データサイエンス特論
Data Science

統計の基礎

1．　統計とは
  - 基本統計量
  - 点推定，区間推定

2．　仮説検定

(C) 創造技術コース　橋本洋志／大久保友幸
{hashimoto, ohkubo-tomoyuki}@aiit.ac.jp
http://hhlab.org/

統計とは

統計とは、集団の傾向・性質などを数量的に明らかにすること（多数の定義あり）。

統計解析とは標本を調べることによって，母集団の性質を解析することが目的である（統計とほぼ同義）。

標本から得られた統計量をもとにして，母数の存在する範囲を求めることを統計的推定（statistical inference）　という。

点推定（point estimation）と区間推定（interval estimation）がある

点推定とは，母集団の母数を一つの値で推定すること

区間推定とは，例えば，　６００件調査のTV視聴率が15%±3%以内は，９７．５%の信頼度で確からしい。

点推定

点推定の良否を測る性質

不偏性　推定量 $\hat{\theta}$ の期待値が母数 θ に一致する。すなわち，

$$E[\hat{\theta}] = \theta \tag{5.1}$$

が成立するとき推定量 $\hat{\theta}$ を**不偏推定量**（unbiased estimator）と呼ぶ。この性質を持つとき，θ の周りで $\hat{\theta}$ は分布する。

一致性　標本の数が増えるにつれて，推定量 $\hat{\theta}$ は対応する母数に近づくことが望ましい。この性質を表したのが次の式である[*1]。

$$\lim_{N \to \infty} \mathrm{P}\left(\left|\hat{\theta}_N - \theta\right| < \varepsilon\right) = 1 \tag{5.2}$$

このような性質を持つとき，$\hat{\theta}$ を**一致推定量**（consistent estimator）と呼ぶ。

点推定：標本の平均と分散

標本平均 $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$

標本分散 $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$

ハット"^"が付けば，
真値ではなく，
標本計算による推定を意味する

- 標本平均は，不偏推定量，かつ，一致推定量

$$E[\hat{\mu}] = \frac{1}{N} E\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N} N\mu = \mu \qquad E\left[(\hat{\mu} - \mu)^2\right] = \frac{1}{N}\sigma^2$$

- 標本分散は，不偏推定量，しかし！

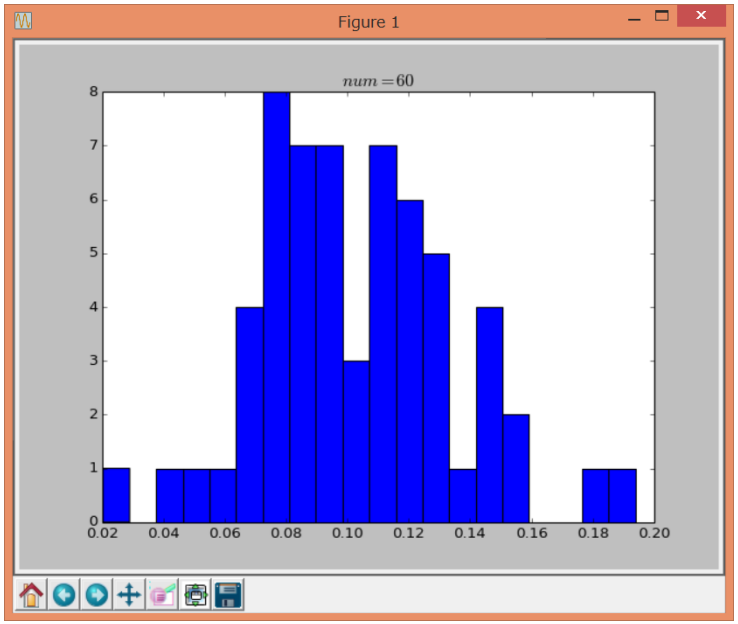
$$\hat{\sigma}^2_{bias} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2 \qquad E\left[\hat{\sigma}^2_{bias}\right] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

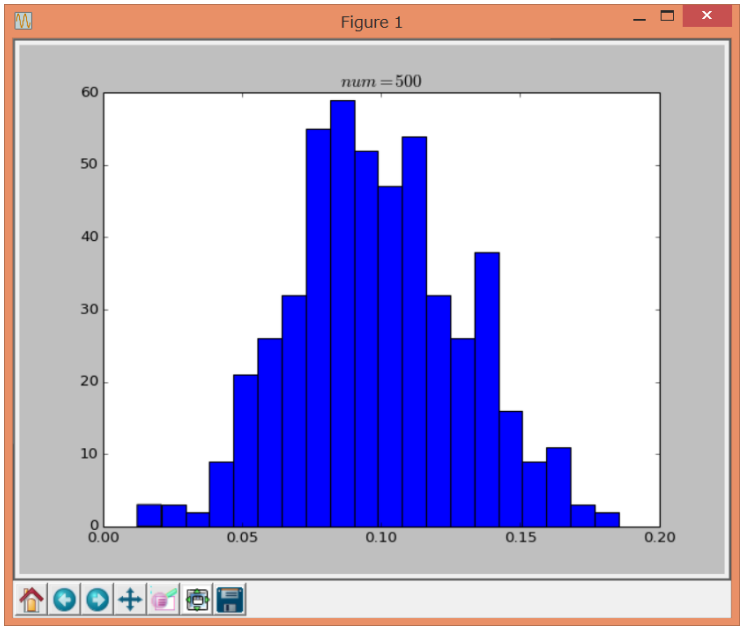
標本平均は確率変数であることを確かめる

PRB_NormalDistribution

```
mean = 0.1
std = 0.1
for i in range(num):
        x=scipy.stats.norm.rvs(loc = mean, scale = std, size=m )
        mu[i] = x.mean()
```

- μ(m のこと)＝0．1の周りで分布している。“分布”は、すなわち、標本平均は確率変数であることを示している。 → 中心極限定理ですでにみた。





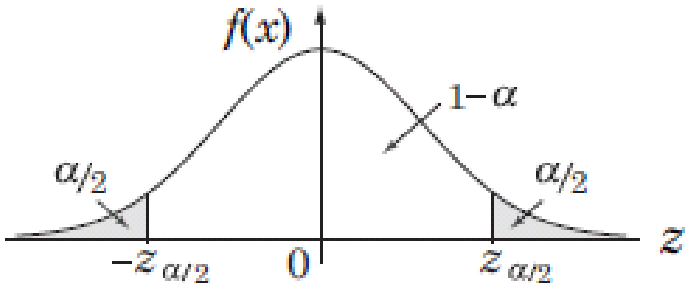
区間推定

【信頼区間】（confidence interval）　これは，$(1-\alpha)$ の確率（または，信頼度）で真の母数の値 θ が区間 $[L, U]$ に入る区間のことを言う。これを定式化すると

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{5.11}$$

で表わされる。このとき，L と U を求めることが主たる目的となる。ここに，L, U はそれぞれ，**下側信頼限界**（lower confidence limit），**上側信頼限界**（upper confidence limit），$(1-\alpha)$ は**信頼度**（confidence level）または**信頼係数**（confidence coefficient）といい，区間 $[L, U]$ を **$100(1-\alpha)\%$ 信頼区間**（または，単に**信頼区間**）と呼ぶ。

$1-\alpha$ は，目的に応じて適当な値が選ばれるが，通常 0.90, 0.95, 0.99 が選ばれる



$1-\alpha$	0.9	0.95	0.99
$\alpha/2$	0.05	0.025	0.005
$z_{\alpha/2}$	1.645	1.960	2.576

図 5.2　標準正規分布における信頼度 $1-\alpha$ と信頼区間

区間推定

図 5.2 に標準正規分布における各信頼度と $z_{\alpha/2}$ を示す。この $z_{\alpha/2}$ を計算で求めるには，例えば，$1-\alpha=0.9$ の場合には，次のようにすればよい。

既に説明した「パーセント点から確率を求める」を参照

```
alp = 0.05
alp2 = alp/2
z_alp2 = scipy.stats.norm.ppf ( 1-alp2 )
print 'alp2 = %f, z_alp2 = %f' % (alp2, z_alp2)
```

関数 `norm.ppf()` は片側を計算する。
両側の場合を計算したいときの工夫が`alp2`である。

結果： alp = 0.05 のときの、 z_alp2 = $z_{\alpha/2}$ が求まる。
alp2 = 0.025000,　z_alp2 = 1.959964

平均値μの信頼区間

- 母分散（真の分散）σ^2　が既知，未知で場合分けする

【母平均の信頼区間】　μ の信頼度 $1-\alpha$ の信頼区間は

母分散 σ^2 が既知の場合　$\left[\hat{\mu} - z_{\alpha/2} \cdot SE(\sigma),\ \ \hat{\mu} + z_{\alpha/2} \cdot SE(\sigma)\right]$　(5.15)

母分散 σ^2 が未知の場合　$\left[\hat{\mu} - t_{\alpha/2} \cdot SE(\hat{\sigma}),\ \ \hat{\mu} + t_{\alpha/2} \cdot SE(\hat{\sigma})\right]$　(5.18)

考察として，信頼区間 (5.15) 式，(5.18) 式の両方に共通して言えることは

- 信頼度を高くすれば信頼区間が広がる。逆もまた然りである。これは，直感的にわかることであろう。しかし，高すぎる信頼度は，あまり意味をなさない。
- $SE(\sigma)$, $SE(\hat{\sigma})$ 共に N の増加とともに小さくなり，信頼区間が狭まる。すなわち，標本数を増やせば，その分，母平均の値の絞り込みができることになる。

$\hat{\mu}$ を計算してから，両式とも　SE()を計算して，次に z_{α/2}または t_{α/2}を計算すればよい。ここに，

$$SE(\sigma) = \sigma / \sqrt{N}\ , \qquad SE(\hat{\sigma}) = \hat{\sigma} / \sqrt{N}$$

【例題 1】　小学校のある学年の全国児童数は 110 万人とする。この児童への全国テストの平均値を推定するため，$N = 10$ 人を無作為抽出して，これに対する標本平均は $\hat{\mu} = 145.2$ 点だった。また，標本標準偏差 $\hat{\sigma}$ は 23.7 点だった。このとき，信頼度が 0.99, 0.95, 0.90 に対する信頼区間を求めよう。

母分散 σ^2 が未知の場合　$$\left[\hat{\mu} - t_{\alpha/2} \cdot SE(\hat{\sigma}),\ \ \hat{\mu} + t_{\alpha/2} \cdot SE(\hat{\sigma})\right] \qquad (5.18)$$

［解］　$N = 10$ の場合，標本数が少ないので，t 分布を用いた (5.18) 式に基づいて信頼区間を求める。信頼度 0.99, 0.95, 0.90 に対する $\alpha/2$ はそれぞれ 0.005, 0.025, 0.05 であるから，自由度 $9(= N - 1)$ の $t_{\alpha/2}$

STA_Estimation

```
N = 10
mu_hat = 145.2
std_hat = 23.7
t1 = t.interval( 0.99, df=N-1)
t2 = t.interval( 0.95, df=N-1)
t3 = t.interval( 0.90, df=N-1)
se = std_hat / np.sqrt(N)

print('1-alp = 0.99, interval:', mu_hat + t1[0]*se, mu_hat + t1[1]*se)
print('1-alp = 0.95, interval:', mu_hat + t2[0]*se, mu_hat + t2[1]*se)
print('1-alp = 0.90, interval:', mu_hat + t3[0]*se, mu_hat + t3[1]*se)
```

ここに、次の計算を用いている。

$$SE\,(\hat{\sigma}) = \frac{23.7}{\sqrt{10}} = 7.4946$$

答え

$$\left[145.2 - t_{\alpha/2} \cdot 7.4946,\ \ 145.2 + t_{\alpha/2} \cdot 7.4946\right] = [120.844,\ \ 169.556]$$
$$= [128.246,\ \ 162.154]$$
$$= [131.462,\ \ 158.938]$$

α

1%
5%
10%

信頼区間が広がるとはどういう意味？

母比率の信頼区間

【母比率の信頼区間】 信頼度 $1-\alpha$ の信頼区間は次となる。 $\boxed{SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}}$

$$\left[\hat{p} - z_{\alpha/2} \cdot SE\left(\hat{p}\right),\quad \hat{p} + z_{\alpha/2} \cdot SE\left(\hat{p}\right)\right] \tag{5.31}$$

【例題 2】 内閣支持率を調べるために，世論調査を行った。サンプル数は 1000 人で，支持する人の数は 550 人であった。信頼度 95 % の信頼区間を求めよう。

［解］ $\hat{p} = 550/1000 = 0.55$, $z_{\alpha/2} = 1.96$, $SE\left(\hat{p}\right) = \sqrt{0.55(1-0.55)}/\sqrt{1000} = 0.01573$ の値を (5.31) 式に代入すると，$\hat{p} \pm 1.96 \times 0.01573 = 0.55 \pm 0.0308$ ，すなわち，$0.5192 \leq p \leq 0.5808$ となる。

信頼度を高めるには，サンプル数は？

この例題において，信頼度を 95 % より高めた調査を行いたいと考えたとき，二つの策がある。1 番目は，信頼区間を広げることである。例えば，信頼区間を $[0, 1]$ にすれば，p はこの間にあることは 100 % 確信できる（現実には，意味のない話である）。2 番目は，標本数 N を大きくすることである。それでは，どれだけ大きくするかについて考えてみよう。

先の母比率の信頼区間を求める例題の場合，信頼度 95 % の信頼区間の幅は $2 \times 1.96\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}$ であった。ここで，

$$\hat{p}(1-\hat{p}) = -\hat{p}^2 + \hat{p} = -\left(\hat{p} - \frac{1}{2}\right)^2 + \frac{1}{4} \leq \frac{1}{4} \tag{5.32}$$

であるから，$\hat{p}(1-\hat{p})$ の最大値は $1/4$ である。したがって，信頼区間の幅は広くても

$$2 \times 1.96\sqrt{\frac{1}{4N}} = 1.96\frac{1}{\sqrt{N}} \tag{5.33}$$

である。(5.33) 式を見てわかるように，信頼区間の幅を半分にしようとするならば，標本数はその 2 乗の 4 倍にする必要があることがわかる。

また, この幅を 0.06 ($= \pm 3\%$)以内に収まるようにすれには, $1.96/\sqrt{N} \leq 0.06$ が成り立つ N を求めればよい。すなわち，$\sqrt{N} \geq 1.96/0.06 = 32.67$ より，$N \geq 32.67^2 \simeq 1067$ を得る。この結果から, 標本数を 1000 にした先の例題では, 信頼区間の幅が $2 \times 0.0308 = 0.0616$ に近い結果を示している。

それでは，信頼度を 99% に引き上げるということは，信頼区間の幅が 0.01 であるから，$\sqrt{N} \geq 1.96/0.01 = 196$，よって，$N = 38416$ となる。信頼度を 95% から 99% に引き上げるには，標本数を約 38 倍にしなければならない。

視聴率の区間推定

- **視聴率はTV番組をどれだけ見ているかの標本調査である。ただし，実際の視聴率に対する推定値である。**
  - 視聴率＝その番組を見ているTV台数／標本数（600台，関東の場合）
  - 関東地区，1995年国勢調査によると1455万世帯
  - 1世帯1台あると仮定すると，母集団の数は1455万台（全部調べるのは無理！）
  - 視聴率をpとおく。このpは，番組Mを見ている確率に他ならない。
  - 確率変数Xは，番組Mを見ている台数とする。
  - このとき，Xは二項分布（600,p）に従う（興味ある学生は自ら調べましょう！）
  - このとき，Xの平均値と標準偏差はそれぞれ次で表される。
  - $m = 600\,p, \quad \sigma = \mathrm{sqrt}\{600p(1-p)\}$
  - 95%の信頼度で，次の推定幅が求められる。
  - $X - 1.96\,\sigma \leqq m \leqq X + 1.96\,\sigma$
  - もし，実際に番組Mを見ていた台数が99台としたとき，$p = 99/600 = 16.5\ \%$
  - よって，$13.5\ \% \leqq p \leqq 19.5\ \%$，すなわち，実際に$p = 16.5\ \%$が得られたとしても，このpの値は，95 %の確率で，この区間に収まっている，というのが正しい。

まとめ

- **マスコミなどで，平均値，支持率がよく出現する。**
- **この値をどう見るかは，見る人間の問題となる。**
- **また，サンプル集団そのものが偏っている場合，得られた平均値，支持率そのものの信頼性が大きくことなる。**
  - マスコミの電話調査の問題
    - 昼間，有線電話にランダムに電話をかける。年齢層に偏りがある。
    - 質問の仕方も結果にバイアスを与える。
  - 支持率調査の問題
    - すでに政党支持がはっきりしている場合，日本人は，答えを濁したり，嘘の回答を示すことが多い。
    - 対処は，ここ数回の選挙結果，支持者の特定，地域性などを考慮する。

仮説検定
（Hypothesis testing）

1. 仮説検定とは
2. 片側検定，両側検定
3. 平均値の検定
4. 2標本の平均値の差の検定

仮説検定（hypothesis testing）とは

◘ 仮説検定とは

⇒ ある仮説に対して、それが正しいのか否かを統計学的に検証する手段

例：新薬の効き目

研究開発された新薬の効用を調べるため、複数の被験者に投与し、医学で定められる基準に基づいて、その効き目を評価した。この場合、仮説検定に基づくと、この評価は次の2つの仮説を立てる。

帰無仮説（H0）：薬の効き目がない。　（null hypothesis）

対立仮説（H1）：薬の効き目がある。　（alternative hypothesis）

◘ 棄却

H0を間違っていると判断することを　H0を棄却する、と言う。この場合、H1が成り立つことを意味する。

- しかし、この棄却が間違っていたら？？？
- または、棄却すべきだったのを棄却しなかったならば？？？？
- 例えば、
  - 新薬の効用があるのに、効き目が無いと判断した
  - 新薬の効用がないのに、効き目があると判断した

データの取得に対する過ち

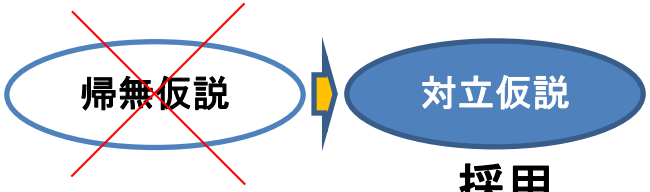
データの取得に対する過ち

- 採用すべきデータを取りこぼした：計測器の一時停止、見過ごした、などなど（第1種の過誤）
- 採用すべきでないデータを取得した：会議で大きな声の意見が採用された（第2種の過誤）

	真実	
判断	H0が正しい	H1が正しい
H0を棄却	第1種の過誤	正しい
H0を棄却しない	正しい	第2種の過誤

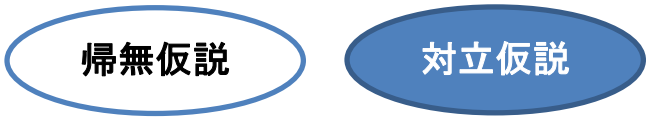
解釈の注意

H0を棄却の場合



採用

H0を棄却しない場合



どちらが正しいかは分からない

帰無仮説を棄却しない
⇒　帰無仮説が正しい、とはならない！

このことから、H0が棄却されなければ，仮説は無に帰する言葉として帰無が充てられた。帰無仮説は棄却されることで意味をもつ。
背理法に近い考え方と言える。背理法の例：√2が無理数であることの証明で，これが有理数と仮定することから論を出発させる。そして，矛盾を引き出し仮定を否定することで，命題を証明する方法論。

有意水準

- **有意水準（significance level）とは，仮説検定を行う場合に，帰無仮説を棄却するかどうかを判定する基準。例を通して説明する。**
  - 有意水準α= 5%のとき，帰無仮説H0は棄却された。
  - しかし，本当のところ，めったに起きないことがたまたま起きただけかもしれず，本当のところは，仮説は正しかったかもしれない。
  - このような誤りを犯す確率は，やはり有意水準α= 5%である。
  - この意味で，有意水準のことを**危険率**ともいう。
  - α＝5%とは，同じ状況下で検定を行うと20回に1回は検定を誤る危険性があることを意味する。
  - あるいは，1回の検定でH0が正しいにもかかわらず，誤ってH0を棄却する確率が0.05であることを意味する。
  - 通常は，α＝0.05（5%），0.01（1%），0.001（0.1%）が用いられる。

手順

1. 命題を立てる
2. 帰無仮説H0，対立仮説H1を設ける
3. 仮説検定の方法を選択する（命題により選定される）
4. 有意水準αを設定（5%, 1%が多い）
5. 検定を行い，検定量p値を求める（検定では，どの方法を用いても，p値と呼ぶことが多い）
6. p＜α：H0を棄却し，H1を採用する
   - すなわち，H0が生じる確率は大変小さいので棄却する，ということである。言い換えれば，滅多におこらないことが，たまたま生じただけであるから，棄却しても大丈夫であろう，しかし，第1種の過誤が生じる危険性は確率pあるとも言える。
7. p＞α：H0を棄却できない。H0を棄却できない場合には，H1を直ちに採択するのではなく，サンプル数，分析方法などを見直して，再検討をするのが良い。

片側検定と両側検定

片側検定と両側検定とは

- 帰無仮説H0は「平均μ=2.0である」というように，等号の形で設定されることが多い。これに対して，対立仮説H1を「μ > 2.0」または「 μ < 2.0」とおくことを片側検定（one tailed test / one side test）という。
- また，「 μ ≠ 2.0」のようにおくと， μは大きいか小さいかのいずれかに含まれるか否かを考えることになる。これを両側検定（two tailed test / two side test）という。
- 対立仮説を特に示さないときは，対立仮説を両側検定とするのが普通である。

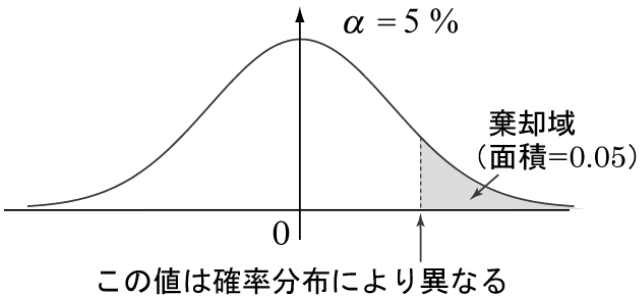
片側検定と両側検定

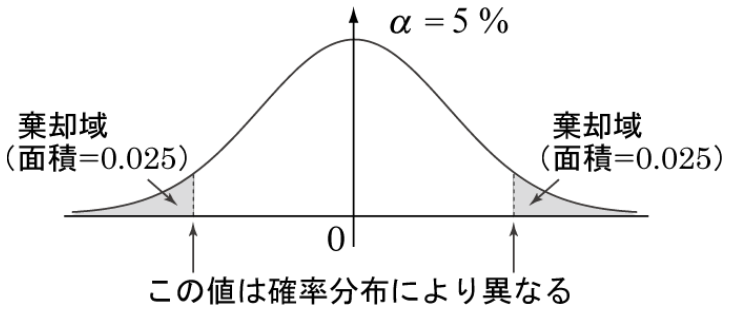
片側検定と棄却域

- 下図左に示すように，有意水準α=5%とは棄却域の面積が5%（0.05）を言う。
- 問題に適する検定統計量（後述）を求め，それに対応する確率分布を用いる。
- 検定統計量が棄却域（図の網掛け部分）に入ればH0を棄却，入らなければH0を棄却しない。
- 片側検定の場合，棄却域が負の領域にある場合もある。

両側検定と棄却域

- 同じαの場合を下図右に示す。両側だから，5%を二つに分けて，一つの棄却域の面積は0.025である。
- 検定統計量が棄却域（図の網掛け部分）に入ればH0を棄却，入らなければH0を棄却しない。





片側検定、両側検定

例：商品や製品の検定

- 片側検定：缶詰の重さは、規格では200g、N = 20（サンプリングが20個という意味）の重さの平均値は198gであった。重い分には消費者は文句を言わないから、これは、軽いのか否かだけを考える。

  H0：缶詰の重さ＝規格の重さ

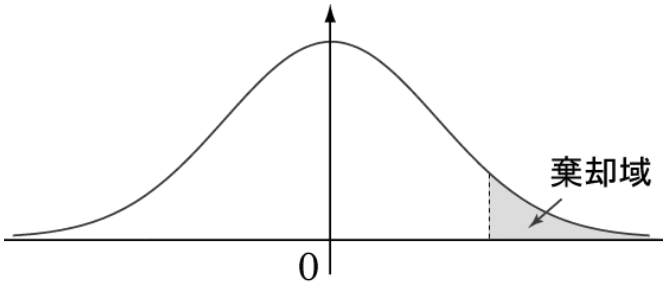
  H1：缶詰の重さ＜規格の重さ

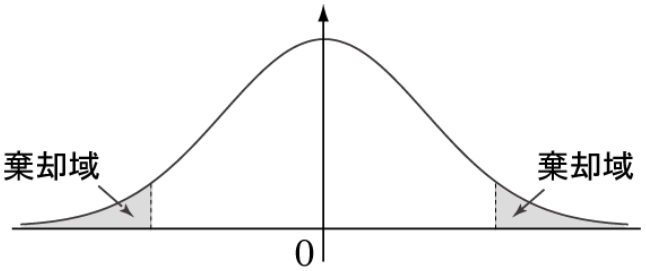
  ただし，20個の値はバラバラであるから，分散を考えなければならない。

- 両側検定：ある製品の長さは、仕様では1cm、N = 10の長さの平均値は1.1cmであった。製品の長さは短くても長くてもダメであるから、仕様の範囲に入っているのか？

まず、上の両者ともある検定統計量を計算し、それが、

- 下に示す棄却域に入るか否かを確かめる。
- また，その検定統計量が起こりうる確率p値を求め，そのp値が，有意水準αより大きいか否かを確かめる。





statsを用いたパーセント点の求め方

- **標準正規分布N(0,1)を対象とする**

```
z = 1.7
prob = scipy.stats.norm.cdf(z, loc=0.0, scale=1.0)
p = 1.0 - prob
print('one  side p value =',p)
print('both side p value =',2*p)

one  side p value = 0.0445654627585
both side p value = 0.0891309255171
```

z_αのとき，棄却域の面積は2*α

グラフの網掛け部分が棄却域，それ以外の部分が採択域

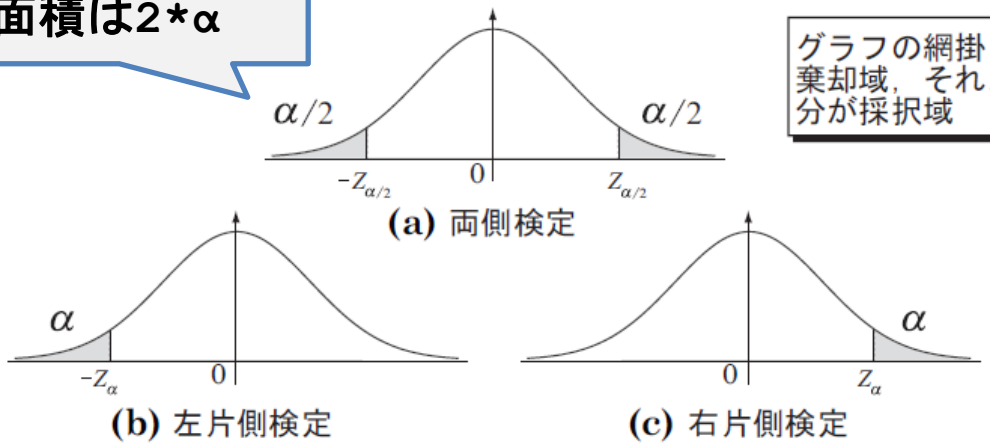


図 5.4　両側検定と片側検定（図中の z は正規分布，t 分布，χ^2 分布などの変数を代表して表す）

平均値の検定

- 母分散が既知の場合はz検定を用いるが，t検定とほぼ同じため省略する。
- 母分散が未知の場合に用いるt検定を説明する。
- 1標本の平均値検定と2標本の平均値の差の検定を説明する。
- Pythonを用いた検定を説明する。

z検定とt検定

標本x1, x2, … , x_N が互いに独立で正規分布N(μ_0, σ^2)に従うとする。この標本から得られる平均値$\hat{\mu}$が母平均μ_0をよく表しているか否かを検定したい。

天下り的に、次の2つの検定統計量を導入する。

□ 母分散σ^2が既知の場合(z検定という)

- 母平均μ_0の検定統計量

$$z = \frac{\hat{\mu} - \mu_0}{\sigma / \sqrt{N}} \quad \sim \quad N(0, 1)$$

- $\hat{\mu}$:標本平均、μ_0:母平均がこの値と仮説する値(帰無仮説で用いる)、σ:母標準偏差、N:サンプル数(標本数)
- 備考:μ_0はわからないのだから、μ_0は仮説として設けた値である。

□ 母分散σ^2が未知の場合(t検定という)

自由度(N−1)の t分布に従う, と読む

- 母平均μ_0の検定統計量

$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma} / \sqrt{N}} \quad \sim \quad t(N-1)$$

- $\bar{x}$; 標本平均、μ_0:母平均がこの値と仮説する値(帰無仮説で用いる)、$\hat{\sigma}$:標本標準偏差
- tは自由度 N −1のt分布に従う。

例：母分散が未知の場合（片側検定）

例1：あるクラスでは，数学の平均点をあげるべく補講の前後でテストを行った。補講前後の点数差は，1, −1, −2, 3, −1, 5, 4, 0, 7, −1であった。補講の効用を知りたい。有意水準 α=5 % で検定しよう。

考え方：補講で平均点が上がったと主張したいならば、これを対立仮説H1におく。よって、帰無仮説は上がっていない（μ=0）とする。これより片側検定となる。

H0：μ=0

H1：μ＞0

上記のデータから　t検定の検定量を手計算して検定が行える。ここではpythonを用いて検定を行う。

```
data = np.array([1, -1, -2, 3, -1, 5, 4, 0, 7, -1])
m = np.mean(data)
s = np.std(data, ddof=1)
N = len(data)
print(m,s,N)
```

ここに、m:平均値、s:標準偏差、N:データ数である。

STA_HypothesisTesting

numpyの標準偏差は、次の注意がある。

- ddof = 0 ならば、偏差平方和をデータの個数でそのまま割った標本標準偏差を求める
- ddof = 1ならば、不偏標準偏差を求めるので、この値とする

t_{5%}=1.833113

例：母分散が未知の場合（片側検定）

次に、t検定量と、有意水準α=5%のときのtalpを求める

```
alp = 0.05
talp = stats.t.ppf((1-alp),N-1)
print('talp (alpha=0.05, df=%d) =%f' %((N-1),talp))

m0 = 0 # null hypothesis
t = (m-m0)/(s/np.sqrt(N))
print('t=', t)
```

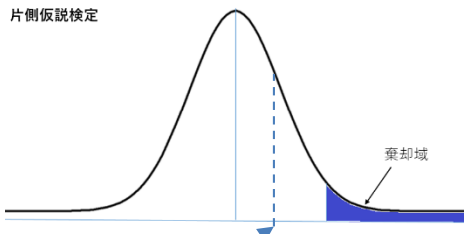
talp =1.833113
t= 1.5480470613460082

ココから

この結果から、図のようにα=5%のとき棄却域にtが入らないので、H0は棄却されない、ということになった。

ということは、どういうことか？　を考えましょう。

・平均点が同じであっても、バラツキが異なるとどうなるか？
　今回の例では m = 1.5
　他の値は、Notebookで確認します。



t =1.548
talp =1.833

例：母分散が未知の場合（両側検定）

例2：ある精密部品の直径の規格は1.54 cmである。製造したものから8個をサンプリングし、直径を測ると 1.5399, 1.5390, 1.5399, 1.5395, 1.5400, 1.5390, 1.5399, 1.5399であった。この部品は規格どおりであるか？　有意水準 α=5%で検定せよ

考え方：この例では仕様の値より大きくても小さくても不合格であるから両側検定を考える。よって、

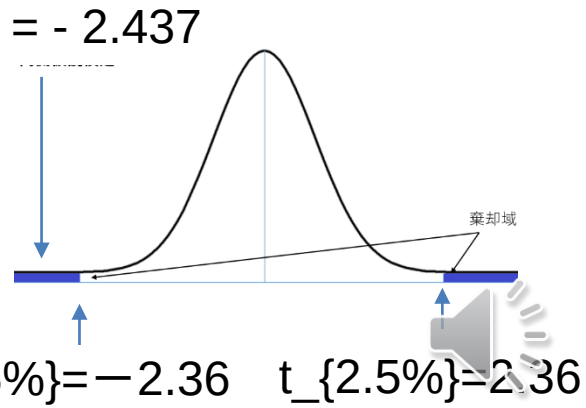
H0：μ= 1.54

H1：μ≠ 1.54

```
data2 = np.array([1.5399, 1.5390, 1.5399, 1.5395, 1.5400, 1.5390, 1.5399, 1.5399])
m = np.mean(data2)          # mean
s = np.std(data2, ddof=1) # std, ddof=1 : unbiased
N = len(data2)              #
df = N - 1                 # DoF (degree of freedom)
m0 = 1.54                   # H0 (null hypothesis)
# t =  -2.43730674672
# both side p =  0.0449361592238
```

STA_HypothesisTesting

これを見ながら，考える



p値が5%より小さい値より，H0は棄却できる。
よって，規格どおりでない。

平均値の差の検定（2標本）

- **2つの標本の母集団（いずれも正規分布とする）の母平均の差（$\mu 1 - \mu 2$）を検定することを考える。この場合，次の3通りが考えられる**
  - 2つの母分散 σ_1^2 , σ_2^2 が既知
  - 2つの母分散は未知であり，この2つは等しい
  - 2つの母分散は未知、この2つは等しいとは限らない ⇒ ウェルチのt検定（Welch's t test）
- **注意**
  - 厳密には，2標本が正規分布に従っているかを確かめる検定などが必要であるが，ここでは，正規分布に従っていると仮定する。
  - 2標本が従属であるか否かの検討も本来必要である（これを対応が有る，無いと言う場合もある）
  - 2つの母分散が共に未知，2標本が対応が無い場合だけを説明する

平均値の差の検定（2標本）

定式化

- 2標本　x = {x_1, …, x_n},　y = {y_1, …, y_m}（それぞれ，n個，m個の標本数）
- それぞれの，平均 $\widehat{\mu_x}$，$\widehat{\mu_y}$，分散$\widehat{\sigma_x^2}$, $\widehat{\sigma_y^2}$，　を求めたとする。
- 2つの平均の差の検定統計量は次式で表される。

$$t = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}}$$

ウェルチのt検定
（Welch's t test）

この自由度は，ウェルチ-サタスウェイトの式より
近似的に求められる。複雑ゆえWikipedia参照。

例：対応の無い２標本，分散が未知

2つの体温計の性能検定

- 二つの体温計（s1）と（s2）の測定精度に差があるかを確かめた10回の測定結果である。

s1 = np.array([37.1, 36.7, 36.6, 37.4, 36.8, 36.7, 36.9, 37.4, 36.6, 36.7])

s2 = np.array([36.8, 36.6, 36.5, 37.0, 36.7, 36.5, 36.6, 37.1, 36.4, 36.7])

- この二つの温度計は同じような体温測定精度を示すと言えるのか？
- このため，両方の平均値の差に有意差があるか否かの検定を行う。
- この問題は，　石村：統計解析の話，p.213を引用，ただし，本の分散値は間違っていることと，片側検定を行っている。
- 仮説
  - H0:　$\widehat{\mu_x} = \widehat{\mu_y}$
  - H1:　$\widehat{\mu_x} \neq \widehat{\mu_y}$　（両側検定）
- 関数　scipy.stats.ttest_ind
  - 2標本の平均値の差の検定を行う。ウェルチのt検定
  - equal_var = True: 2標本の分散が等しい，False:等しくない
- 考察
  - p値が11%強より，有意水準5%とすれば，H0は棄却できない。すなわち，二つの体温計の平均値が等しいという仮説は棄却できない。

STA_HypothesisTesting

```
s1 = np.array([37.1, 36.7, 36.6, 37.4, 36.8, 36.7, 36.9, 37.4, 36.6, 36.7])
s2 = np.array([36.8, 36.6, 36.5, 37.0, 36.7, 36.5, 36.6, 37.1, 36.4, 36.7])
t, p = scipy.stats.ttest_ind(s1, s2, equal_var = False)
print('t = ',t, ' p value = ',p)
# t =  1.66538214496  p value =  0.114776580923
```

まとめ

メソッド	説明	Rとの対応
ttest_1samp(x, m)	xの平均に対する1群のt検定	t.test(x,mu=m)
ttest_ind(x,y)	対応のない2群の平均の差に対するt検定(Welchのt検定？)	t.test(x,y,var.equal=TRUE)
ttest_rel	対応のある(等分散)2群の平均の差に対するt検定	
kstest(rvs, cdf)	コルモゴロフ-スミノフ検定。rvsとcdfのずれを検定。cdfは分布関数を指定？ 't':t分布。'norm':正規分布など	
ks_2samp(data1,data2)	2群のコルモゴロフ-スミノフ検定	

引用　http://seesaawiki.jp/met-python/d/%C5%FD%B7%D7%B2%F2%C0%CF

t検定の関数

scipy.stats　が提供するt検定の種類

メソッド	説明
ttest_1samp(x, m)	xの平均に対する1群のt検定
ttest_ind(x,y)	対応のない2群の平均の差に対するt検定(Welchのt検定？)
ttest_rel	対応のある(等分散)2群の平均の差に対するt検定
kstest(rvs, cdf)	コルモゴロフ-スミノフ検定。rvsとcdfのずれを検定。cdfは分布関数を指定？　't':t分布。'norm':正規分布など
ks_2samp(data1,data2)	2群のコルモゴロフ-スミノフ検定

- 引用: http://seesaawiki.jp/met-python/d/%C5%FD%B7%D7%B2%F2%C0%CF

おわりに

他の検定

- 分散、比率、相関係数などを対象とした検定
- 適合度、独立性などの検定
- ノンパラメトリック検定
  - 母集団の分布の型に関する情報なしに検定を行う方法
- これらは他の成書を参照してください

余話

- z, tの検定統計量の式を見て，特に，z検定統計量を見て
  - 検定の精度を高めるためには，サンプル数Nを多くとれ，と良く言われる。
  - しかし，式を見ると，Nが大きくなるとzの値は大きくなる。
  - したがって，棄却域に入りやすくなる。これは，どこか矛盾していないか？
- 母分散が未知の場合，t検定を用いると説明した。
  - 他書の中では，N＞30，または，N>100の場合では，標本分散を母分散とみなしてz検定を用いても良い，と書いてある場合がある。
  - N>30では，t分布がz分布に良く似通ってくるということと，コンピュータが発達していなかった昔は，z分布やt分布の値は表を参照していて煩雑だった，という理由からであった。
  - N>100の場合，もともと，データが正規分布に近くなるだろうという前提に基づく。
  - 現在は，コンピュータで簡単に計算できるのだから，N数に依らずにt検定を用いることを勧める。

$$z = \frac{\hat{\mu} - \mu_0}{\sigma / \sqrt{N}} \quad \sim \quad N(0, 1) \qquad t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma} / \sqrt{N}} \quad \sim t(N-1)$$

"ttest_ind" "ttest_rel" 違い

- 丁寧で詳しい，「対応があるか否か」など
  - https://stats.biopapyrus.jp/stats/t-test.html
- 宝くじの例
  - https://www.codereading.com/statistics/t-test.html

本講義は、ここまでです。
直ちに、課題に取り組んでください。