統計の基礎の基礎

統計

- 記述統計
  - 合計、平均、標準偏差…
- 推測統計

推測統計

e.g. 日本語全体
(母集団)

コーパス =
標本 (サンプル)

前提

- 無作為抽出（ランダムサンプリング）
  - 実際には完全な無作為抽出は困難であり、データがそれほど偏っていないという想定のもと研究が行われることが多い

推測統計

「表さない」 2
「表しない」 5

母集団の差を反映？
それとも、この
サンプルの偶然？

検定

- **帰無仮説** (null hypothesis) ：母集団に偏りはない。サンプルに数字の違いが出たとしても偶然である
- **対立仮説** (alternative hypothesis)：母集団に本当に偏りがあり、サンプルの数字はそれを反映している

検定

- 帰無仮説が正しい（母集団に差がない）として、偶然サンプルで差がつく確率を計算する。これを**p値**という。
- p値が十分小さい場合に、差が**有意** (significant) であるといい、帰無仮説が棄却され、対立仮説が支持される
- 通常、p値が5%以下のとき有意とする。5%という数字に特に意味はない。

検定

- 注意
  - p値は帰無仮説が正しい確率ではない。あくまで、帰無仮説が正しいとしたときのデータの確率。
  - この検定の考え方のもとでは、帰無仮説が正しいことを示すことはできない。

検定の例

- コイン投げ

クロス表の検定

	否定環境	その他	合計
「が」	690	5972	6662
「の」	185	753	938
合計	875	6725	7600

南部 (2007) より

クロス表の検定

- 何がしたいか？
  - 「が」を使うか「の」を使うかと、否定環境かどうかとのあいだに関連があるかどうかを見たい
  - 関連がないとした場合の期待値を求める
  - 期待値と実測値との違いを合計し (**カイ二乗統計量**)、そのようなズレが生じる確率を求める (**p値**)

Excel の場合

- CHITEST - p値を得る
- CHIINV - χ^2 （カイ二乗値）を得る
- 自由度

$$\chi^2 = \sum \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}}$$

- 論文では、**χ^2、自由度 df、p値 の3つを報告する**

注意

- カイ二乗の計算式で使うのは必ずもとの頻度です。比率（パーセントとか、100万語あたりなど）のデータを使ってはいけません。
- データが少なすぎる場合（数値が5以下のセルがある場合などと言われます）は正確さが落ちます。フィッシャーの正確確率検定 (Fisher's Exact Test) などを使うほうが良いと言われます。

注意2

- 有意検定は万能ではありません
  - とくにコーパス研究の場合、「有意差がたまたま出るまでデータを恣意的に追加する」というような操作が簡単なため、有意差だけに頼るのは危険と言われています
- 効果量 (effect size) を報告するなど、別の手法を組み合わせることが望ましいと言われています

他にもさまざまな統計手法が使われます

- 例えば
  - 回帰（とくにロジスティック回帰）

統計ソフト

SPSS

- 見た目はExcelのような表計算ソフト風
- 心理学で利用が盛ん
- 高価
- Windows/Mac

R

- フリー
- コマンド操作が基本
- 自分で複雑な一連の操作を関数として定義したり、それを世界の研究者と共有したりといった、自由度・拡張性が高い
- Windows/Mac/Unix

VARBRUL

- アメリカ社会言語学を中心にスタンダード
- 性別/年齢/社会階層などを変数に、ロジスティック回帰を行うのに向いている

コメントシート：架空のデータで練習

	Aさん	Bさん	合計
um	24	36	60
uh	6	24	30
合計	30	60	90

$$\chi^2 = \sum \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}}$$

p値	カイ二乗値 (df = 1)
0.5	0.455
0.1	2.710
0.05	3.84
0.025	5.02
0.01	6.63
0.005	7.88

有意