentised range分布に従う。
- 全k個の差について上記分布をもちいた検定を行う。
- 最大値が統計的に有意でないとき、残りの(k-1)個も統計的に有意でない。⇒ 最大値についてαを設定すれば、自動的にファミリーとしての第一種の誤り確率がαで抑えられる。

どこかのシステム間に差があるか?

Tukey HSD 検定の前に分散分析による検定を行う必要があるか?→ NO

詳しくは 永田靖: 多重比較法の実際, 応用統計学27(2), 1998 or http://sakailab.com/leirbook/ (Chapter 4) など参照

対応のあるデータのTukey HSD検定（1）

二元配置分散分析(繰り返しなし)のデータ構造

System	Per-topic scores			
1	x_{11}	x_{12}	...	x_{1n}
2	x_{21}	x_{22}	...	x_{2n}
⋮	⋮			
m	x_{m1}	x_{m2}	...	x_{mn}

システムiの
サンプル平均

$$\bar{x}_{i\bullet} = \sum_{j=1}^{n} x_{ij}/n$$

トピックjの
サンプル平均

$$\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{m} x_{ij}}{m}$$

対応のあるデータのTukey HSD検定（2）

$$S_{E2} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2 , \quad V_{E2} = \frac{S_{E2}}{\phi_{E2}}$$

帰無仮説 $H_{i,i'}$:

システムi, i’ の母平均は等しい

$$\phi_{E2} = (m-1)(n-1)$$

検定統計量 $t'_{i,i'} = \dfrac{\bar{x}_i - \bar{x}_{i'}}{\sqrt{V_{E2}/n}}$

Studentised range分布より

$|t'_{i,i'}| \geq q_{inv}(m, \phi_{E2}; \alpha)$ のとき $H_{i,i'}$ を棄却

対応のあるデータのTukey HSD検定（3）

```
> mat2 <- data.frame(Topic=1:20, mat)
> head(mat2, 3)
  Topic System1 System2 System3
1     1  0.4695  0.3732  0.3575
2     2  0.2813  0.3783  0.2435
3     3  0.3914  0.3868  0.3167
> tail(mat2, 3)
   Topic System1 System2 System3
18    18  0.3028  0.3400  0.3253
19    19  0.3752  0.4895  0.3205
20    20  0.2796  0.2335  0.2240
```

対応のあるデータのTukey HSD検定（4）

```
> library(tidyr)
> mat2tidy <- gather( mat2, key=System, value=Score, -Topic )
> head(mat2tidy, 3)
  Topic  System  Score
1     1 System1 0.4695
2     2 System1 0.2813
3     3 System1 0.3914
> tail(mat2tidy, 3)
   Topic  System  Score
58    18 System3 0.3253
59    19 System3 0.3205
60    20 System3 0.2240
```

「Topicの列以外について値を列挙せよ」

対応のあるデータのTukey HSD検定（5）

```
> TukeyHSD( aov(Score ~ factor(System) + factor(Topic), data=mat2tidy),
+ "factor(System)", ordered=TRUE )
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = Score ~ factor(System) + factor(Topic), data = mat2tidy)

$`factor(System)`
                      diff         lwr         upr     p adj
System2-System3 0.061470   0.02813169 0.09480831 0.0001829
System1-System3 0.083845   0.05050669 0.11718331 0.0000011
System1-System2 0.022375 -0.01096331 0.05571331 0.2428822
```

「出力のdiff（平均値の差）の値が負にならないよう、どちらからどちらを引くか決めて」

システム2,3間、1,3間のみ有意

対応のないデータのTukey HSD検定（1）

グループサイズが等しい場合

グループサイズが等しくない場合は http://sakailab.com/leirbook/ (Chapter 4)

グループサイズが等しい一元配置分散分析のデータ構造

システムiの
サンプル平均

$$\bar{x}_{i\bullet} = \sum_{j=1}^{n} x_{ij}/n$$

System	Per-topic scores
1	$x_{11}, x_{12}, \ldots, x_{1n}$
2	$x_{21}, x_{22}, \ldots, x_{2n}$
⋮	⋮
m	$x_{m1}, x_{m2}, \ldots, x_{mn}$

グループサイズが全てnだが対応なし

対応のないデータのTukey HSD検定（2）

$$S_{E1} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2 , \quad V_{E1} = \frac{S_{E1}}{\phi_{E1}}$$

帰無仮説 $H_{i,i'}$:

システムi, i’ の母平均は等しい

$$\phi_{E1} = m(n-1)$$

検定統計量 $t'_{i,i'} = \dfrac{\bar{x}_i - \bar{x}_{i'}}{\sqrt{V_{E1}/n}}$

$|t'_{i,i'}| \geq q_{inv}(m, \phi_{E1}; \alpha)$ のとき $H_{i,i'}$ を棄却

対応のないデータのTukey HSD検定（3）

```
> library(tidyr)
> mattidy <- gather(mat, key=System, value=Score)
> head(mattidy, 3)
   System  Score
1 System1 0.4695
2 System1 0.2813
3 System1 0.3914
> tail(mattidy, 3)
    System  Score
58 System3 0.3253
59 System3 0.3205
60 System3 0.2240
```

対応のないデータのTukey HSD検定（3）

```
> TukeyHSD( aov(Score ~ factor(System), data=mattidy), ordered=TRUE )
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = Score ~ factor(System), data = mattidy)

$`factor(System)`
                       diff           lwr       upr     p adj
System2-System3 0.061470 -0.019844676 0.1427847 0.1725122
System1-System3 0.083845  0.002530324 0.1651597 0.0418683
System1-System2 0.022375 -0.058939676 0.1036897 0.7862427
```

システム1,3間のみ有意

ランダム化(Tukey　HSD)検定

データの正規性などの仮定からスタートして分布を数学的に導出する古典的検定とは違い、観測されたデータをもとに計算機をぶんまわして分布を描く

簡単な実装例: https://research.nii.ac.jp/ntcir/tools/discpower-ja.html

NTCIR Project
Tools

Discpower

[ENGLISH] [NTCIR Home] [NTCIR Tools Home]

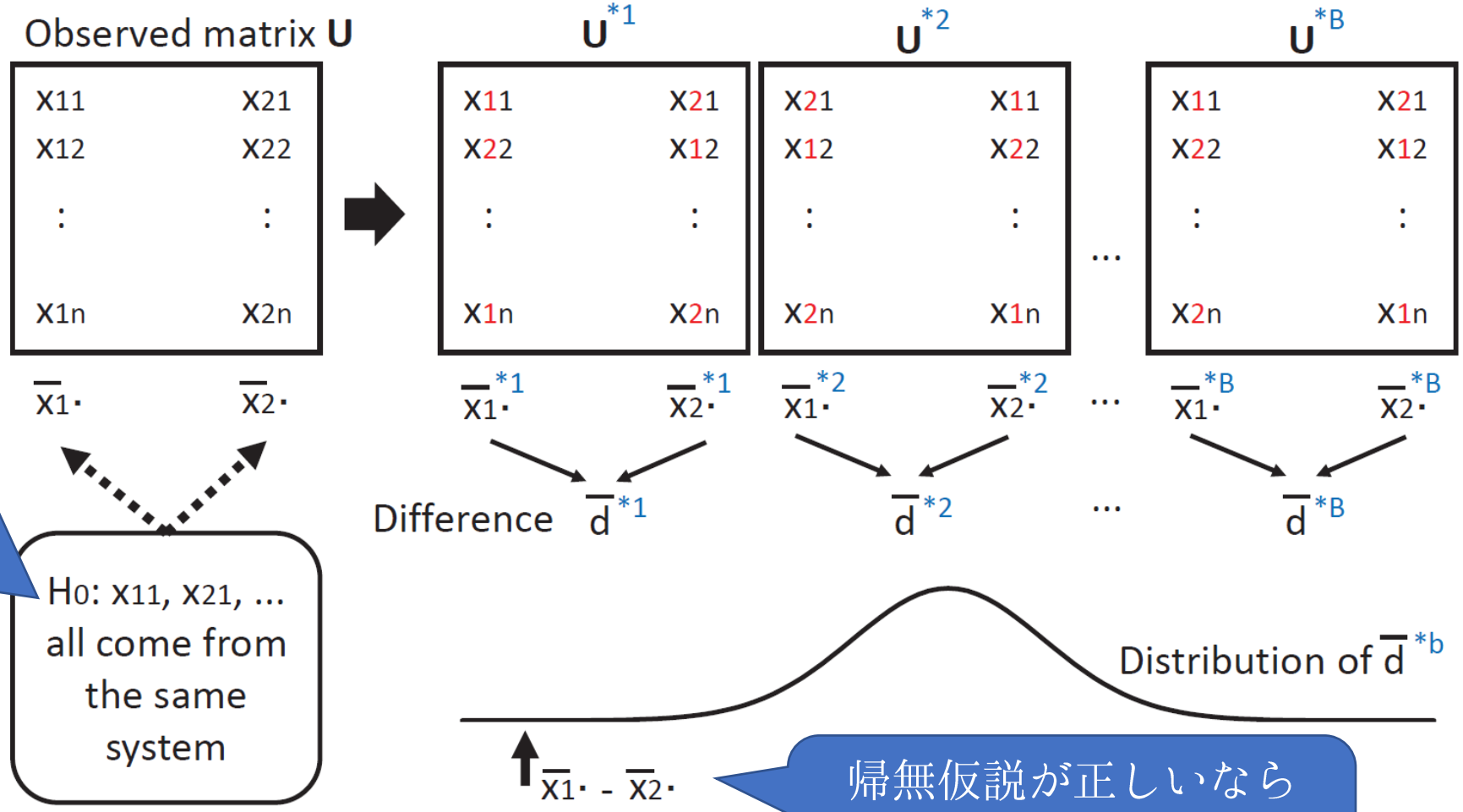
Discpower by Tetsuya Sakai

Discpowerはrandomised Tukey HSD testを用いて評価指標の判別能力を計算するツールキットです。判別能力は与えられた有意水準αに対して計算することもできますが、p-valueをシステム対に対してプロットするAchieved Significance Level (ASL) 曲線のほうが情報量は多いです。Discpowerが出力するcsvファイルから、簡単にMicrosoft Excelなどによ りASL曲線を描くことができます。
詳細についてはtarファイルに含まれるREADMEおよび下記の論文を参照してください。

Download

2つのシステムの場合の原理
(対応のあるデータのt検定に相当)

3つ以上の場合は
対応のあるデータの
Tukey HSD検定に相当



帰無仮説:「各トピックについて2つの値を生成し、コインを投げていずれのシステムに割り当てるかを決めている(システム間の差に意味なし)」

帰無仮説が正しいなら
平均の差は
こんな分布に従うはず

詳細: http://sakailab.com/leirbook/ (Chapter 4)

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

統計的検定結果の適切な報告の仕方

論文を読んだ人がその結果を検証するに足る情報を提供:
サンプルの収集方法、検定の種類、サンプルサイズ、自由度、検定統計量、p-value、効果量

×「提案手法はベースラインを統計的に有意に上回った」

→どんな検定をやったのかさえわからない

×「提案手法1，2，ベースラインを比較したところその差は統計的に有意であった」

→多重比較法を用いたのか? 何と何の間に有意差があったのか?

×「提案手法1はベースラインを有意水準α=0.05にて統計的に有意に上回り、提案手法2はα=0.01にて上回った」

→αは事前に決めるもの。またp-valueその他の情報がない(二元論)

第一種の誤りをどこまで許容するか

(対応のないデータの)t検定結果の報告例

表の目的を明記

Table 5.2 Comparison of retrieval effectiveness

各サンプルサイズを明記。本文中に書いてあっても
のちの研究者のために表にも明記推奨

System	#topics	Mean nDCG
Proposed (System 1)	$n_1 = 20$	0.4500
Baseline (System 2)	$n_2 = 20$	0.4277

数字がなんなのか明記。
平均なら平均と明記。
「nDCG」は不正確

"We conducted a Student's t-test for the difference between our proposed system and the baseline in terms of mean nDCG. The difference is not statistically significant ($t(38) = 0.639$, $p=0.530$, 95%CI[-0.049, 0.094]), with Hedge's $g = 0.200$."

t値と自由度

信頼区間: http://sakailab.com/leirbook/ (Chapter 2)

(ランダム化) Tukey HSD検定結果の報告例

このスライドでは
平均nDCGの表を省略

Table 5.3 Comparison of systems 1–3 in terms of mean nDCG over $n = 20$ topics. In each cell, the difference in mean nDCG, ES_{E2} (effect size) and the p-value (randomised Tukey HSD test with $B = 5{,}000$ trials), are shown

	System 2		System 3	
System 1	0.0224	($ES_{E2} = 0.5182$)	0.0838	($ES_{E2} = 1.9386$)
		($p = 0.4996$)		($p \approx 0.0000$)
System 2	–		0.0615	($ES_{E2} = 1.4227$)
				($p = 0.0024$)

各平均値の差を
標準偏差の推定値
で割った効果量
(ただし等分散性を仮定)

$$ES_{E2}(i,i') = \frac{\bar{x}_{i\bullet} - \bar{x}_{i'\bullet}}{\sqrt{V_{E2}}}$$

"We conducted a paired randomised Tukey HSD test with B=5,000 trials to compare every system pair [⋯] System 3 statistically significantly underperforms Systems 1 (p≒0.0000) and 2 (p=0.0024). Moreover, Systems 1 and 3 are almost two standard deviations apart from each other (Sakai 2018). "

詳細: http://sakailab.com/leirbook/ (Chapter 5)

p値による結論はp.46 と一致

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

α、β、効果量、サンプルサイズ

	H0 採択	H0 棄却
H0 は真 (母平均は等しい)	正しい結論 (確率 1-α)	第一種の誤り α (あわてもの)
H0 は偽 (母平均は異なる)	第二種の誤り β (ぼんやりもの)	正しい結論 (確率 1-β)

検出力: 本物の差を見出せる確率

有意水準α (例: 5%)、確実に検出したい効果 (例: 標準偏差1/5分の差は実用上意味があると見なす)を定めたもとで、
サンプルサイズを大きくする → 検出力が上がる (見逃しが減る)

検出力過小な実験: サンプルサイズが小さすぎて本当の差を見逃しまくる
検出力過大な実験: サンプルサイズが大きすぎて実用的有意差なしでも統計的に有意に

今やったt検定の検出力は? 今後のサンプルサイズは?

https://waseda.box.com/SIGIR2016PACK

豊田秀樹: 検定力分析入門, 東京図書, 2009

上記パッケージは豊田先生の本のRコードをハックしたもの。
例: 対応のあるデータのt検定の結果について:

```
> future.sample.pairedt( 0.953, 28 )
INPUT: t= 0.953 n= 28 Alt= two.sided ALPHA= 0.05 POWER= 0.8

EShat= 0.1801001

Achieved_power= 0.1510342

Sample size n per group = 244
```

t値とサンプルサイズを入力

効果量: 標準偏差いくつぶん? (p.27)

デフォルトは有意水準5%
検出力80%
(Cohenの慣例)

「今回の検出力は15%
ぜんぜんだめ」

「これくらいの効果量について検出力80%を保証し
たいならサンプルサイズはこれくらいないとね」

多重比較における「検出力」はややこしいので (永田靖: 多重比較法の実際, 応用統計学 27(2), 1998)
最も関心のあるシステム対に注目してt検定の検出力を議論するのが個人的なおすすめ

Tutorial outline

(I) 復習: t検定 [15分]
- ギネスビールの社員の話
- 対応のあるデータのt検定
- 対応のないデータのt検定

(II) 「統計的に有意である」と言うだけでいいか？ [15分]
- 統計的検定の弊害
- 効果量
- 統計的検定に代わるアプローチ

(III) 三つ以上のシステムの比較においてt検定を繰り返していいか？ [15分]
- ワインを飲みすぎた客の話
- 全体としての第一種の誤り
- Bonferroni補正はちょっと古い

(IV) Tukey HSD検定 [15分]
- 対応のあるデータのTukey HSD検定
- 対応のないデータのTukey HSD検定
- ランダム化検定とランダム化Tukey HSD検定

(V) 実験結果の適切な報告の仕方 [5分]
- 2つのシステムの比較
- 3つ以上のシステムの比較

(VI) さらにレベルアップ [10分]
- 有意水準、検出力、効果量、サンプルサイズの関係
- 今やったばかりのt検定の検出力は? 今後のサンプルサイズは?

(VII) Q&A・バッファ [15分]

まとめ

- 統計的検定（(今回は平均値の差の検定)）の前提と意味を正しく理解し、正しく報告しよう。P-valueだけでなく効果量を議論!
- 適切な多重比較法を使おう。
- 検出力とサンプルサイズを気にしよう。
- 統計的検定にはたくさん批判もあることを知ろう。

ただしこれは論文で検定も何もやらない言い訳にはならない。

- 永田靖先生、豊田秀樹先生の本などで勉強しよう！

英語でよければ　http://sakailab.com/leirbook/　も！

（是非引用してください）