第3章の内容

1. はじめに・推測統計の基本
2. 統計的推定（点推定・区間推定）
3. 統計的仮説検定
   - 基本的な仮説検定 –
   - 2つの標本問題に関する仮説検定 -

105

1.はじめに・推測統計の基本

推測統計とは

母集団から抽出した標本の情報を用いて母集団の情報を推測すること。つまり，観測対象全体の統計的性質を，その観測対象の一部分のみを使って推測する。

母集団（population）：推測したい観測対象全体のこと
標本（sample）：推測に使う観測対象の一部分
標本抽出（sampling）：母集団から標本を取り出すこと
推定量（estimator）：推定に用いられる統計量
推定値（estimate）：標本データを用いて計算した結果，推定量の実現値

106

2.統計的推定（点推定・区間推定）

基本的な分布

- 正規分布
  - 統計解析において最もよく使われ，重要な推定や検定の理論は全て正規分布を基礎にしていると言っても過言ではない。
  - 自然界の多くの現象を表現できる連続分布（確率分布）。
  - 18世紀にガウス（Gauss）によって誤差の研究から誘導されたものでガウス分布（Gaussian distribution）とも呼ばれる。

107

2.統計的推定（点推定・区間推定）

基本的な分布

- 確率密度関数（probability density function, pdf）
  - 連続型確率変数がある値をとるという事象の確率の密度を表す関数

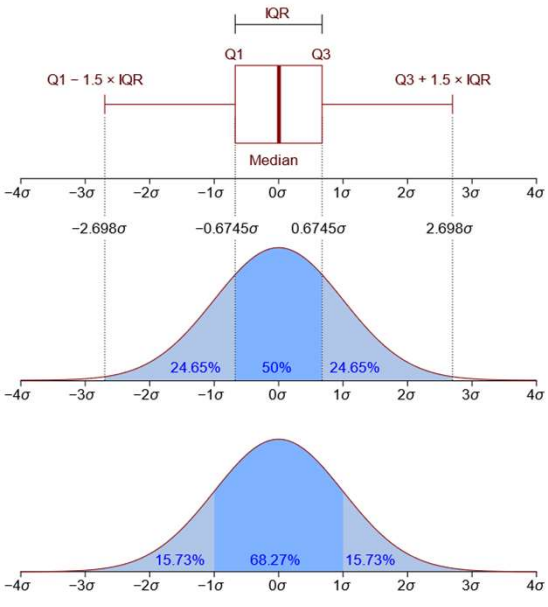
ある事象が起きる確率が決まっているという性質をもつ変数
→確率変数（random variable）

確率変数がどのような値になるかという法則性を与えるもの
→確率分布（probability distribution）
※母集団のばらつき具合を確率分布としてとらえる

108

2.統計的推定（点推定・区間推定）

基本的な分布



出典：Wikipedia（確率密度関数）

109

2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：確率密度関数）

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
```

※必要なパッケージ等の読み込み

```
norm.pdf(0)
```

```
0.3989422804014327
```

0が発生する確率

110

2.統計的推定（点推定・区間推定）

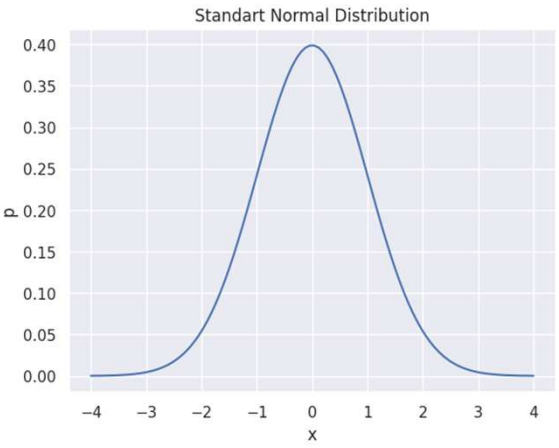
基本的な分布（正規分布：確率密度関数）

```
x = np.linspace(-4, 4, 100)
```

-4から4までの区間で100個の値で近似

```
y_pdf = norm.pdf(x)
```

```
plt.plot(x,y_pdf)
plt.xlabel('x')
plt.ylabel('p')
plt.title('Standart Normal Distribution')
pass
```



111

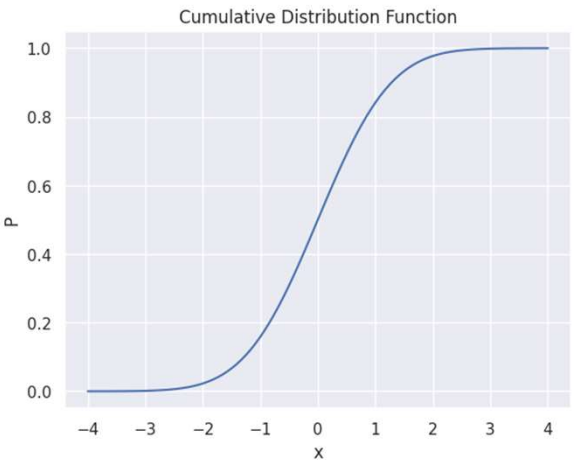
2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：累積分布関数）

```
y_cdf = norm.cdf(x)
```

Cumulative Distribution Function

```
plt.plot(x, y_cdf)
plt.xlabel('x')
plt.ylabel('P')
plt.title('Cumulative Distribution
Function')
pass
```



112

2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：累積分布関数）

```
norm.cdf(0)
```

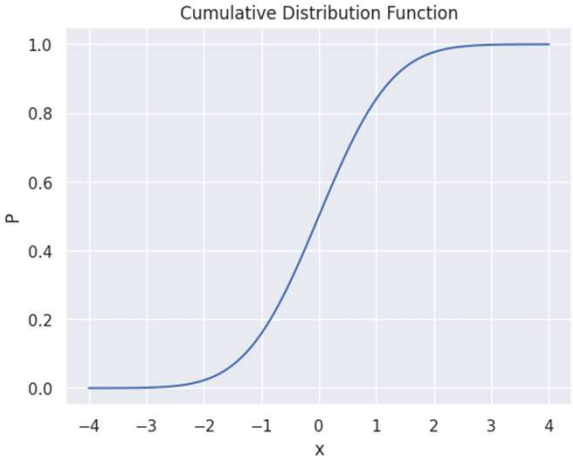
0以下を取る確率
（平均0を中心に左右対称のため）

```
0.5
```

```
norm.cdf(-4)
```

xが-4以下の場合

```
3.167124183311986e-05
```



113

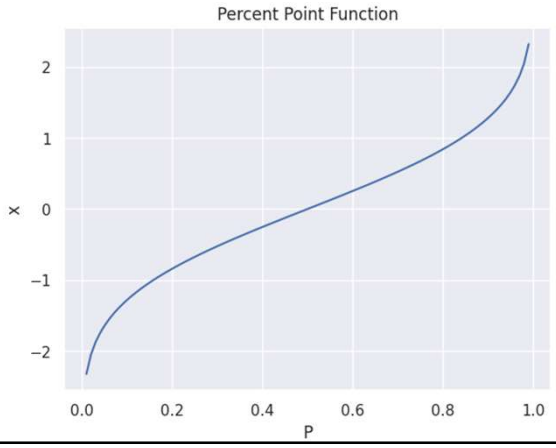
2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：パーセント・ポイント関数）

```
p = np.linspace(0,1,100)
y_ppf = norm.ppf(p)
```

Percent Point Function

```
plt.plot(p,y_ppf)
plt.xlabel('P')
plt.ylabel('x')
plt.title('Percent Point Function')
pass
```



114

2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：パーセント・ポイント関数）

```
norm.ppf(0.5)
```

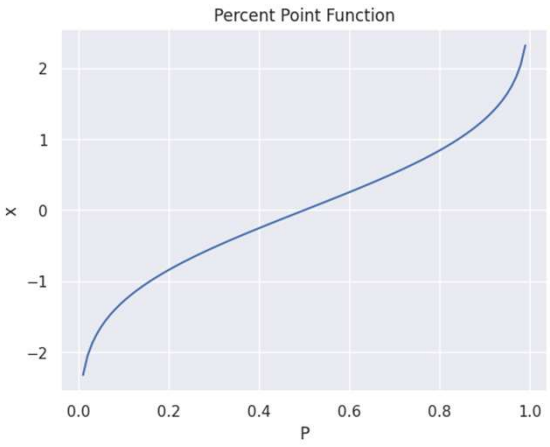
P=0.5の場合

```
0.0
```

```
norm.ppf(0.025)
```

P=0.025の場合

```
-1.9599639845400545
```



115

2.統計的推定（点推定・区間推定）

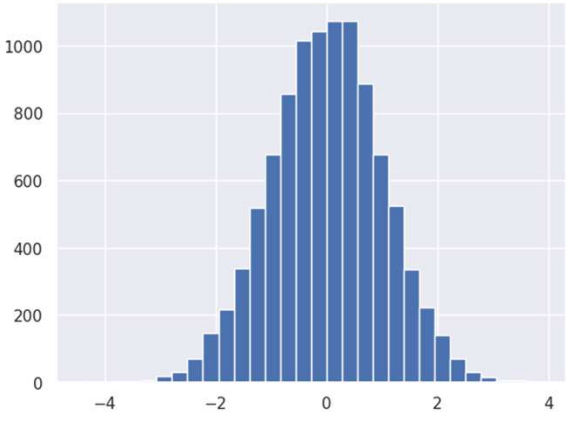
基本的な分布（正規分布：ランダム変数生成関数）

```
y_rvs = norm.rvs(size=10_000)
```

10,000個のランダム変数を生成

```
plt.hist(y_rvs, bins=30)
pass
```

bins=表示する棒の数



116

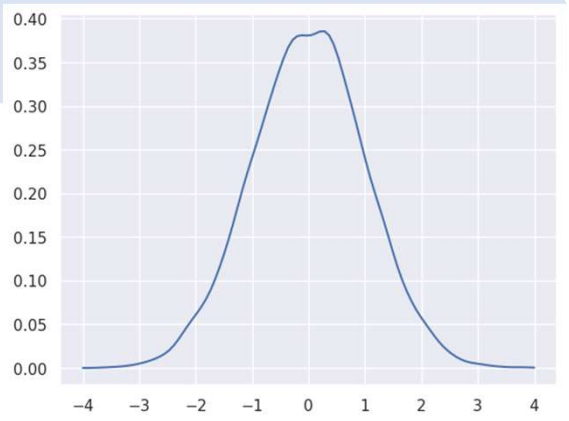
2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：ランダム変数生成関数）

```
from scipy.stats import gaussian_kde
kde = gaussian_kde(y_rvs)
plt.plot(x, kde(x))
pass
```

サブパッケージの読み込みと
y_rvsから確率密度関数を推定

カーネル密度推定：
統計学において、確率変数の確率密度関数を推定するノンパラメトリック手法のひとつ



117

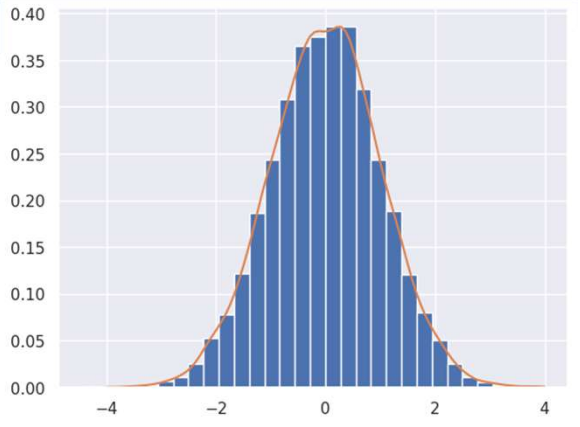
2.統計的推定（点推定・区間推定）

基本的な分布（正規分布：ランダム変数生成関数）

```
plt.hist(y_rvs, bins=30, density=True)
plt.plot(x, kde(x))
pass
```

density=Trueでplt.histとplt.plotの
カーネル密度関数が同じスケールに

推定なので標準正規分布の確率密度と全く同じにならないが，非常に近い



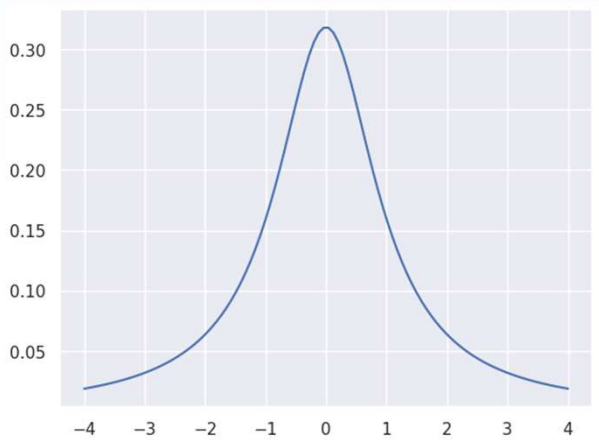
118

2.統計的推定（点推定・区間推定）

基本的な分布（ t 分布）

```
from scipy.stats import t
x = np.linspace(-4, 4, 100)
y = t.pdf(x, df=1)
plt.plot(x,y)
pass
```

dfn:自由度（degree of freedom）



母分散が未知の場合の母平均に関する推論
・母集団が正規分布に従う
・σが未知
・n<30

119

2.統計的推定（点推定・区間推定）

基本的な分布（ t 分布）

```
t.cdf(-3, df=1)
```

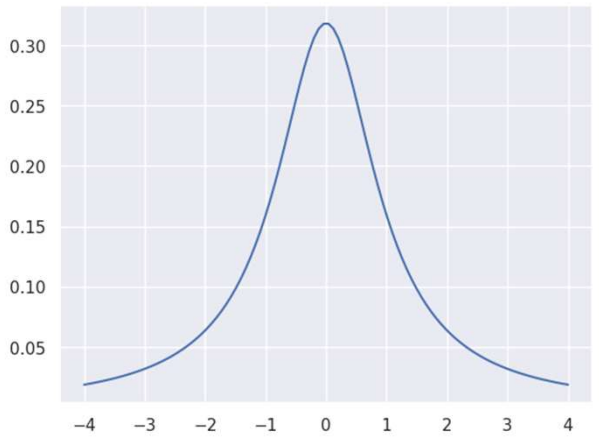
```
0.10241638234956672
```

自由度1の場合にxが-3以下となる確率

```
1-t.cdf(3, df=1)
```

```
0.10241638234956672
```

自由度1の場合にxが3以上となる確率



120

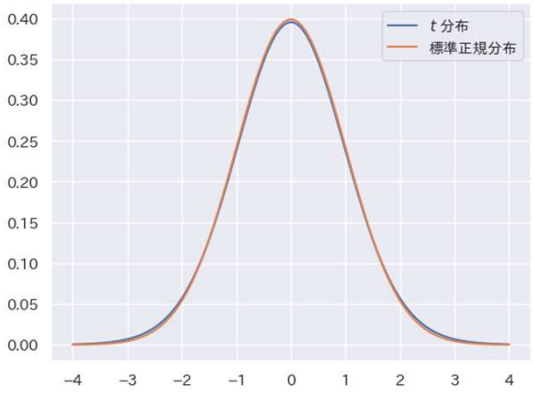
2.統計的推定（点推定・区間推定）

基本的な分布（ t 分布）

```
!pip install japanize-matplotlib
```
japanize-matplotlibのインストール

```
import japanize_matplotlib
```
インポート

```
x = np.linspace(-4, 4, 100)
plt.plot(x, t.pdf(x, 30), label=r'$t$ 分布')
plt.plot(x, norm.pdf(x), label='標準正規分布')
plt.legend()
pass
```
自由度30で標準正規分布と重ねてプロット

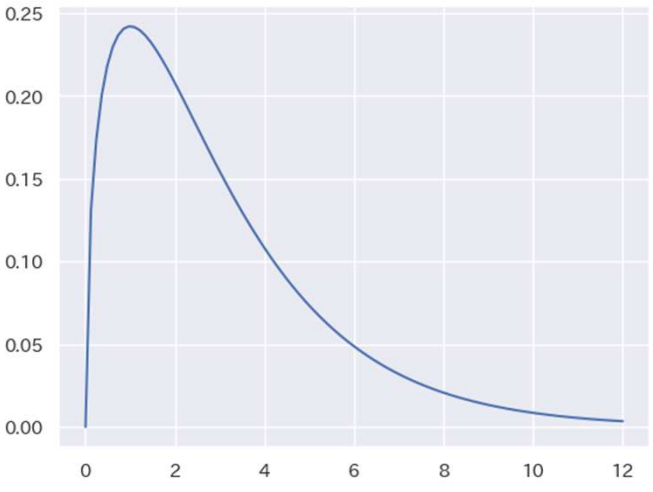


121

2.統計的推定（点推定・区間推定）

基本的な分布（X^2分布）

```
from scipy.stats import chi2
x = np.linspace(0,12,100)
y = chi2.pdf(x, df=3)
plt.plot(x,y)
pass
```



122

2.統計的推定（点推定・区間推定）

基本的な分布（X²分布）

```
chi2.cdf(1, df=3)
```

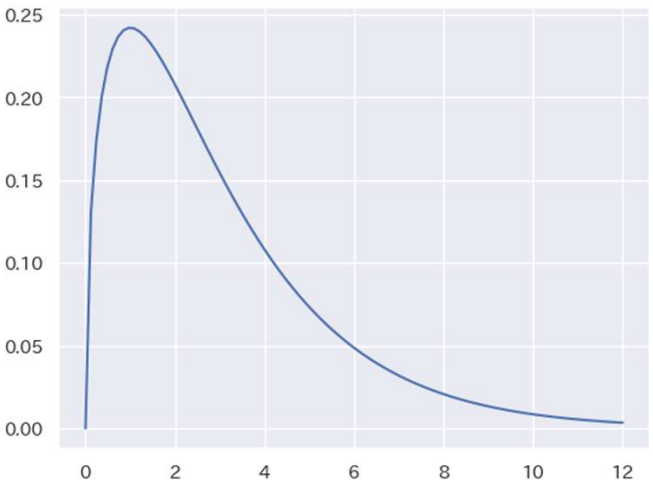
```
0.19874804309879915
```

自由度3の場合にxが1以下の確率

```
1-chi2.cdf(10, df=3)
```

```
0.0185661354630432
```

自由度3の場合にxが10以上となる確率



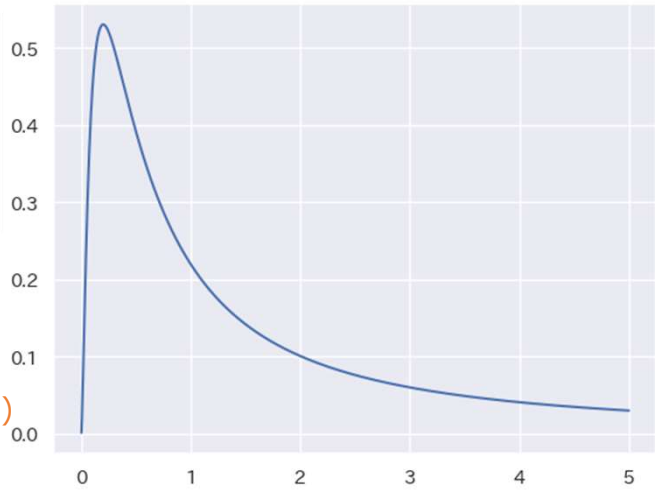
123

2.統計的推定（点推定・区間推定）

基本的な分布（F分布）

```
from scipy.stats import f
x = np.linspace(0.001, 5, 1000)
y = f.pdf(x, dfn=5, dfd=1)
plt.plot(x, y)
pass
```

dfn:分子の自由度
（numerator degree of freedom）
dfd:分母の自由度
（denominator degree of freedom）



124

2.統計的推定（点推定・区間推定）

基本的な分布（X^2分布）

```
f.cdf(0.1, dfn=5, dfd=1)
```

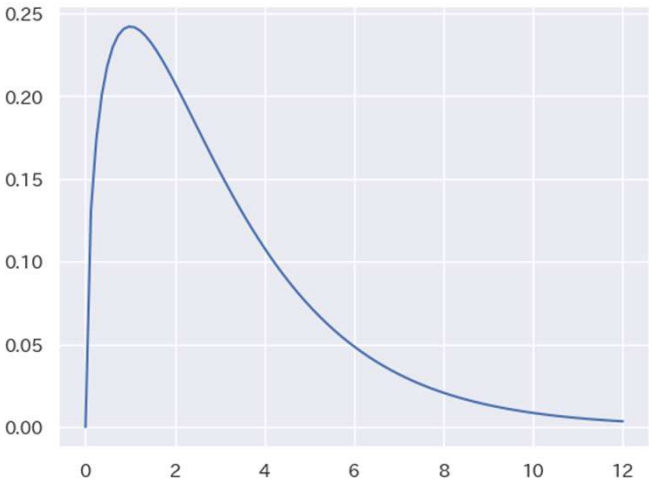
```
0.02503101581845294
```

dfn=5, dfd=1の場合にxが0.1以下の確率

```
1-f.cdf(5, dfn=5, dfd=1)
```

```
0.32657156446244606
```

dfn=5, dfd=1の場合にxが5以上の確率



125

2.統計的推定（点推定・区間推定）

点推定

- 母集団分布のパラメータである1つの値として指定する推定方法

区間推定

- 推定値に幅を持たせた推定方法

126

2.統計的推定（点推定・区間推定）

点推定

- 架空のデータのサイズ10，このデータは正規母集団からの無作為標本であると仮定する

```
weight = [8.033, 7.298, 6.223, 7.538, 2.546, 9.251, 5.006, 5.769, 9.628, 6.512]
```

- 母平均を推定する場合は標本平均を，母分散を推定する場合は不偏分散を，推定量として扱う

127

2.統計的推定（点推定・区間推定）

点推定

- 母平均を推定する場合は標本平均を，母分散を推定する場合は不偏分散を，推定量として扱う

```
x_bar = np.mean(weight)
u2 = np.var(weight, ddof=1)

print('標本平均 : ', round(x_bar, 3))
print('不偏分散 : ', round(u2, 3))
```

```
標本平均: 6.78
不偏分散: 4.345
```

128

2.統計的推定（点推定・区間推定）

区間推定（用語の整理）

- 信頼係数
  - 区間推定の幅における信頼の度合いを確率で表現したもの
- 信頼区間
  - ある信頼係数を満たす区間
- 信頼限界
  - 信頼区間の下限値（下側信頼限界）と上限値（上側信頼限界）

129

2.統計的推定（点推定・区間推定）

区間推定（手順）

- 信頼区間95%として，母平均の区間推定
  - 母分散が明らかであれば標準正規分布が利用できるが普通はない
    → t 分布を活用する

130

2.統計的推定（点推定・区間推定）

区間推定（手順）

1. 標本平均$\bar{X}$と標準誤差SEを計算
2. サンプルサイズをnとするとき，自由度$n-1$のｔ分布における2.5%点と97.5%点を計算する
   - ｔ分布における2.5%点を$t_{0.025}$と表記
   - ｔ分布における97.5%点を$t_{0.975}$と表記
   - ｔ分布における従う確率変数が$t_{0.025}$以上$t_{0.975}$以下になる確率は95%
   - このとき95%が信頼係数となる
3. $\bar{X} - t_{0.975} * SE$が下側信頼限界となる
4. $\bar{X} - t_{0.025} * SE$が上側信頼限界となる

131

2.統計的推定（点推定・区間推定）

区間推定

```
n = len(weight)
df = n - 1
u = np.std(weight, ddof=1)
se = u / np.sqrt(n)

print('サンプルサイズ  : ', n)
print('自由度          : ', df)
print('標準偏差        : ', round(u, 3))
print('標準誤差        : ', round(se, 3))
print('標本平均        : ', round(x_bar, 3))
```

区間推定に必要なのは
自由度，標本平均，標準誤差
標本平均は計算済み

```
サンプルサイズ : 10
自由度       : 9
標準偏差     : 2.085
標準誤差     : 0.659
標本平均     : 6.78
```

132

2.統計的推定（点推定・区間推定）

区間推定

```
t_025 = stats.t.ppf(q=0.025, df=df)
t_975 = stats.t.ppf(q=0.975, df=df)

print('t分布の2.5%点　: ', round(t_025, 3))
print('t分布の97.5%点 : ', round(t_975, 3))
```

自由度n-1の t 分布における2.5%点と97.5%点を計算

```
t分布の2.5%点 : -2.262
t分布の97.5%点 : 2.262
```

133

2.統計的推定（点推定・区間推定）

区間推定

```
lower_mu = x_bar - t_975 * se
upper_mu = x_bar - t_025 * se

print('下限信頼区間： ', round(lower_mu, 3))
print('上限信頼区間： ', round(upper_mu, 3))
```

t 分布は左右対称なので，
$t_{0.025}$ - $t_{0,975}$で信頼区間を計算

```
下限信頼区間: 5.289
上限信頼区間: 8.272
```

母平均の95%信頼区間は，5.289から8.272となった

134

2.統計的推定（点推定・区間推定）おまけ問題

区間推定

1. 標本$n = 10$
   405g, 395g, 374g, 410g, 417g, 426g, 383g, 398g, 390g, 402g
   母標準偏差：15g
   信頼水準95%で信頼区間を推定せよ

2. 工場で生産している製品Aがある
   以下のデータがわかっている
   標本$n = 100$，標本平均=150g，母分散＝15^2g
   母平均μを信頼水準95%で信頼区間を推定せよ

135

3.統計的仮説検定

統計的仮説検定

- データを使って何かを判断したいときに使われる手法
- さまざまな種類があり判断する対象も手法によって
  さまざま，単に検定と呼ぶ場合もある

- 統計的推定 → 母集団分布のパラメータを言い当てる試み
- 統計的検定 → 母集団のパラメータについて判断を下す
  例）「母平均が50か，あるいは50ではないか」を判断

136

3.統計的仮説検定 - 基本的な仮説検定 -

t検定（母平均に関する1標本のt検定）

- データは正規母集団からの無作為標本，と仮定
- 平均値が「ある値」と異なると言えるかどうか，を判断

例

内容量が135gと書かれたスナック菓子，測ってみると134gの場合もあれば，136gの場合もある。工場の機械に問題がないか検査する必要がある。
「スナック菓子の内容量の母平均が135gと異なっていると言えるかどうか」という判断をサポート。

137

3.統計的仮説検定 - 基本的な仮説検定 -

仮説検定の流れ

1. 仮説を立てる
2. 有意水準を決める
3. 検定統計量を計算
4. p値を計算
5. p値より有意水準が大きいか
   - （Yes）帰無仮説を採択
   - （No）帰無仮説を棄却

138

3.統計的仮説検定 - 基本的な仮説検定 -

仮説検定の流れ

1. 仮説を立てる
2. 有意水準を決める
3. 検定統計量を計算
4. p値を計算
5. p値より有意水準が大きいか
   - （Yes）帰無仮説を採択
   - （No）帰無仮説を棄却

139

3.統計的仮説検定 - 基本的な仮説検定 -

1. 仮説を立てる

- 帰無仮説H_0：スナック菓子の母平均は135gである
- 対立仮説H_1：スナック菓子の母平均は135gと異なる

- 帰無仮説が棄却されたならば，有意差あり，つまり「スナック菓子の母平均は135gと異なる」と判断する
  （135gより大きい or 少ない，ではない）

140

3.統計的仮説検定 - 基本的な仮説検定 -

仮説検定の流れ

1. 仮説を立てる
2. 有意水準を決める
3. 検定統計量を計算
4. p値を計算
5. p値より有意水準が大きいか
   - （Yes）帰無仮説を採択
   - （No）帰無仮説を棄却

141

3.統計的仮説検定 - 基本的な仮説検定 -

有意水準を決める

- 有意である：偶然ではなく，何か意味があるということ
- 帰無仮説が間違っていると判断（棄却）する区間のことを棄却域（rejection region），採択される区間を採択域（acceptance region）といい，この境の基準となる確率のこと
- つまり，帰無仮説を棄却する基準
- 伝統的にαと表記し，5%や1%が使われることが多い

142

3.統計的仮説検定 - 基本的な仮説検定 -

有意水準を決める（危険率）

- 第一種の過誤：帰無仮説が正しいのに誤って帰無仮説を棄却してしまうこと
- 第二種の過誤：帰無仮説が間違っているのに，誤って帰無仮説を採択してしまうこと

- 有意水準は第一の過誤を許容できる確率

143

3.統計的仮説検定 - 基本的な仮説検定 -

仮説検定の流れ

1. 仮説を立てる
2. 有意水準を決める
3. 検定統計量を計算
4. p値を計算
5. p値より有意水準が大きいか
   - （Yes）帰無仮説を採択
   - （No）帰無仮説を棄却

144

3.統計的仮説検定 - 基本的な仮説検定 -

検定統計量を計算（準備）

```
import numpy as np
import pandas as pd
from scipy import stats

food = [111, 124, 125, 126, 127, 134, 135, 136, 139, 141]
```

- 帰無仮説H_0：スナック菓子の母平均は135gである
- 対立仮説H_1：スナック菓子の母平均は135gと異なる
- 有意水準は5%

145

3.統計的仮説検定 - 基本的な仮説検定 -

標本平均

```
x_bar = np.mean(food)
round(x_bar, 3)
```

```
129.8
```

自由度

```
n = len(food)
df = n - 1
df
```

```
9
```

146

3.統計的仮説検定 - 基本的な仮説検定 -

標準誤差（標準偏差 / サンプルサイズの平方根）

```
u = np.std(food, ddof=1)
se = u / np.sqrt(n)
round(se, 3)
```

```
2.839
```

t値の計算

```
t_sample = (x_bar - 135) / se
round(t_sample, 3)
```

```
-1.831
```

147

3.統計的仮説検定 - 基本的な仮説検定 -

棄却域の計算

```
round(stats.t.ppf(q=0.025, df=df), 3)
```

```
-2.262
```

```
round(stats.t.ppf(q=0.975, df=df), 3)
```

```
2.262
```

$-t_{0.025} < |t_sample| < t_{0.975}$の範囲が採択域となる，
よって帰無仮説を棄却することはできない

148

3.統計的仮説検定 - 基本的な仮説検定 -

仮説検定の流れ

1. 仮説を立てる
2. 有意水準を決める
3. 検定統計量を計算
4. p値を計算
5. p値より有意水準が大きいか
   - （Yes）帰無仮説を採択
   - （No）帰無仮説を棄却

149

3.統計的仮説検定 - 基本的な仮説検定 -

p値の計算

```
p_value = stats.t.cdf(-np.abs(t_sample), df=df) * 2
round(p_value, 3)
```

```
0.1
```

p値が有意水準0.05を上回っているので，帰無仮説を支持する確率が高い

スナック菓子の平均重量は135gと有意に異なっていないと判断することができる

150

3.統計的仮説検定 - 基本的な仮説検定 -

stats.ttest_1samp関数（1標本のt検定）

```
stats.ttest_1samp(food, 135)
```

```
TtestResult(statistic=-1.831369433567421, pvalue=0.10027730072021666, df=9)
```

statisticがt値，pvalueがp値

151

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

2郡のデータに対するt検定

- 2つの変数の間で平均値に差があるかどうか，を判断

例

あるボディビルにより体重の増減がおこるかどうか調べるために，10人についてその効果を調べる場合など，
「同じ対象を異なった条件で２回測定して，その違いをみる」とった場合

152

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

2郡のデータに対するt検定

	正規分布を仮定できる	正規分布を仮定できない
対応あり	対応のあるｔ検定※1	ウィルコクソンの符号付き順位検定
対応なし	対応のないｔ検定※2	マン・ホイットニーのU検定

※1　「データの差」を取ってから母平均に関する１標本のｔ検定
※2　「平均値の差」に注目する

153

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

1. 仮説を立てる

- 帰無仮説H_0：ボディビルにより体重の増減は起こっていない
- 対立仮説H_1：ボディビルにより体重の増減は起こっている

2. 有意水準を決める

- 有意水準5%とし，p値が0.05を下回れば，帰無仮説が棄却され，ボディビルでの体重増減への有意な変化が認められると主張できる

154

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

3. 検定統計量を計算する

No.	1	2	3	4	5	6	7	8	9	10	平均
前	57.6	88.5	73.5	77.1	64.9	93.0	76.2	79.4	89.4	61.7	
後	61.2	90.7	73.4	82.6	66.7	93.7	78.0	84.4	88.0	64.0	
後－前	3.6	2.2	-0.1	5.5	1.8	0.7	1.8	5	-1.4	2.3	2.14

155

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

3. 検定統計量を計算する（ライブラリの読み込み等）

```
from google.colab import files
uploaded = files.upload()
```

```
import numpy as np
import pandas as pd
from scipy import stats
```

```
bodybuilding = pd.read_csv('sample_t-test.csv')
print(bodybuilding)
```

```
    person bodybuilding  weight
0        1     before    57.6
1        2     before    88.5
2        3     before    73.5
3        4     before    77.1
4        5     before    64.9
5        6     before    93.0
6        7     before    76.2
7        8     before    79.4
8        9     before    89.4
9       10     before    61.7
10       1      after    61.2
11       2      after    90.7
12       3      after    73.4
13       4      after    82.6
14       5      after    66.7
15       6      after    93.7
16       7      after    78.0
17       8      after    84.4
18       9      after    88.0
19      10      after    64.0
```

156

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

3. 検定統計量を計算する

```
before = bodybuilding.query('bodybuilding == "before"')['weight']
after = bodybuilding.query('bodybuilding == "after"')['weight']
```

ボディビル前と後の標本平均

```
before = np.array(before)
after = np.array(after)
```

アレイに変換

```
diff = after - before
diff
```

差を計算

```
array([ 3.6,  2.2, -0.1,  5.5,  1.8,  0.7,  1.8,  5. , -1.4,  2.3])
```

157

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

対応のない t 検定（不等分散）

```
#平均値
x_bar_bef = np.mean(before)
x_bar_aft = np.mean(after)

#分散
u2_bef = np.var(before, ddof=1)
u2_aft = np.var(after, ddof=1)

#サンプルサイズ
m = len(before)
n = len(after)

#t値
t_value = (x_bar_aft - x_bar_bef) / np.sqrt((u2_bef/m + u2_aft/n))
round(t_value, 3)
```

```
0.406
```

158

3.統計的仮説検定 - 2つの標本問題に関する仮説検定 -

対応のない t 検定（不等分散）

```
df = (u2_bef / m + u2_aft / n)**2 / ((u2_bef / m)**2 / (m-1) + (u2_aft / n)**2 / (n-1))
round(df, 3)
```

```
17.963
```

```
p_value = stats.t.cdf(-np.abs(t_value), df=df)*2
round(p_value,5)
```

```
0.68988
```

```
stats.ttest_ind(after, before, equal_var=False)
```

```
TtestResult(statistic=0.4055321209937369, pvalue=0.6898758363640115, df=17.963028980728833)
```

159

今日の内容

1. データ分析に必要な統計学の基礎を学ぶ
2. 比較して2変数の関係を考える
3. データに基づいて判断を下すための手法を学ぶ
4. ビジネスにおける予想と分析結果の報告

160