統計学(基礎)

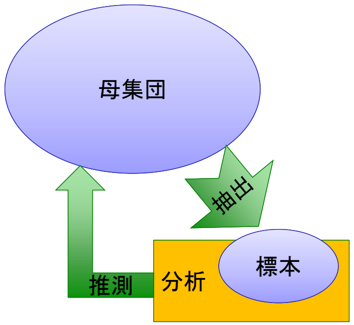
第7回
母集団の等分散性の判断と
2群の平均値の差の検定

2群の平均値の差の検定

t検定

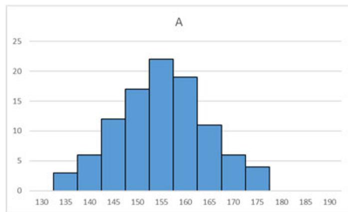
推測統計の基本的な考え(再)

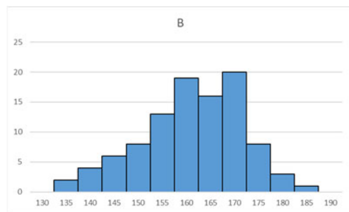
- 母集団と標本抽出
  - 全体(母集団)から偏りなく得られた(抽出された)データ(標本)を使えば、全体を測定しなくても全体をある程度の精度で推測できる
  - 手元にあるデータが抽出された標本であると仮定できれば、全体を推測できるとする
  - そもそも、対象となるデータは大きすぎて現実的にデータがとれない



疑問

- A　平均値 153.0　　標準偏差 9.1
- B　平均値 159.4　　標準偏差 10.4



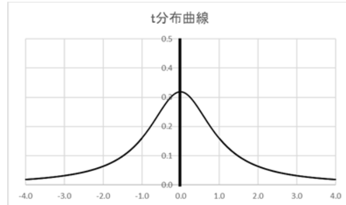


疑問

- グループ間で平均値に違いがあるのか
  →AとBの間で違いがあるか
- t検定(平均値の差の検定)
  - 群によって平均値が異なっているかの違いを見る
  - データの分布が異なっているか
  - 同じ母集団からの標本と言えるか言えないか

t検定

- 2群の平均値の差の検定
  - 2つのグループ間で平均値に差があるかどうかを調べる
  - t値を計算して、それが帰無仮説が棄却できるかできないかを調べる
  - 2群間に差がなければt値は0
  - t分布は0を中心に左右対称



t検定は2種類あるの

t検定の進め方

t検定は2種類ある

- 普通のt検定　スチューデント(Student)のt検定
  - 2群が等分散である
- ウエルチ(Welch)の検定
  - 等分散であることを仮定しない

2つのt検定(JASP)

2つのt検定(jamovi)

スチューデントのt検定

- t値の計算方法

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

nはデータ数 sは標準偏差(s^2は分散) $\bar{x}$は平均値

- 自由度は (一方のデータ数-1)+(もう一方のデータ数-1) なので、n_1+n_2 −2

ウエルチのt検定

- t値の計算方法

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)}}$$

nはデータ数 sは標準偏差(s^2は分散) $\bar{x}$は平均値

- 自由度は…

ウエルチのt検定の自由度(d.f)

$$d.f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}\right)}$$

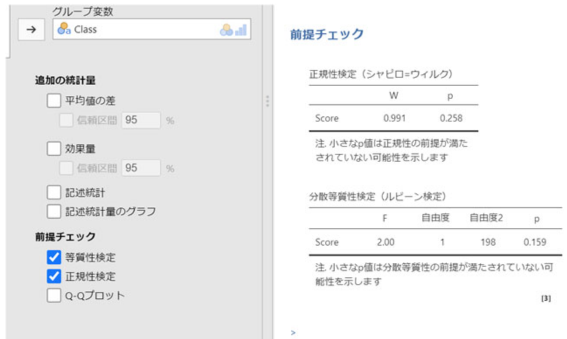
- nはデータ数　sは標準偏差(s^2は分散)

t検定と等分散性

- 2群間の分散が等しければ、スチューデントのt検定
- 分散が等しくなければ、ウエルチの検定
- t検定の前に、分散が等しいかどうかの検定をする
  - ってことになっていたんですが

仮定のチェック(JASP)

前提チェック(jamovi)



t検定前の検定

- 正規性の検定
  - シャピローウイルク(Shapiro-Wilk)検定
- 等分散性の検定
  - リーベン(Levene)の検定
  - ブラウン・フォーサイス(Brown-Forsythe)検定
- これやるの？

正規性や等分散性の判断

- 等分散性の検定をして、帰無仮説が棄却されたら分散が等しくない、棄却されないなら等分散として検定を選択する。
- データが正規分布であるかどうかを判断する。

- 同じデータで検定を繰り返すと、間違う確率が高くなるので、最初から等分散でないと仮定して検定を実施する
  - 第1種の過誤

第1種の過誤とは

- 例えば同じデータで2回検定をする
  - 最初の有意水準を5%(0.05)とすると、帰無仮説が棄却された場合、それが正しい確率は95%(0.95)
  - 2回目も棄却された場合、それが正しい確率は$0.95^2 = 0.903$と約90%まで下がる
  - 3回目は$0.95^3 = 0.857$となって、約86%

正規性や等分散性の判断

- 正規性
  - t検定はロバストだから、まあだいたいで大丈夫
- 等分散性
  - 最初からウエルチの検定でやればいい

スチューデントとウエルチ

- スチューデントのt検定は元々簡易版
  - ちゃんとやると計算が大変だから、等分散ってことで
  - だって、ウエルチの検定とか自由度の計算大変なんだもん

$$d.f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}\right)}$$

スチューデントとウエルチ

- コンピュータで分析を行うようになって、ウエルチも使われるようになる
- だったら全部それでやればいいじゃん
  - 等分散性の検定をする意味なし
  - 過去との比較でどうしてもスチューデントを使いたいときだけ、等分散性の検定
    - リーベン 外れ値やnが小さいと弱い(平均値の偏差)
    - ブラウン･フォーサイス リーベンよりはロバスト(中央値の偏差)

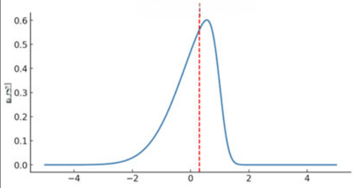
正規性の検定について(前回参照)

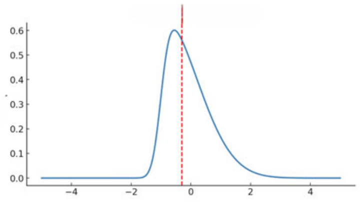
- ヒストグラムで十分
  - シャピローウイルクはnが小さいと棄却力が低く、nが大きいと何でも棄却されるので使いにくい
- そもそもt検定はロバストだから、少々のゆがみは大丈夫
- ダメな場合はノンパラ(マンホイットニーのU検定)

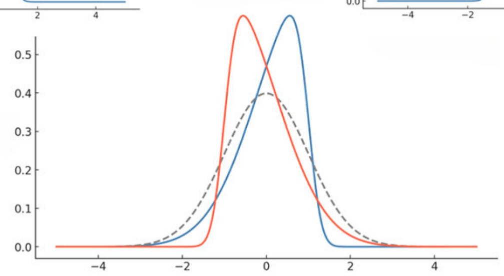
今の動向

- ヒストグラムを作る
- ある程度正規性が見られるなら、ウエルチのt検定
- ゆがんでいたりする(裾が長い)場合は、マンホイットニーのU検定
- 正規性の検定や等分散性の検定はしない

歪み







このデータの本来の手順

- 度数分布表を作る→ヒストグラムを作る
- 判断をする→t検定ができるならウエルチで
  – data07_01

	Class	Score
1	A	151
2	A	155
3	A	157
4	A	143
5	A	156
6	A	132
7	A	174
8	A	172
9	A	150
10	A	172

分析の選択

分析の指定と結果(JASP)

分析の指定と結果(jamovi)

t検定の結果の書き方

t=4.626　d.f= 194.7　p<0.01

t値は絶対値(正の値)で書くのが一般的

