再現性と検定

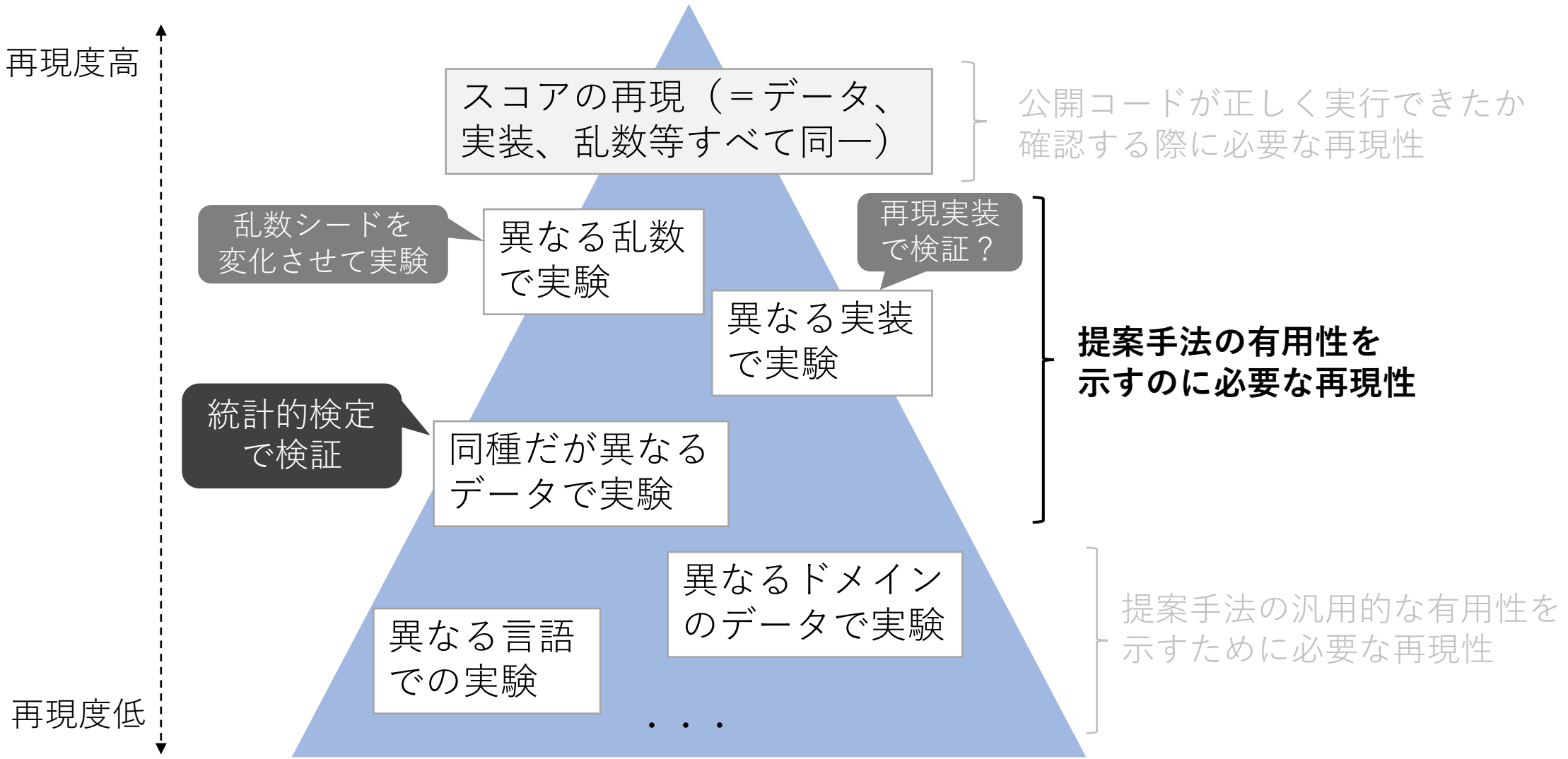
– なぜ検定を行うか –

名古屋大学大学院情報学研究科
笹野遼平

もくじ

- なぜ検定を行うか(5分)
- NLPのための検定入門(10分)
- 検定を実施する上での注意点(10分)

様々なレベルでの再現実験

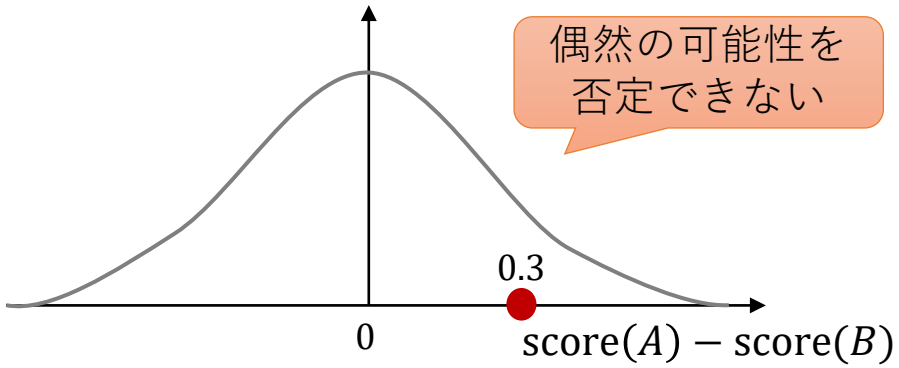


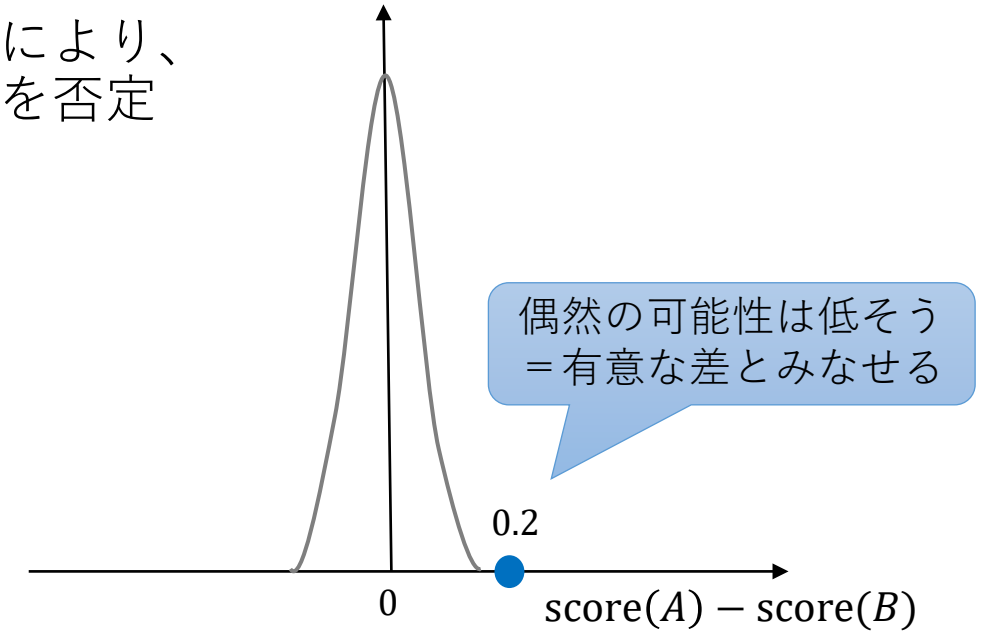
検定の目的

- 母集団に関する仮説を検証
  - 実験で得られた仮説が標本の選び方による偶然の結果でないか検証

e.g., 手法Aと手法Bを比較する場合

- 実験サンプルに対するスコアの分散により、スコア差が大きくても偶然の可能性を否定できない場合が存在



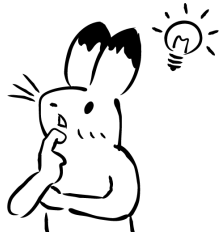


検定を行うモチベーション

- 論文を通すため
  - 実験結果の信頼性は査読者に良い印象を与えるが…
  - 良い研究の結果であって目的ではないはず

- 論文の読者のため
  - 読者が実験結果を適切に解釈するのに有用な情報

- 自分のため
  - **研究の目的は役立つ技術の開発（工学）、真理の追究（理学）**
  - 正しくない仮説に時間を費やすのはもったいない
    ⇔ 有意かどうかの直感を身につけることは非常に有益



NLPのための検定入門

統計的仮説検定の基本的な流れ

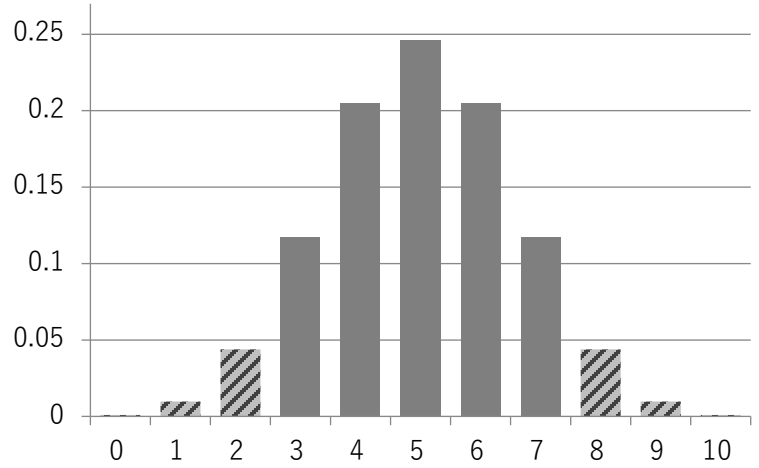
1. 示したい仮説（対立仮説: Alternative hypothesis）に対し、その逆の帰無仮説（null hypothesis）を考える

2. 『**帰無仮説が正しいとした場合に、観測された結果と同等、またはそれより極端な結果が起こる確率**』であるp値を算出

3. p値が事前に定めた有意水準α(本発表では0.05とする)以下の場合、帰無仮説を棄却し、有意水準αで有意であると判定

例1. コインを10回投げた結果、表が出た回数が2回であったとき

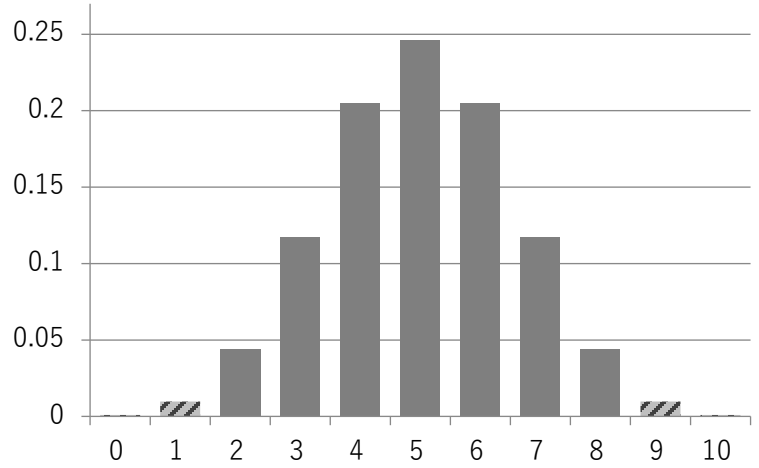
- 対立仮説:
  - 表と裏が出る確率は異なる
- 帰無仮説:
  - 表と裏が出る確率が同じである



- p値（観測値と同等 or より極端な結果となる確率）
  = 表裏いずれかの出る回数が2回以下となる確率:
  $= 2 \times \sum_{k=0,1,2} \frac{{}_{10}C_k}{2^{10}} \approx 0.1094 > 0.05$ ⇒ 有意差なし

例2. コインを10回投げた結果、表が出た回数が1回であったとき

- 対立仮説:
  - 表と裏が出る確率は異なる
- 帰無仮説:
  - 表と裏が出る確率が同じである



- p値（観測値と同等 or より極端な結果となる確率）
  = 表裏いずれかの出る回数が1回以下となる確率:
  $= 2 \times \sum_{k=0,1} \frac{{}_{10}C_k}{2^{10}} \approx 0.0215 < 0.05$ ⇒ 有意差あり

二項検定（符号検定）

仮説検定の結果と対立仮説の真理値

	帰無仮説が棄却される	帰無仮説が棄却されない
対立仮説が真	正しい判定（1 −β）	第2種の過誤（β）
対立仮説が偽	第1種の過誤（α）	正しい判定（1 − α）

- 第1種の過誤 (Type I error):
  - 対立仮説が実際には偽であるのに帰無仮説が棄却されてしまう誤り
  - 有意水準 α と一致
- 第2種の過誤 (Type II error):
  - 対立仮説が実際には真であるのに帰無仮説が棄却されない誤り
- 確率 α と確率βはトレードオフの関係

NLPで利用される検定

- 実験結果が特定の分布に従うことが仮定できない場合が多い

  ⇒ ノンパラメトリック検定が一般的

  - 特定の母集団分布を仮定しないので原理は比較的簡単

    （何らかの分布を仮定してパラメトリック検定を使う場合もある）

- NLPで利用されるノンパラメトリック検定は大きく２種類
  - 並べ替え検定 (paired permutation test)
  - ブートストラップ法 (paired bootstrap)

マクネマー検定やウィルコクソンの符号順位和検定はこれの特別な場合

マクネマー検定

- システムAとBが共通のN個の問題を解くとする（正解/不正解で評価）
- 表に示す結果が得られた場合に２システムの優劣を検定
  - 対応のある2群の検定
  - **優劣に差がないならば、いずれか一方のみ正解となった24事例に対し、各0.5の確率で一方のシステムが正解となるはず**

⇒ 事例数が18:6の符号検定

$$p = 2 \times \sum_{k=0}^{6} \frac{{}_{24}C_k}{2^{24}} \approx 0.0227 < 0.05$$

⇒ 有意な差

	B正解	B不正解	計
A正解	54	**18**	72
A不正解	**6**	22	28
計	60	40	100

※ 同じスコア差でも出力が類似している場合の方が有意となりやすい

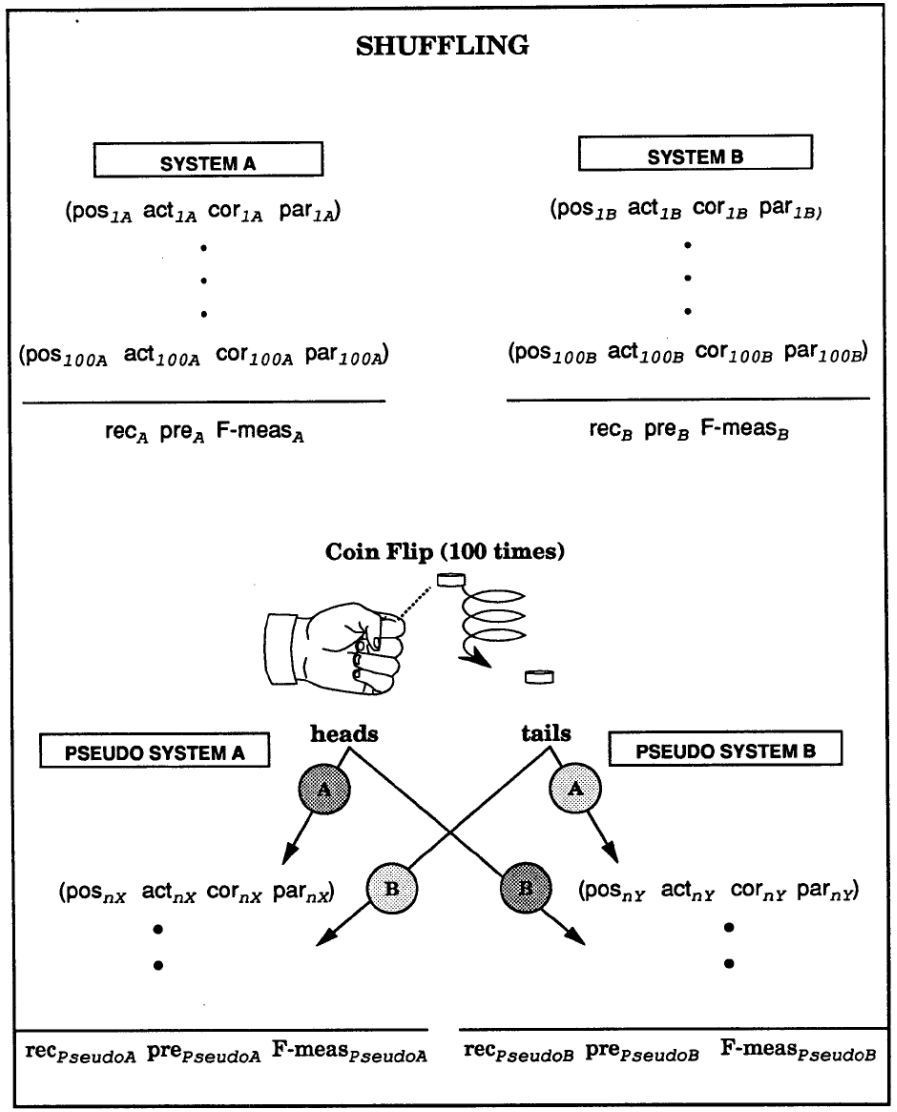
並べ替え検定 (paired permutation test)

[Chinchor'92]

- ２システムに差がない場合、各インスタンスごとに、出力を入れ替えた結果も同じ確率で発生
  ⇒ 2^N通りの並べ替えに対するスコアからp値を算出
- 近似的な算出法が一般的（**ランダム化検定**, randomization test）

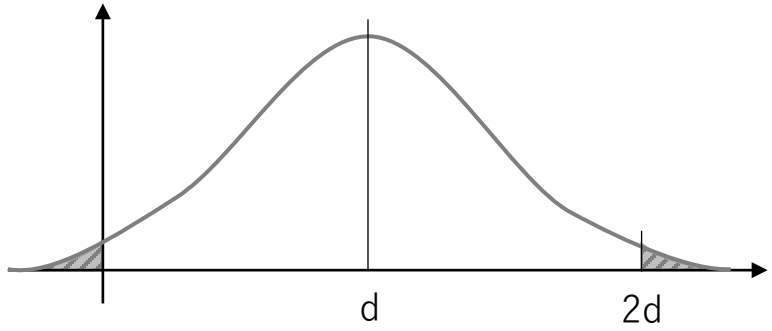
1. 実際に観測されたデータの評価スコアの平均の差の絶対値|d|を算出・カウンタCを1に初期化
2. 以下の処理をM − 1 回繰り返す
   1) 各入力インスタンスごとに1/2の割合で出力を入れ替えたデータを生成
   2) 生成されたデータに対し、２つのシステムのスコアの平均の差の絶対値$|d_m|$ を算出
   3) $|d_m| \geqq |d|$ の場合、カウンタCを1加算
3. CをMで割った値をp値として出力

- インスタンスを正解/不正解で評価⇒マクネマー検定



ブートストラップ法 (paired bootstrap)

- 観測データ（N個のインスタンスで構成）から、重複を許しN'個のインスタンスをリサンプリングし、擬似データセットを生成（本発表ではN'=N）
- 各セットにおける2システムのスコアを計算
- **『スコアの大小関係が実際に観測された大小関係と一致しない割合（or スコア差d'が実際のスコア差dの2倍以上である割合）』**がp値



- なぜ『』内をp値とみなせるか？
  - 実際に観測されたデータからサンプリングを行うのでシステムの差の平均はd
  - 擬似データにおけるスコア差d'が0 以下 or 2d以上の場合、平均dからd以上離れていることから「観測された結果と同等、または、それより極端な結果」とみなせる

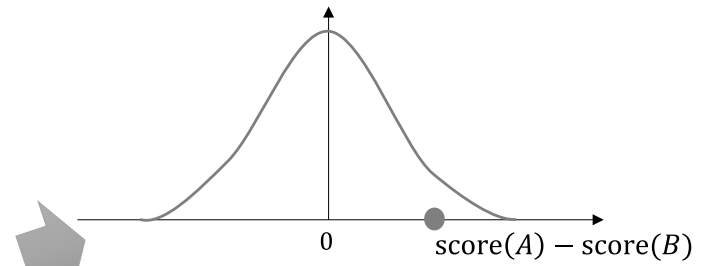
ランダム化検定とブートストラップ法

インスタンス番号 i	A のスコア（s_{a_i}）	B のスコア（s_{b_i}）			
1	0.527 9	0.600 1	×	1	2
2	0.889 1	0.586 0	×	-1	
3	0.732 5	0.699 6	×	1	1
4	0.652 7	0.604 7	×	1	1
5	0.623 8	0.597 9	×	1	
6	0.792 5	0.620 0	×	-1	
7	0.615 2	0.622 4	×	1	3
8	0.583 7	0.600 0	×	-1	2
9	0.499 5	0.570 5	×	1	1
10	0.839 6	0.662 0	×	-1	
11	0.705 5	0.601 0	×	-1	
12	0.641 4	0.592 5	×	-1	1
13	0.621 5	0.732 3	×	1	
14	0.619 8	0.609 8	×	1	4
15	0.610 6	0.599 4	×	1	
16	0.567 0	0.546 7	×	-1	1
17	0.999 1	0.635 2	×	-1	
18	0.796 5	0.600 0	×	1	
19	0.683 1	0.581 8	×	-1	3
20	0.633 0	0.686 1	×	1	1
平　均	0.681 7	0.617 4			

- どちらも観測データを基にサンプリングし分布を推定
- ２つの違いはサンプリング法
- 基本的に得られるp値は類似

ランダム化検定
- 全事例を使用
- 出力を入れ替え

ブートストラップ法
- 事例を重複を認めリサンプリング



検定を実施する上での注意点

アメリカ統計学会(ASA)によるp値に関する声明(2016)

- p値の望ましくない使われ方が目立ったことからASAにより出された声明

1. p値はそのデータが特定の統計モデルと適合しない度合いを示し得る
2. p値は調査対象とした仮説が正しい確率やそのデータが偶然のみにより生成された確率を計測するものではない
3. 科学的な結論やビジネス・政策上の意思決定はp値がある特定の閾値を越えたかどうかのみに拠るべきではない
4. 適切な結論の導出には行ったことの完全な報告と透明性が必要である
5. p値や統計的有意性は効果の大きさや結果の重要性を測るものではない
6. p値はそれだけではモデルや仮説の正しさに関するエビデンスのよい指標とはならない

検定を実施する上での注意点

1. p値の小ささと2群の差の大きさを区別する
2. p値がどのくらいの大きさであったかも意識する
3. 検定統計量/帰無仮説が何であるか意識する
4. どのような独立性を仮定したか意識する
5. 有意差のある組合せを取り出していないか注意する

1. p値の小ささと2 群の差の大きさ

- 真の差が大きいほど小さなp値が得られやすい
- しかしp値が小さいことは2手法の差が大きいことを意味しない
  - 一般に出力が類似している場合の方がp値は大きくなりやすい

e.g., 2つの改善手法A, Bがあるとする

A）改善度合いは小さいがベースライン手法と比較した場合の有意差検定におけるp値は非常に小さい

B）改善度合いは大きいがp値は小さくない

改善が偶然でなかった可能性は高いのはAであるが
改善度合いの期待値が大きいのはB

2. p値の大きさについて

- 検定では事前に設定した有意水準α（=0.05）よりp値が小さいかどうかで有意性を判定
  - しかしp値が0.051と0.049の場合で本質的な差があるわけではない
  - 一方、p値が0.049と0.001の場合では大きな差がある

- p値は報告してほしい
- **実験結果をどう捉えるか読者が判断可能に**

3. 検定統計量および帰無仮説が何であるか

- 全体のスコアの平均は $a < b$ （⇒合計スコアに着目するならbの方が優秀）
- しかし、マクネマー（っぽい）検定を行うと…
  - インスタンスレベルでは、$s_a > s_b$ が20事例、逆が5事例
  - $2 \times \sum_{k=0}^{5} \frac{{}_{25}C_k}{2^{25}} \approx 0.004 < 0.05$ ⇒ aが有意に良いことになる
  - 大小の検定ならこれでOK（=aがbより高いスコアとなる確率は1/2より有意に大きい）

$r(\lvert s_a - s_b \rvert)$	$s_a - s_b$
1	+0.003
2	+0.009
3	+0.015
4	+0.016
5	+0.020
6	+0.021
7	−0.022
8	+0.023
9	−0.026

$r(\lvert s_a - s_b \rvert)$	$s_a - s_b$
10	+0.027
11	+0.028
12	+0.029
13	+0.030
14	+0.031
15	+0.033
16	+0.034
17	+0.040
18	+0.044

$r(\lvert s_a - s_b \rvert)$	$s_a - s_b$
19	+0.049
20	+0.061
21	+0.070
22	+0.080
23	−0.203
24	−0.217
25	−0.234
total	−0.039

4. 仮定した独立性について

- どのような独立性を仮定して検定を行ったかは重要

- どのような仮定をおいたかは報告してほしい
- 少なくとも何を１単位としたか報告してほしい

- 簡単なように見えて意外と難しい話

e.g., 固有表現認識
- 文単位で良いのか ⇒ 多くの場合、良くない
- 記事単位で良いのか ⇒ 記事のトピックに偏りがある場合は良くない
- ドメイン単位が良いのか ⇒ 事例が少なくなり有意差を確認できない

5. 有意差のある組合せを取り出していないか

- 素性の組合せを試した結果、ベースラインに対して精度が向上した上でp値が有意水準以下となる組合せが見つかった

Q これらの素性の組合せはベースライン手法に対して有意な精度向上を達成しているか？

e.g., 10個の新たな素性の組み合わせを試す

A 実際には偶然にすぎない場合が多い

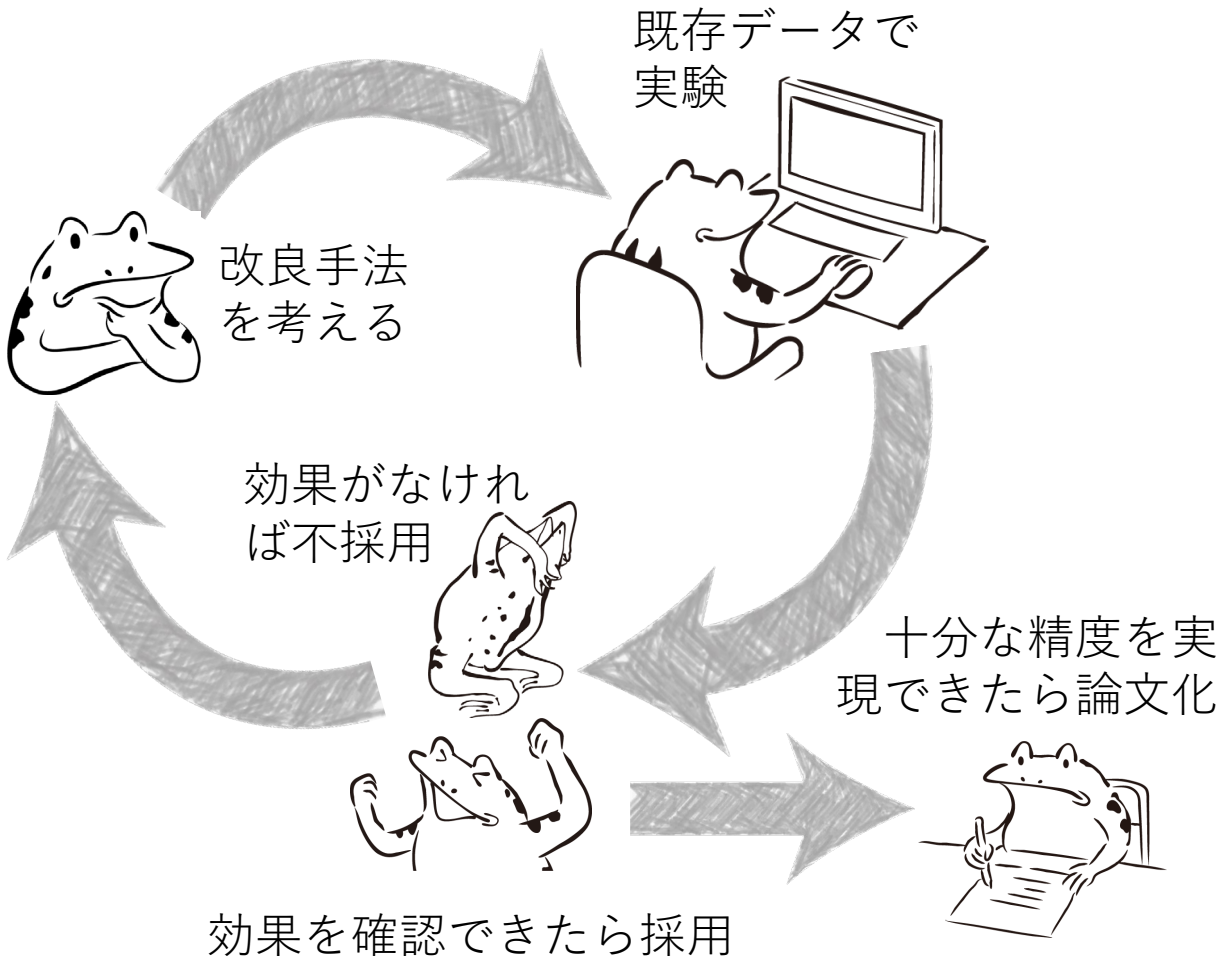
- 10個の素性を試したということは$2^{10} = 1024$通り
- 確率的に50通りくらいは有意であると判定される

１つの解決法：多重検定 (multiple testing)

- 課題に対し検定を繰り返し実施する設定
- 第1種の過誤の発生確率は以下のようになる
  - $1-(1-\alpha)^N$ （e.g., α=0.05, $N = 10 \Rightarrow 0.401$）
- なんらかの補正が必要
  - ボンフェローニ法: 事前に設定した有意水準を検定を行う回数Nで割った値を有意水準として使用
  - 改良手法であるホルム法なども存在

実際には簡単な話ではない…

- 一般的(?)な研究の進め方
  - 開発データは存在するが…
  - 既存研究を上回るために通常、テストデータで複数回実験
- 同一のサンプルで実験することが通常あり得ない“サイエンス”の実験と異なる点
- 絶対的な解決法はないもののp値が十分に小さいかを確認することである程度判断できる可能性
  - 自分で有意な差と思えるかどうか



その他:
複数の乱数シードで実験した場合の検定法

- M個の乱数シードで実験し分布を分析すれば良い？

- シード選択に対する偶然性は判断できる
- **データ選択に対する偶然性は判断できない**

- 見たことのある対処法/考えられる対処法
  - インスタンスごとの全シードの平均値を算出し検定[Mithun+'21]
    - 実験設定・評価尺度によっては適用が難しそう
  - シードごとに検定を実施し有意と判定された割合で判断[Huang+'20]
  - サンプリング時に全乱数シードの結果から無作為に抽出？

まとめ: なぜ検定を行うか

- 直接の目的:
  - 実験で得られた仮説が標本の選び方による偶然の結果でないか検証
- 検定を行うモチベーション
  - 自分のため＝**技術の開発（工学）、真理の追究（理学）**を達成するために正しくない仮説に時間を費やすことを防ぐ

  ⇒ 個人的にはあまり学会等でガチガチにルールを決めてほしくない
- 正しくない仮説であることを認識するのに必要なこと
  - 正しく検定を行うための知識・技術
  - 有意差に対する直観（このくらいの標本サイズだと…）

参考文献

- 検定手法に関する論文・書籍
  - Nancy Chinchor, The Statistical Significance of the MUC-4 Results, MUC4, 1992
  - Philipp Koehn: Statistical Significance Tests for Machine Translation Evaluation, EMNLP, 2004
  - Taylor Berg-Kirkpatrick, David Burkett, Dan Klein: An Empirical Investigation of Statistical significance in NLP, EMNLP-CONLL 2012
  - 笹野遼平, 飯田龍: 文脈解析 -述語項構造・照応・談話構造の解析-, コロナ社, 2017
  - Rotem Dror, Gili Baumer, Segev Shlomov, Roi Reichart: The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing, ACL, 2018
- 複数の乱数シードを用いた実験の検定法について参考にした論文
  - Mitch Mithun, Sandeep Suntwal, Mihai Surdeanu: Students Who Study Together Learn Better: On the Importance of Collective Knowledge Distillation for Domain Transfer in Fact Verification, EMLNP, 2021
  - Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, Xiaodan Liang: GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems, EMNLP, 2020