測定値から母集団の推定の例

10人の男子学生の身長を測った結果

学生	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
身長 cm	175	170	169	165	164	179	172	168	175	170

仮定：　母集団の分散 σ^2 が分かっていない

- 母平均 X の点推定値＝標本平均 $(X_1+\ldots+X_{10})/10$
- 母平均 X の区間推定値は
  - 信頼係数 $1-\alpha=0.95$（$\alpha=5\%$）に対する信頼区間は
    $[X-t_{\alpha/2}(n-1)\cdot s/\sqrt{n},\ X+t_{\alpha/2}(n-1)\cdot s/\sqrt{n}]$
    但し $t_{\alpha/2}(n-1)$ は自由度 n-1 の t 分布で外側の確率が $\alpha/2$ の点
    $t_{\alpha/2}(n-1)$ の値は、表を見るか、プログラムで計算

データの準備　CSVファイルの読込み

CSVファイル＝カンマ区切りのデータ（のファイル）

168	57
173	62
177	63
169	58

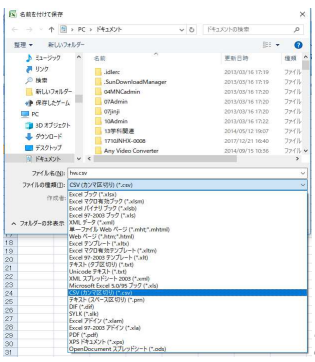
168,57（改行）
173,62（改行）
177,63（改行）
169,58（改行）
Enterキー

- 列（欄）はカンマで区切られている
- 行は行で区切られている
- 基本的にデータのみ書かれている
  （ヘッダ行を入れてもいい）
- Excelなどとの交換が可能

データを作ってみよう

- Excelから作る　　Excelを起動 → 表を作成
  →「ファイル名を指定して保存」で
  D:¥WPy-3662¥notebooksフォルダの中に
  「ファイルの種類」⇒CSV(カンマ区切り)
  を選んで、ファイル名「hw.csv」を書いて
  「保存」する
- メモ帳で作る　　メモ帳を起動 → データ入力
  →D:¥WPy-3662¥notebooksフォルダ中に保存
  →ファイル名をhw.csvに変更する

	A	B
1	168	57
2	173	62
3	177	63
4	169	58



データの準備　CSVファイルの読込み

CSVファイル = カンマ区切りのデータ (のファイル)

pandas の read_csv を使って読み込んでみよう

import pandas as pd　　をしておいて

hw = pd.read_csv('hw.csv', header=None) で読む

ファイル名指定　　**csvファイル中にヘッダー行が無いことを指定**

print(hw)　で、読み込んだ結果のhwを表示してみる

ファイル hw.csv

168,57 (改行)
173,62 (改行)
177,63 (改行)
169,58 (改行)

print の結果

	0	1
0	168	57
1	173	62
2	177	63
3	169	58

では「身長の平均」に戻って、データを自分で作ろう　→

175
170
169
165
164
179
172
168
175
170

→ height.csv というファイルに保存しよう

↓

hw = pd.read_csv('hw.csv', header=None) で読んでみよう

- 母平均 の点推定値＝標本平均 $(X_1+\ldots+X_{10})/10$
  - 手で計算できる程度ですが…

    標本平均Xは

    X = (175 + 170 + 169 + 165 + 164 + 179 + 172 + 168 + 175 + 170) / 10
  - Pythonだと？
    - pandas の read_csv で読み込む

      h = pd.read_csv(' ….… ', … … …)

      print(h)　　結果 h を表示してみる
    - pandasの平均値関数で平均を計算してみる

      hm = h[0].mean()　データフレーム h に対して mean()を作用させる

      print(hm)　　結果 hm を表示してみる
    - 標本平均の結果は X = 170.7

- 母平均の区間推定値は
  - 信頼係数 $1-\alpha=0.95$（α=5%）に対する信頼区間は

    $[X - t_{\alpha/2}(n-1)\cdot s/\sqrt{n},\ X + t_{\alpha/2}(n-1)\cdot s/\sqrt{n}]$

    但し $t_{\alpha/2}(n-1)$ は自由度 n-1 の t 分布で外側の確率が $\alpha/2$ の点

    $t_{\alpha/2}(n-1)$ の値は、表を見るか、プログラムで計算
- この通りに計算してみる

$$\text{分散} = \frac{\sum_i (x_i - \overline{X})^2}{n}$$

  - Xは前ページで 170.7
  - s は標本の分散で定義は ⇒ 計算してみると 21.79

    pandasの平均値関数で分散・標準偏差を計算してみると

    hv = h[0].var()　　データフレーム h に対して var()を作用（分散）

    hs = h[0].std()　　データフレーム h に対して std()を作用（標準偏差）

    標本分散 hv は 21.79、　標本標準偏差 hs は 4.668
  - $t_{\alpha/2}(n-1)$ は、

```
import scipy.stats as st
t_critical = st.t.ppf(q=0.975, df=n-1) * h[0].std() / math.sqrt(n)（α=0.05で）
    t_critical = 3.339
```


- では、データが h に読み込んであるとして、母平均の区間推定値

  $[X - t_{\alpha/2}(n-1)\cdot s/\sqrt{n},\ X + t_{\alpha/2}(n-1)\cdot s/\sqrt{n}]$

  を計算してみよう

  X = h[0].mean()　標本平均

  s = h[0].std()　　標本標準偏差

  $t_{\alpha/2}(n-1)$ の値は、

  t_critical = st.t.ppf(q=0.975, df=n-1)　（α=0.05として）

  $\sqrt{n}$ = math.sqrt(n)　　で計算できる　（import mathが必要）
- プログラムを完成させて計算してみよう

  $L = X - t_{\alpha/2}(n-1)\cdot s/\sqrt{n}$

  $U = X + t_{\alpha/2}(n-1)\cdot s/\sqrt{n}$　　推定区間は [L, U]

  あとは自分で…

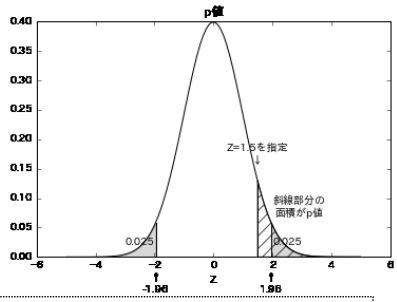
TAチェック

t_crit = 3.3392
[L, U]=[167.3608 , 174.0392]

仮説検定はこうする

仮説： 母集団の平均 μ ＝175、分散 σ^2は未知

ファイルの測定値データは95%の有意水準で
この仮説を棄却するか？



母分散未知 → t 分布による仮説検定
この仮説を棄却するか？

⇒

仮説が正しければ、標本平均Xについて
$t = (X-\mu)/(s/\sqrt{n})$ としたtが

＊有意水準95%の区間推定 [U, L]
から外れれば棄却される

＊t より外側の起こる確率 (面積) を
pとするときp＜2.5%なら棄却される

$t = (X-\mu)/(s/\sqrt{n})$ の計算は
mu = 175　　仮定する
X = h[0].mean();　s = h[0].std()
$\sqrt{n}$ は math.sqrt(n)

あとは自分で…　　print(t)しておくこと

求めた $t = (X-\mu)/(s/\sqrt{n})$ に対する
分布の外側の面積 (p値) は
import numpy as np
p = st.t.cdf(-np.abs(t), n-1)

np.abs(…) は絶対値を計算する
この p は片側だけの面積なので、
両側ならこれを2倍して5%と比較

あとは自分で…　　TAチェック

このデータに対する結果：
母平均=170という仮説
⇒ p=0.3233>0.025　棄却しない
母平均=175という仮説
⇒ p=0.0086<0.025　棄却される

一発で計算するライブラリ関数

scipyの中の scipy.stats.ttest_1samp を使う

マニュアルページは
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html
← 一般にマニュアルページを見ること（種類が多過ぎて網羅的に紹介されていないので）

```
import scipy.stats as st      scipy.stats を用意する

(pandasでcsvを読んで h に入れてあるとする)
res = st.ttest_1samp(h[0], 170)    母平均=170の仮説
print('統計値=', result[0], 'p=', result[1])
```

TAチェック

各自試すこと
結果は
母平均 170 とすると　　統計値= 0.47422068978599446　　p= 0.6466366136116495
母平均 175 とすると　　統計値= -2.9130699515425924　　p= 0.017224970631213674