令和2年度「専修学校による地域産業中核的人材養成事業」

統計学II

令和2年度「専修学校による地域産業中核的人材養成事業」

統計学II

Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

目 次

AIシステム科　　シラバス　　20XX年度入学生

科目名	統計学Ⅱ					週合計駒数	1駒	作成日	
区分	必修 講義	開講時期	1年次 後期	週講義駒数 週実習等駒数	1駒 0駒	総時間数 総単位数	30時間 2単位	担当教員	

目標	概要
人工知能を学ぶ上で必要な基礎数学を習得するとともに、推測統計学の習得を目標とする。	本講義では、統計学および人工知能を学ぶ上で必須となる基礎数学を学習するとともに、データ分析や統計学的機械学習の基本知識となる推測統計学について学習する。

履修前提	テキスト・参考文献
※選択・エクステンションのみ記入	オリジナルテキスト

評価方法	関連科目
小テスト／中間テスト／期末テスト、提出課題、授業に取り組む姿勢（出席率、授業態度）	データマイニング、AIプログラミングⅠ・Ⅱ、機械学習Ⅰ・Ⅱ・Ⅲ、AIシステム開発

1	学習目標 微分の計算が出来る。偏微分の計算が出来る。関数の極値を求めることができる。	学習項目 微積分を中心とした解析学について学習する。ここでは、微分の定義から始め、練習問題を解くことにより、その幾何学的意味の理解および計算方法に慣れる。また、偏微分、停留点、極値、最大最小についても学習する。
	理解度確認：練習問題、小テスト	
2	学習目標 積分の計算が出来る。	学習項目 微積分を中心とした解析学について学習する。ここでは、積分の定義から始め、練習問題を解くことにより、その幾何学的意味の理解および計算方法に慣れる。
	理解度確認：練習問題、小テスト	
3	学習目標 確率変数とは何かを説明出来る。確率分布・確率密度関数について説明出来る。期待値・分散の計算が出来る。	学習項目 確率変数の定義と意味を理解し、確率分布、確率密度関数、期待値、分散について学習する。
	理解度確認：練習問題、小テスト	
4	学習目標 多次元の確率分布の特徴・性質について説明出来る。同時確率密度関数と周辺確率密度関数について説明出来る。	学習項目 多次元における確率分布およびその関連事項である同時確率密度関数、周辺確率密度関数、確率変数の独立性について学習する。
	理解度確認：練習問題、小テスト	
5	学習目標 大数の法則について説明出来る。	学習項目 大数の法則について学習する。併せて、大数の法則のコンピュータシミュレーションも行う。
	理解度確認：練習問題、小テスト	
6	学習目標 中心極限定理について説明出来る。	学習項目 中心極限定理について学習する。併せて、中心極限定理のコンピュータシミュレーションも行う。
	理解度確認：練習問題、小テスト	
7	学習目標 代表的なサンプリングについて説明出来る。母集団とサンプリングされた集団の統計量の計算およびそれらの関係を説明出来る。	学習項目 推測統計学と記述統計学の違い、推測統計学の歴史と発展を踏まえた上で、母集団、標本とサンプリング、統計量（母平均、母分散、標本平均、標本分散、不偏分散）について学習する。
	理解度確認：練習問題、小テスト	
8	学習目標 母分散既知／未知の場合の標本平均の標本分布について説明出来る。Studentのt分布について説明出来る。	学習項目 Gauss分布からのサンプリングについて学習する。具体的には、母分散既知／未知の場合の標本平均の標本分布、Studentのt分布について学習する。
	理解度確認：練習問題、小テスト	
9	学習目標 これまでに学習した内容を復習し、理解を確実なものにする。	学習項目 これまでの学習内容の総復習を実施する。
	理解度確認：練習問題、小テスト	
10	学習目標 推測統計学における推定とは何かを説明出来る。点推定が出来る。最尤法が出来る。	学習項目 点推定と区間推定の違いを概説した上で、点推定について学習する。具体的には、点推定の方法と基準、推定量と推定値、最尤法について学習する。
	理解度確認：練習問題、小テスト	
11	学習目標 区間推定が出来る。	学習項目 区間推定について学習する。具体的には、信頼区間、信頼係数、正規母集団の母平均、母分散の区間推定について学習する。
	理解度確認：練習問題、小テスト	

12	学習目標 仮説検定とは何かを説明出来る。統計的仮説の有意性、有意水準、帰無仮説と対立仮説、第一種・第二種の誤りについて説明出来る。	学習項目 仮説検定について学習する。具体的には、統計的仮説の有意性、仮説の棄却、有意水準、帰無仮説と対立仮説、第一種・第二種の誤りについて学習する。
	理解度確認： 練習問題、小テスト	
13	学習目標 検定統計量と棄却域、採択域について説明ができる。	学習項目 仮説検定における検定統計量、棄却域、採択域、両側・片側検定について学習する。
	理解度確認： 練習問題、小テスト	
14	学習目標 Studentのt検定とは何かを説明出来る。	学習項目 Studentのt検定について学習する。
	理解度確認： 練習問題、小テスト	
15	学習目標 これまでに学習した内容を復習し、理解を確実なものにする。	学習項目 これまでの学習内容の総復習を実施する。
	理解度確認： 確認テスト	

第1回：微分

アジェンダ

- 微分とは
- 偏微分
- 停留点、極値、最大値最小値

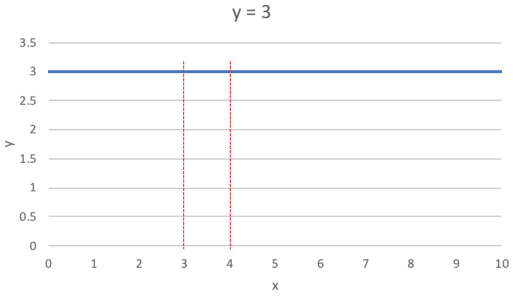
微分とは

- 微分とは、任意の関数の各点における変化の割合（傾き）を求めることです。
- 傾きは以下の式で定義されます。
  - 変化の割合（傾き）$= \frac{\text{yの増加量}}{\text{xの増加量}}$
- 上式の変化の割合のことを、導関数といいます。

微分の例：傾きがない関数

- 下図のような「y = 3」という関数を考えます。
- この関数ではxの値によらずyの値は3のため、xの変化に対するyの変化の割合は0となります。

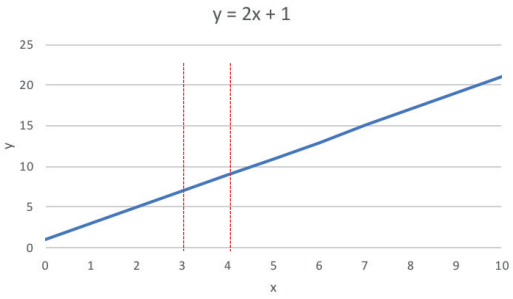
- 例えばxが3から4に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0}{4-3} = 0$



微分の例：1次関数

- 下図のような「y = 2x + 1」という関数を考えます。
- この関数ではxの値が1増加すると、yの値は2増加します。

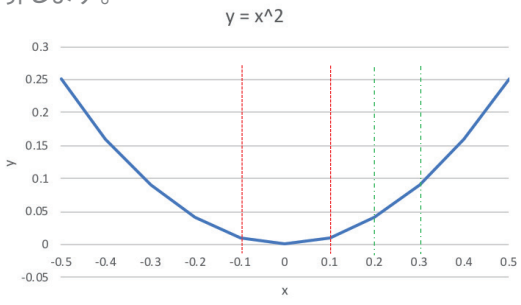
- 例えばxが3から4に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{9-7}{4-3} = 2$



微分の例：2次関数

- 下図のような「$y = x^2$」という関数を考えます。
- この関数ではxの増加量に対するyの増加量は、xの場所によって異なります。

- 例えばxが−0.1から0.1に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0.01-0.01}{0.1-(-0.1)} = 0$
- 次に、xが0.2から0.3に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0.09-0.04}{0.3-0.2} = 0.5$



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

微分の定義

微分は、以下の式のように変化の割合だと述べました。

$$\text{変化の割合（傾き）} = \frac{\text{yの増加量}}{\text{xの増加量}}$$

また、変化の割合はxの場所によって変化することも学習しました。

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

微分の定義

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

$f(x)=x^2$のとき、上式にしたがって微分を実施すると、以下のようになります。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}=\lim_{h\to 0}\frac{(x+h)^2-x^2}{h}=\lim_{h\to 0}\frac{x^2+2hx+h^2-x^2}{h}=\lim_{h\to 0}(2x+h)=2x$$

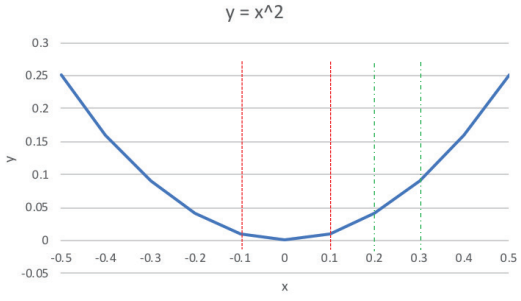
ここで求めた導関数を基に計算すると、

x=0.2のとき0.4

x=0.3のとき0.6

となります。「微分の例：2次関数」ではxが0.2から0.3に増えるときの傾きを計算し0.5となりました。

これは上の数字のちょうど真ん中にあることが確認できます。



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

可微分性

関数$f(x)$の任意の場所aにおける微分は、以下の式で表せました。

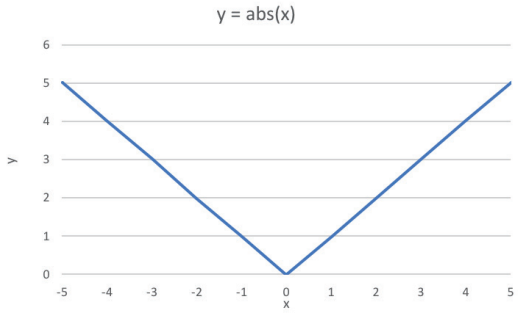
$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

絶対値関数$f(x)=|x|$の、x=0における傾きは

h＞0のときは1

h＜0のときは−1

となり、x=0で微分可能ではありません。



偏微分

偏微分とは

偏微分とは，変数が複数ある関数の「特定の変数以外は定数だとみなして」微分することです。
2変数（X, Y）の場合、偏微分は以下のように表されます。

$\frac{\partial f(x,y)}{\partial x}, f_x$: xによる偏微分
$\frac{\partial f(x,y)}{\partial y}, f_y$: yによる偏微分

偏微分の計算例

「$f(x,y) = x^2 + y^3 + 5y + xy$」についての偏微分を取り扱います。

上式をxで偏微分します（xのみ変数として扱い、yは定数とみなします）。

$$\frac{\partial f(x,y)}{\partial x} = 2x + y$$

上式をyで偏微分します（yのみ変数として扱い、xは定数とみなします）。

$$\frac{\partial f(x,y)}{\partial y} = 3y^2 + 5 + x$$

停留点・極値・最小値最大値

停留点とは

2変数の関数について、

$$\frac{\partial f}{\partial x}(a,b)=\frac{\partial f}{\partial y}(a,b)=0$$

を満たす点

$$(x,y)=(a,b)$$

を停留点、または臨界点といいます。

極値とは

関数$f(x,y)$が点(a,b)で極大値$f(a,b)$をとるとは、
(a,b)の近くの(x,y)では常に$f(a,b) > f(x,y)$
が成り立つことをいいます。

関数$f(x,y)$が点(a,b)で極小値$f(a,b)$をとるとは、
(a,b)の近くの(x,y)では常に$f(a,b) < f(x,y)$
が成り立つことをいいます。

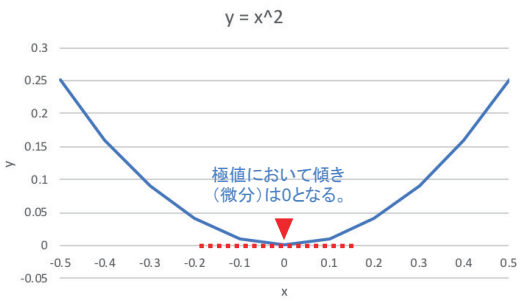
極大値と極小値のことを極値といいます。

停留点と極値

$f(x,y)$が微分可能な関数で、(a,b)が極値ならば、(a,b)は停留点となります。

$$\frac{\partial f}{\partial x}(a,b) = \frac{\partial f}{\partial y}(a,b) = 0$$

ただし、その逆は必ず成り立つとは言えません。



停留値と極値

極値とならない停留点もあります。例えば

$$f(x,y) = x^2 + y$$

において

$$(x,y) = (0,0)$$

は停留点です。

ただし、$f(0,y)$はyが正なら正の値、yが負なら負の値を取りますので、$(0,0)$の周辺には$f(0,0) = 0$より大きい点も小さい点もあります。よって、

$$(x,y) = (0,0)$$

は停留点ですが、極大でも極小でもありません。

極値と最大値/最小値

極値が最大値と成り得る条件を考えます。

例えば右下の図にある関数について

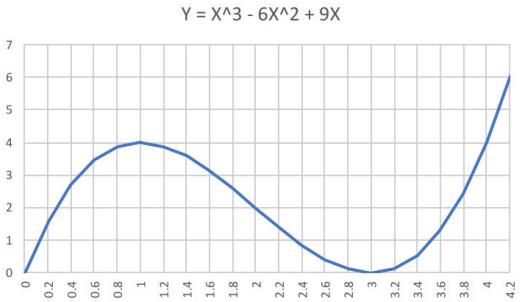
$$0 \leq x < 4$$

の範囲においては、$x = 1$のときに$y = 4$となるため、極値が最大値となります。

しかし、

$$0 \leq x \leq 4.2$$

の範囲においては、$x = 4.2$のときに$y = 6.048$となるため、極値が最大値とはなりません。



演習問題

演習1：定数関数の微分

- $f(x) = 4$となる定数関数の導関数を求めてください。

演習2：1次関数の微分

- $y = 5x + 3$の導関数を求めてください。

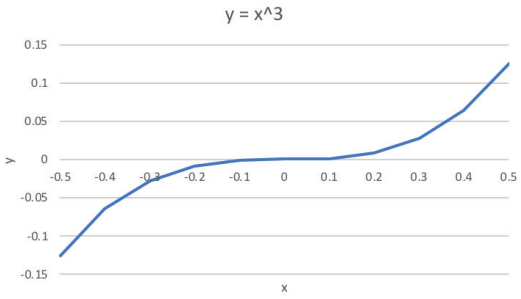
演習3：2次関数の微分

- $y = 2x^2$の導関数を求めてください。
- 求めた導関数より、$x=1$と$x=2$の傾きをそれぞれ求めてください。

演習4：3次関数の微分

- y = x3の導関数を求めてください。
- 求めた導関数より、x=0とx=0.3の傾きをそれぞれ求めてください。

x	y
-0.5	-0.125
-0.4	-0.064
-0.3	-0.027
-0.2	-0.008
-0.1	-0.001
0	0
0.1	0.001
0.2	0.008
0.3	0.027
0.4	0.064
0.5	0.125



演習5：偏微分

- 「$f(x,y) = x^2 + 2y^3 + 4y + xy$」について、xで偏微分を実施してください。
- 上式に対し、yで偏微分を実施してください。

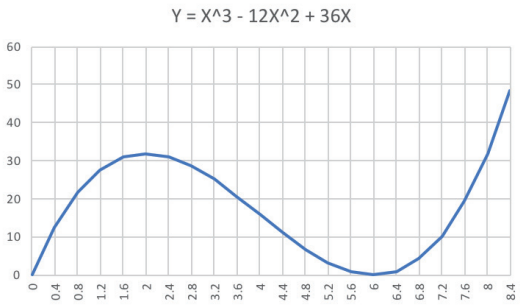
演習6：停留点、極値、最大値最小値

右下の図の関数について

$$0 \leq x < 8$$

の範囲内で停留点と極値を求めてください。

また、上記の範囲内で極大値は最大値となるのか、確認してください。



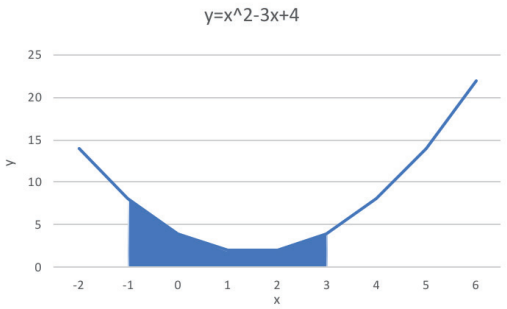
第2回：積分

アジェンダ

- 積分とは
- 積分の例
- 複雑な関数の定義
- 微分と積分の関係

積分とは

- 積分とは、任意の関数f(x)で囲まれた部分の面積を求めることを意味しています。
  - $\int_a^b f(x)dx$
- 例えば$f(x) = x^2 - 3x + 4$、a=−1、b=3の場合、下図の青い部分の面積を求めることができます。

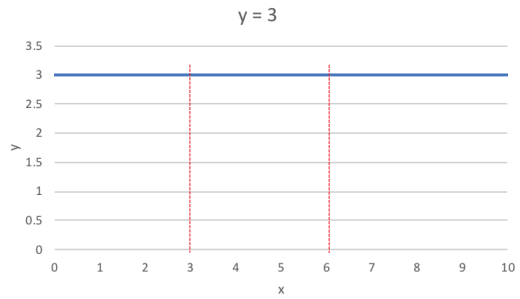


積分の例：傾きがない関数

- 下図のような「y = 3」という関数を考えます。
- この関数ではxの値によらずyの値は3のため、長方形の面積を求めることと同じになります。

- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 3dx = [3x]_3^6 = 3 \times 6 - 3 \times 3 = 9$

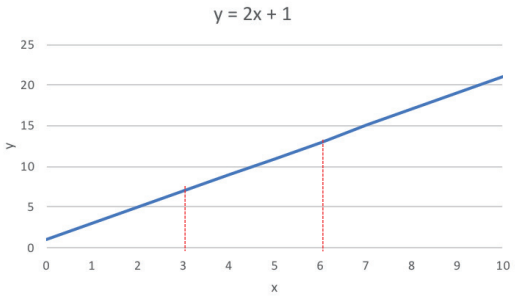
ここでは「3」を導関数とする原始関数「3x」を求めています。



積分の例：1次関数

- 下図のような「y = 2x + 1」という関数を考えます。
- この関数ではxの値が1増加すると、yの値は2増加します。

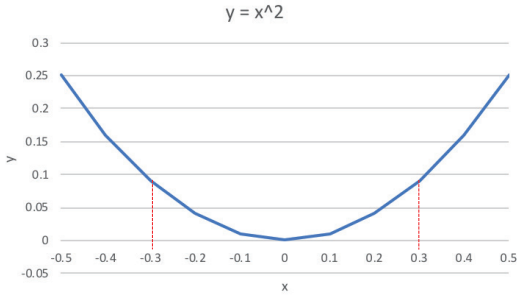
- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 (2x+1)dx = [x^2+x]_3^6 = (6\times 6+6)-(3\times 3+3) = 42-12 = 30$



積分の例：2次関数

- 下図のような「$y = x^2$」という関数を考えます。
- この関数ではxの増加量に対するyの増加量は、xの場所によって異なります。

- 例えばxが−0.3から0.3の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_{-0.3}^{0.3} x^2 dx = \left[\frac{1}{3}x^3\right]_{-0.3}^{0.3} = \left(\frac{0.3\times 0.3\times 0.3}{3}\right) - \left(\frac{(-0.3)\times(-0.3)\times(-0.3)}{3}\right) = 0.009+0.009 = 0.018$



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

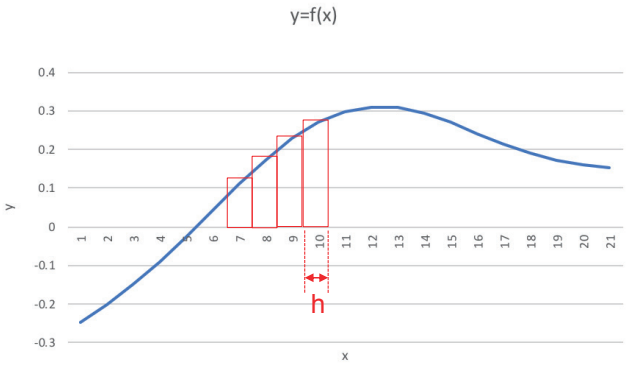
複雑な関数の積分

これまでの例では、導関数の原始関数を解析的に求めることができました。

解析的に求めることができない関数に対して面積を算出する際は、コンピュータプログラムなどで以下のようにして面積の近似値を求めます。

$$\sum_{i=a}^{b} f(i)h$$

hの間隔を徐々に狭くしていけば、上式の値は真の値に近づいていきます。



微分と積分の関係

「複雑な関数の積分」のページでは、面積の近似値をプログラムで求める方法を記載しましたが、より厳密に面積を求めていきます。

下図のオレンジ色の面積は、青い部分より大きく、青＋赤より小さいことがわかります。

$$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$$

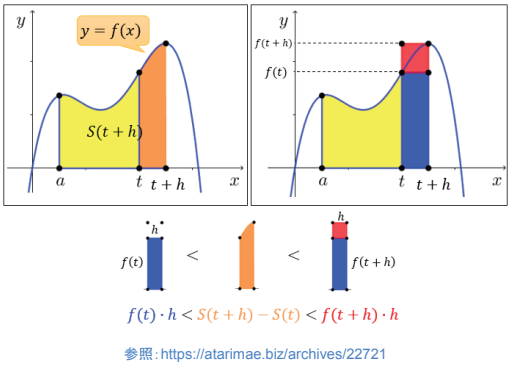
$$f(t) < \frac{S(t+h)-S(t)}{h} < f(t+h)$$

hの極限をとると、

$$f(t) < \lim_{h\to 0}\frac{S(t+h)-S(t)}{h} < \lim_{h\to 0} f(t+h) = f(t)$$

$$\lim_{h\to 0}\frac{S(t+h)-S(t)}{h} = f(t)$$

最後の式は、f(x)の導関数を求める式と同じであることが確認できます。



$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$

参照：https://atarimae.biz/archives/22721

演習問題

演習1：定数関数の積分

- $f(x) = 4$となる定数関数の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

演習2：1次関数の積分

- y = 5x + 3の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

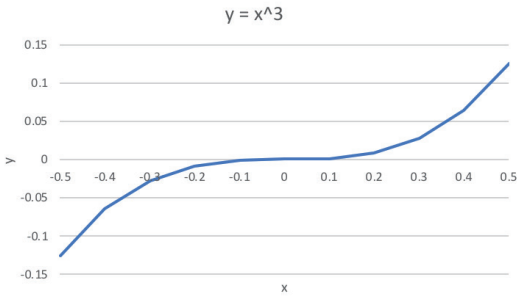
演習3：2次関数の積分

- $y = 2x^2$の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

演習4：3次関数の積分

- y = x3の原始関数を求めてください。
- −0.4<=x<=0.4の範囲で面積を求めてください。

x	y
-0.5	-0.125
-0.4	-0.064
-0.3	-0.027
-0.2	-0.008
-0.1	-0.001
0	0
0.1	0.001
0.2	0.008
0.3	0.027
0.4	0.064
0.5	0.125



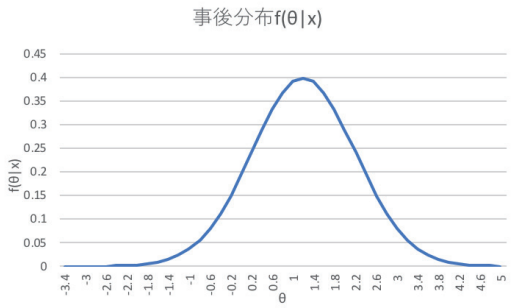
演習問題：機械学習への応用

積分と機械学習

ベイズ推定において下図のように事後分布が得られたとします。
事後分布の解釈を行うために、点推定量としてEAP（Expected a Posterioi：事後期待値）推定量というものを計算することがあります。

$$EAP\text{推定量} = \int f(\theta|x) \cdot \theta \, d\theta$$

本演習では、右表のデータを基にEAP推定量を計算します。



θ	f(θ|x)
-3.4	1.0141E-05
-3.2	2.4942E-05
-3	5.8943E-05
-2.8	0.00013383
-2.6	0.00029195
-2.4	0.0006119
-2.2	0.00123222
-2	0.00238409
-1.8	0.00443185
-1.6	0.00791545
-1.4	0.01358297
-1.2	0.02239453
-1	0.03547459
-0.8	0.05399097
-0.6	0.07895016
-0.4	0.11092083
-0.2	0.14972747
0	0.19418605
0.2	0.24197072
0.4	0.28969155
0.6	0.3332246

θ	f(θ|x)
0.8	0.36827014
1	0.39104269
1.2	0.39894228
1.4	0.39104269
1.6	0.36827014
1.8	0.3332246
2	0.28969155
2.2	0.24197072
2.4	0.19418605
2.6	0.14972747
2.8	0.11092083
3	0.07895016
3.2	0.05399097
3.4	0.03547459
3.6	0.02239453
3.8	0.01358297
4	0.00791545
4.2	0.00443185
4.4	0.00238409
4.6	0.00123222
4.8	0.0006119
5	0.00029195

演習5：事後分布とパラメータθの積

EAP推定量を求めるため、$f(\theta|x) \cdot \theta$を各データについて計算してください。

$$EAP\text{推定量} = \int f(\theta|x) \cdot \theta \, d\theta$$

θ	f(θ|x)	f(θ|x)*θ
-3.4	1.0141E-05	-3.44789E-05
-3.2	2.4942E-05	-7.98159E-05
-3	5.8943E-05	-0.000176829
-2.8	0.00013383	-0.000374725
-2.6	0.00029195	-0.000759062
-2.4	0.0006119	-0.001468565
-2.2	0.00123222	-0.002710882
-2	0.00238409	-0.004768176
-1.8	0.00443185	-0.007977327
-1.6	0.00791545	-0.012664723
-1.4	0.01358297	-0.019016157
-1.2	0.02239453	-0.026873436
-1	0.03547459	-0.035474593
-0.8	0.05399097	-0.043192773
-0.6	0.07895016	-0.047370095
-0.4	0.11092083	-0.044368334
-0.2	0.14972747	-0.029945493
0	0.19418605	0
0.2	0.24197072	0.048394145
0.4	0.28969155	0.115876621
0.6	0.3332246	0.199934762

演習6：事後分布とパラメータθとdθの積

EAP推定量を求めるために計算した$f(\theta|x)\cdot\theta$に、$d\theta$（右表の場合は0.2刻みなので0.2）をかけてください。

$$EAP\text{推定量} = \int f(\theta|x)\cdot\theta\ d\theta$$

θ	f(θ|x)	f(θ|x)*θ	f(θ|x)*θ*⊿θ
-3.4	1.0141E-05	-3.44789E-05	-6.89578E-06
-3.2	2.4942E-05	-7.98159E-05	-1.59632E-05
-3	5.8943E-05	-0.000176829	-3.53658E-05
-2.8	0.00013383	-0.000374725	-7.49449E-05
-2.6	0.00029195	-0.000759062	-0.000151812
-2.4	0.0006119	-0.001468565	-0.000293713
-2.2	0.00123222	-0.002710882	-0.000542176
-2	0.00238409	-0.004768176	-0.000953635
-1.8	0.00443185	-0.007977327	-0.001595465
-1.6	0.00791545	-0.012664723	-0.002532945
-1.4	0.01358297	-0.019016157	-0.003803231
-1.2	0.02239453	-0.026873436	-0.005374687
-1	0.03547459	-0.035474593	-0.007094919
-0.8	0.05399097	-0.043192773	-0.008638555
-0.6	0.07895016	-0.047370095	-0.009474019
-0.4	0.11092083	-0.044368334	-0.008873667
-0.2	0.14972747	-0.029945493	-0.005989099
0	0.19418605	0	0
0.2	0.24197072	0.048394145	0.009678829
0.4	0.28969155	0.115876621	0.023175324
0.6	0.3332246	0.199934762	0.039986952

演習7：事後分布とパラメータθの積の積分

EAP推定量を求めるために計算した$f(\theta|x)\cdot\theta d\theta$全データについて合計（積分）してEAP推定量を計算してください。

$$EAP\text{推定量} = \int f(\theta|x)\cdot\theta\ d\theta$$

※EAP推定量は1.2に近い値となります。

θ	f(θ|x)	f(θ|x)*θ	f(θ|x)*θ*⊿θ
-3.4	1.0141E-05	-3.44789E-05	-6.89578E-06
-3.2	2.4942E-05	-7.98159E-05	-1.59632E-05
-3	5.8943E-05	-0.000176829	-3.53658E-05
-2.8	0.00013383	-0.000374725	-7.49449E-05
-2.6	0.00029195	-0.000759062	-0.000151812
-2.4	0.0006119	-0.001468565	-0.000293713
-2.2	0.00123222	-0.002710882	-0.000542176
-2	0.00238409	-0.004768176	-0.000953635
-1.8	0.00443185	-0.007977327	-0.001595465
-1.6	0.00791545	-0.012664723	-0.002532945
-1.4	0.01358297	-0.019016157	-0.003803231
-1.2	0.02239453	-0.026873436	-0.005374687
-1	0.03547459	-0.035474593	-0.007094919
-0.8	0.05399097	-0.043192773	-0.008638555
-0.6	0.07895016	-0.047370095	-0.009474019
-0.4	0.11092083	-0.044368334	-0.008873667
-0.2	0.14972747	-0.029945493	-0.005989099
0	0.19418605	0	0
0.2	0.24197072	0.048394145	0.009678829
0.4	0.28969155	0.115876621	0.023175324
0.6	0.3332246	0.199934762	0.039986952

第3回：確率変数

アジェンダ

- 確率変数
- 確率分布
  - 離散型確率分布
  - 連続型確率分布
- 期待値
- 分散

確率変数

確率変数とは

ある現象がいろいろな値を取り得るとき、取り得る値全体を確率変数といいます。

例えば、サイコロを振ったときに出る目は[1, 2, 3, 4, 5, 6]のいずれかとなります。
この場合、確率変数Xは

$$\mathrm{X} = 1, 2, 3, 4, 5, 6$$

と表します。

確率変数をXと置くことで、サイコロの目が取りうる値の確率を、以下のように記載することができます。

$$\mathrm{P}(x) = \frac{1}{6}(\mathrm{X} = 1, 2, 3, 4, 5, 6)$$

サイコロを振って4が出る確率は以下のように書きます。

$$\mathrm{P}(x = 4) = \frac{1}{6}$$

離散型の確率変数

離散型確率変数は、「とびとびの値」を指します。
隣り合った数値の間には、数値は存在しません。
例えばサイコロの目、コインの裏表、ルーレットの番号などが該当します。

連続型の確率変数

連続型確率変数は、「連続した値」を指します。
例えば速度であれば、5km/hと6km/hの間には5.1km/hや5.01km/h、5.0001km/hなど無数の値が存在します。

その他の連続確率変数には温度、湿度、高度、体重などがあります。

確率分布

確率分布とは

確率変数のそれぞれの値に対し、その確率変数をとる確率の分布のことです。

離散型確率変数に対する確率分布として、以下のような確率分布があります。

- ポアソン分布
- 二項分布
- 幾何分布
- 一様分布

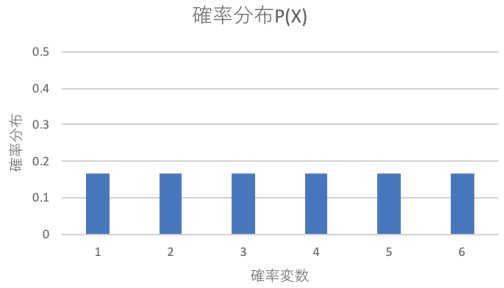
連続型確率変数に対する確率分布として、以下のような確率分布があります。

- 正規分布
- 指数分布
- 一様分布

離散型確率分布

確率変数が離散型の場合の確率分布を、離散型確率分布といいます。
サイコロの例だと、以下のようになります。
確率変数Xを横軸、Xが起きる確率をP(X)とすると、値は全て1/6となります。

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6



離散型確率分布と確率質量関数

確率変数Xが離散型の場合の確率分布P(X)を、離散型確率分布といいました。
確率分布P(x)を関数f(x)で表現した場合、f(x)を「確率質量関数」といいます。

サイコロの例だと、以下のようになります。

$$\sum_{i=1}^{6} P(X = x_i) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

連続型確率分布

確率変数が連続型の場合の確率分布を、連続型確率分布といいます。

サイコロの例は離散型確率変数のため、値は全て1/6となりました。
仮に、1～6までの連続確率変数を考えてみます
1～6の間には、1.1、1.01、1.001、、、と無数の数が存在します。

サイコロの目のように6つの値を持つ離散型確率変数であれば、目が1になる確率は

$$P(X=1)=\frac{1}{6}$$

となります。ですが、連続型確率変数であれば値は無数にあるため、

$$P(X=1)=\frac{1}{\infty}=0$$

となります。

連続型確率分布と確率密度関数

確率変数Xが連続型の場合の確率分布P(X)を、連続型確率分布といいました。
前出の様に、確率変数が連続型の場合には、確率変数が特定の値をとる確率は0になることから、縦軸は確率ではなく「確率密度」という考え方を使います。
確率密度は、連続型確率変数が取りうる範囲内での、特定の値の「相対的な出やすさ」を表しています。

連続型確率分布P(x)を関数f(x)で表現した場合、f(x)を「確率密度関数」といいます。

連続型確率分布と確率密度関数

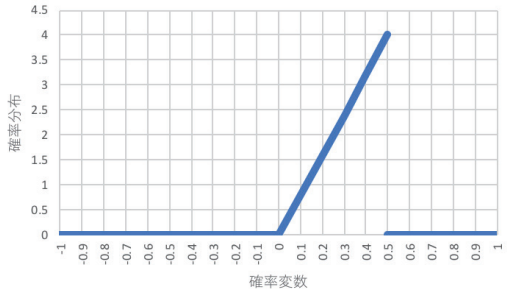
例えば、以下のような確率変数Xが$0 \leq X \leq 0.5$の範囲で8xとなる確率密度関数を考えます。

$$f(x) = \begin{cases} 8x(0 \leq X \leq 0.5) \\ 0(X < 0, X > 0.5) \end{cases}$$

確率変数Xが0.1、0.4のときの値は下記のようになります。

$$f(0.1) = 0.8$$
$$f(0.4) = 3.2$$

この確率密度関数です、X=0.4の状態は、X=0.1の状態より4倍起こりやすいということが言えます。



期待値

期待値とは

期待値とは、1回の試行で得られる値の平均値のことです。
得られうるすべての値（すべての確率変数）とそれが起こる確率の積を足し合わせて計算できます。

離散型確率変数の期待値

離散型確率変数の期待値は、確率変数がとり得る値に対応する確率を掛け、掛けた結果を全て足します。

$$\mathrm{E}(X) = \sum_{i=1}^{n}(x_i \cdot P_i)$$

サイコロの例だと、期待値は以下のようになります。

$$\mathrm{E}(X) = \sum_{i=1}^{n}(x_i \cdot P_i)$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6
X・P(X)	1/6	1/3	1/2	2/3	5/6	1
					ΣX・P(X)	3.5

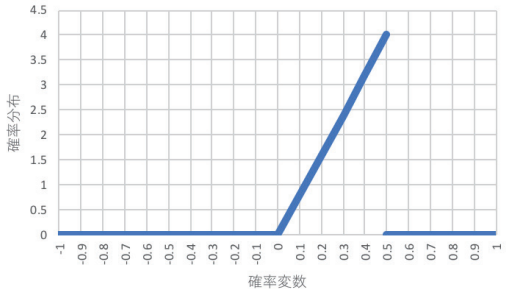
連続型確率変数の期待値

連続型確率変数の期待値は、積分によって計算します。

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

例えば確率密度関数f(x)=8x、確率変数Xが0から0.5の値を取る場合は、以下のようになります。

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$
$$= \int_{0}^{0.5} x \cdot 8xdx$$
$$= \left[\frac{8}{3}x^3\right]_0^{0.5}$$
$$= \frac{1}{3}$$



分散

分散とは

分散とは、「[確率変数の全ての値と期待値（平均値）の差]の2乗」と「確率」との積を、全て足し合わせたものです。
分散は英語でVarianceと表記するので、頭文字を使って$V(X)$と表記します。

離散型確率変数の分散

離散型確率変数の分散は、以下の式で表されます。

$$V(X) = \sum_{i=1}^{n}\left((x_i - \mu)^2 \cdot P_i\right)$$

ここで、Xの期待値$E[X] = \mu$とします。

サイコロの例だと、分散は以下のようになります。

$$V(X) = \sum_{i=1}^{n}\left((x_i - \mu)^2 \cdot P_i\right)$$

$$= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} + (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} = \frac{35}{12}$$

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6
X・P(X)	1/6	1/3	1/2	2/3	5/6	1
					ΣX・P(X)	3.5

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6
$(X-\mu)^2$	6.25	2.25	0.25	0.25	2.25	6.25
$(X-\mu)^2$・P(X)	6.25/6	2.25/6	0.25/6	0.25/6	2.25/6	6.25/6
					Σ((X-μ)2・P(X))	35/12

連続型確率変数の分散

連続型確率変数の分散は、積分によって計算します。

$$V(X) = \int_{-\infty}^{\infty}(x-\mu)^2 \cdot f(x)dx$$

例えば確率密度関数f(x)=1/6、確率変数Xが0から6の値を取る場合は、以下のようになります。
※期待値E[X]は3であることを利用します。

$$V(X) = \int_{-\infty}^{\infty}(x-\mu)^2 \cdot f(x)dx$$
$$= \int_{0}^{6}(x-3)^2 \cdot \frac{1}{6}dx$$
$$= \left[\frac{1}{3}(x-3)^3 \cdot \frac{1}{6}\right]_0^6$$
$$= \frac{(6-3)^3}{18} - \frac{(-3)^3}{18}$$
$$= \frac{27+27}{18}$$
$$= 3$$

演習問題

演習1：離散型の確率変数

- 離散型確率変数の例を挙げてください。
- 確率変数と確率分布の表を作成してください。
- 期待値を求めてください。
- 分散を求めてください。

演習2：連続型の確率変数

- 連続型確率変数の例を挙げてください。
- 確率変数と確率分布の表を作成してください。
- 期待値を求めてください。
- 分散を求めてください。

第4回：多次元の確率分布

第4回の目的

- 確率変数が2つ以上ある場合に、それぞれの確率変数がとる値とその確率の分布を「同時確率分布」といいます。
- 確率変数が離散型の場合には「離散型同時確率分布」といい、確率変数が連続型の場合には「連続型同時確率分布」といいます。

- 第4回講義においては確率変数が2つの場合の同時確率分布について学習します。

アジェンダ

- 離散型同時確率分布
- 連続型同時確率分布
- 確率変数の独立性
- 2変数のガウス分布

離散型同時確率分布

離散型同時確率分布とは

2つの離散型確率変数XとYが、それぞれある値をとるときの確率を表したものを「離散型同時確率分布」といいます。

例えば、男子20名、女子20名のあるクラスがあるとします。生徒の居住地区を表にしてみました。
性別をX、居住地区をYとすると、2つの離散型確率変数とみなせます。

	A地区	B地区	C地区	D地区	計
男子	4	6	6	4	20
女子	6	8	4	2	20
				総合計	40

全生徒40人に対する各マスの数字の割合を計算してみました。これは2つの離散型確率変数XとYがそれぞれの値を同時にとる、離散型同時確率分布となります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

離散型同時確率分布とは

2つの確率変数からなる同時確率分布は、以下のように表記します。

$$f(x_i, y_j) = P(X = x_i, Y = y_j) \quad i = 1,2,3,,,;j = 1,2,3,,,$$

例えば、男子でD地区に住む生徒の確率は、以下のようになります。

$$P(X = \text{男子}, Y = \text{D地区}) = 0.1$$

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

ここで$f(x_i, y_j)$のことを同時確率関数といいます。各iとjについて全ての確率を足すと総和は1になります。

$$\sum_i \sum_j f(x_i, y_j) = 1$$

周辺確率分布

性別X、居住地区Yのそれぞれの値について、確率の合計を計算してみます。
男子の割合は0.5、A地区に居住する生徒の割合は0.25であることがわかります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5
計	0.25	0.35	0.25	0.15	1.00

このようにある1つの確率変数を固定し、別の確率変数が取りうる全ての確率を合計したものを周辺確率分布といいます。

$$f_x(x_i) = \sum_j f(x_i, y_j) = P(X = x_i) \quad i = 1,2,3,,,$$
$$f_y(y_j) = \sum_i f(x_i, y_j) = P(Y = y_j) \quad j = 1,2,3,,,$$

ここで、$f_x(x_i)$と$f_y(y_j)$をそれぞれXとYの周辺確率関数といいます。

連続型同時確率分布

連続型同時確率分布とは

XとYが連続型確率変数であるとき、それぞれある値をとるときの確率を表したものを「連続型同時確率分布」といいます。

XとYの同時確率分布を表す関数を「同時確率密度関数」といいます。

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$$

確率の総和は1になるため、同時確率密度関数に関して以下の式が成り立ちます。

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

連続型確率変数XとYの周辺確率密度関数は、以下の式で求めることができます。

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y) = \begin{cases} x+y \ (0 \leq x \leq 1, 0 \leq y \leq 1) \\ 0 \ (\text{上記以外の} x,y \text{の場合}) \end{cases}$$

同時確率変数XとYの全範囲についての確率を求めてみます。

$$P(0 \leq x \leq 1, 0 \leq y \leq 1) = \int_0^1 \int_0^1 (x+y)\,dx\,dy = \int_0^1 \left[\frac{x^2}{2} + yx\right]_0^1 dy = \int_0^1 \left(\frac{1}{2} + y\right) dy$$

$$= \left[\frac{y}{2} + \frac{y^2}{2}\right]_0^1 = 1$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Xの周辺確率密度関数を求めてみます。

$$f_x(x)=\int_0^1 (x+y)\,dy=\left[x+\frac{y^2}{2}\right]_0^1=x+\frac{1}{2}$$

確率変数の独立性

独立な確率変数とは

2つの確率変数XとYの同時確率分布（同時確率密度関数）f(x, y)が、それぞれの確率変数の周辺確率分布（周辺確率密度関数）g(x)とh(y)の積に分解できる時、その2つの確率変数は独立(independent)であると言います。

$$f(x,y) = g(x)h(y)$$

直感的な理解としては、「XとYの動きは、お互いに影響を及ぼさない」ということです。

共分散とは

共分散とは、二組の対応するデータの関係を表す指標です。

例えば、各地区における男女別の生徒数の関係を考えてみます。
共分散を参照すると、「男子生徒の人数が多い地区は、女子生徒の人数も多いのか？」などの傾向を分析することができます。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

共分散の計算

共分散の定義は「[Xの偏差 × Yの偏差]の平均」です。
また、偏差とは「平均との差」のことです。

男子生徒の各地区の平均人数は35人、女子生徒は21.5人です。
A地区における男子生徒の偏差と女子生徒の偏差は、それぞれ

10 − 35 = −25 : 男子生徒の偏差
6 − 21.5 = −15.5 : 女子生徒の偏差

各地区について男女の偏差を計算し、それらの
平均を取ります。

(387.5 + 462.6 + 127.5 + 172.5) / 4 = 287.5

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

偏差

	A地区	B地区	C地区	D地区	平均
男子	-25	25	15	-15	
女子	-15.5	18.5	8.5	-11.5	
地区ごと偏差の積	387.5	462.6	127.5	172.5	287.5

共分散の意味

共分散の定義は「[Xの偏差 × Yの偏差]の平均」ですので、

共分散が大きい　->Xが大きいとYも大きい傾向がある。
共分散が0付近　->XとYにあまり関係はない。
共分散が小さい　->Xが大きいとYは小さくなる傾向がある。

今回の例では共分散が287.5ですので、「男子生徒が多い地区は女子生徒も多い傾向がある」といえます。

共分散は「Covariance」と言いますので、XとYの
共分散のことを

$Cov(X,Y)$

と書くことがあります。もしくは

σ_{XY}

と書くこともあります。また、期待値の記号では

$E[(X-\mu_X)(Y-\mu_Y)]$

と書きます。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

偏差

	A地区	B地区	C地区	D地区	平均
男子	-25	25	15	-15	
女子	-15.5	18.5	8.5	-11.5	
地区ごと偏差の積	387.5	462.6	127.5	172.5	287.5

共分散の簡単な計算方法

分散の式を展開します。

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y]$$

「和の期待値」は「期待値の和」ですので、以下のようになります。

$$\mathrm{Cov}(X,Y) = E[XY] - E[X\mu_Y] - E[Y\mu_X] + E[\mu_X\mu_Y]$$

定数倍は期待値の外に出して、

$$\mathrm{Cov}(X,Y) = E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y$$

$E[X] = \mu_X$、$E[Y] = \mu_Y$なので、

$$\mathrm{Cov}(X,Y) = E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y$$

となります。

共分散の欠点

先程の地区ごと生徒数の人数を、各マスともに単純に10倍します。
10倍した人数で共分散を計算すると、28750になります。

人数は10倍していますが、「男子生徒が多い地区は女子生徒も多い傾向がある」という関係性においては本質的に何も差がありません。それなのに、共分散の数値は大きくなってしまいます。

この様に、共分散には
スケール変換に対して不変でない
という欠点があります。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	100	600	500	200	350
女子	60	400	300	100	215

偏差

	A地区	B地区	C地区	D地区	平均
男子	-250	250	150	-150	
女子	-155	185	85	-115	
地区ごと偏差の積	38750	46250	12750	17250	28750

相関係数

共分散は「スケール変換に対して不変でない」という欠点がありました。
この欠点を解消するため、共分散を規格化して相関係数という指標にします。

二組の対応するデータ(X, Y)に対し、相関係数 ρ は以下のように定義されます。

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$Cov(X,Y)$: 共分散
σ_X : Xの標準偏差
σ_Y : Yの標準偏差

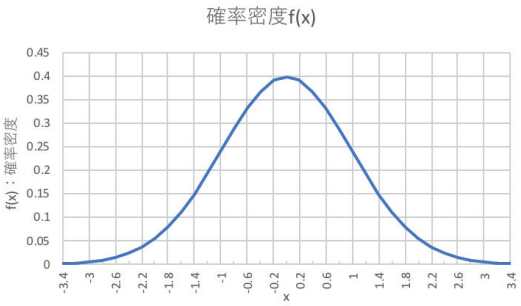
2変数のガウス分布（正規分布）

[第11回講義資料より抜粋]ガウス分布の確率密度関数

ガウス分布に従う確率変数Xの確率密度関数は以下の式で表されます。

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

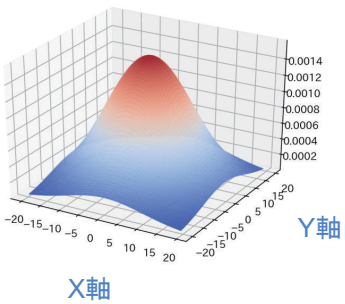
「σ」は標準偏差、「μ」は平均値を表します。
標準偏差σが1、平均μが0のガウス分布を、標準正規分布といいます。

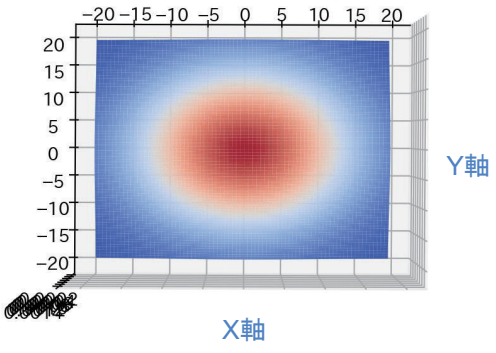


2次元のガウス分布

変数XとYの平均値、それぞれの分散、X-Y間の共分散を操作した際に、2次元のガウス分布の形状がどの様に変わるのかを観察します。

$$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix} = \begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$$の例

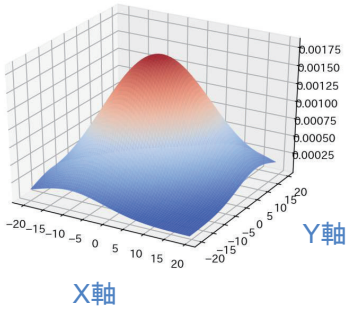


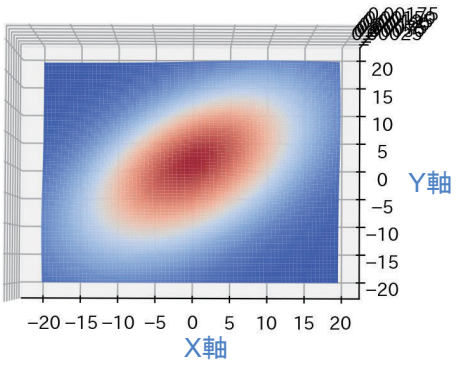


2次元のガウス分布

X–Y間の共分散を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & 50\\ 50 & 100\end{pmatrix}$の例

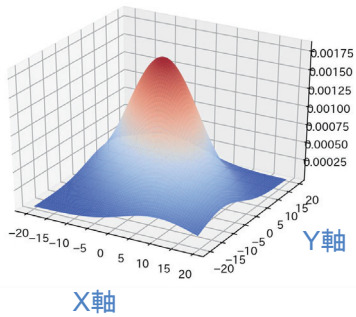


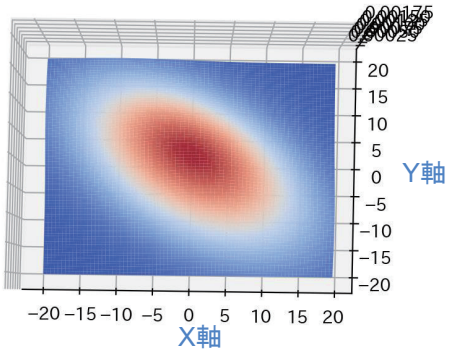


2次元のガウス分布

X–Y間の共分散を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & -50\\ -50 & 100\end{pmatrix}$の例

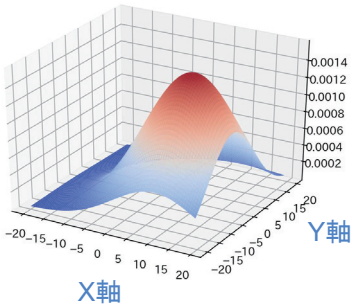


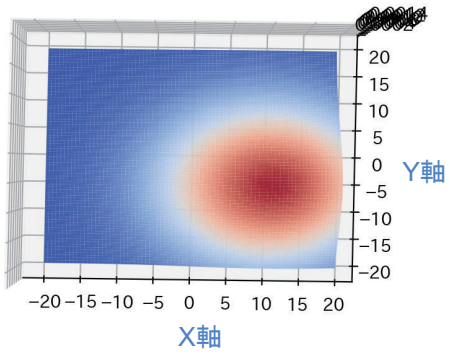


2次元のガウス分布

X、Yの平均値を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}10\\ -5\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$の例





演習問題

演習1：離散型同時確率分布

- 2変数の離散型同時確率分布の例を挙げてください。
- 上記で挙げた例の2つの変数それぞれに対し、周辺化を実施してください。

演習2：連続型同時確率分布

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\leq x\leq 1, 0\leq y\leq 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Yの周辺確率密度関数を求めてください。

演習3：確率変数の独立性

- 本資料で扱った男女別地区別生徒居住数のデータに対し、各セルの人数を変更して共分散を計算してください。
- 上記で作成した表に対し、各セルの人数を100倍して共分散を計算し、共分散値がスケールの変更に不変でないことを確認してください。

演習4：2変数のガウス分布

- [演習/Chapter12_Multidimensional_probability_distribution.ipynb]の平均値、共分散値を変更して実行し、2次元ガウス分布の形状がどのように変化するかを確認してください。

第5回：大数の法則

アジェンダ

- 大数の法則
- 大数の法則のコンピュータ・シミュレーション（サイコロの目）
- 大数の法則のコンピュータ・シミュレーション（コインの表裏）

大数の法則

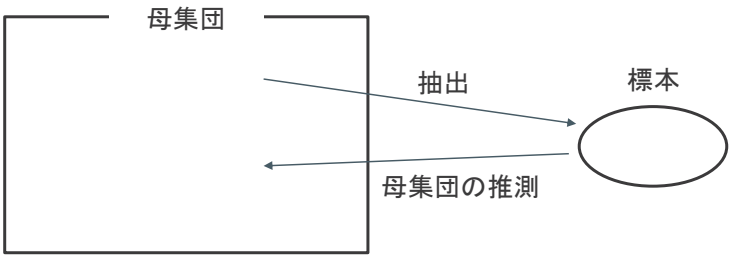
母集団と標本

日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。



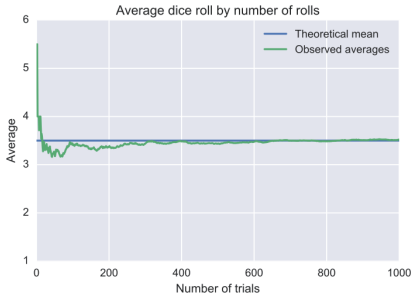
大数の法則とは

大数の法則とは、「ある独立した試行において、試行回数が大きくなるにつれて標本平均は母平均（期待値）に収束する」ということを意味します。

サイコロを何度も投げ続けることを考えます。サイコロの目の期待値は

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

なので、試行を繰り返すと標本平均は3.5に近づいていきます。

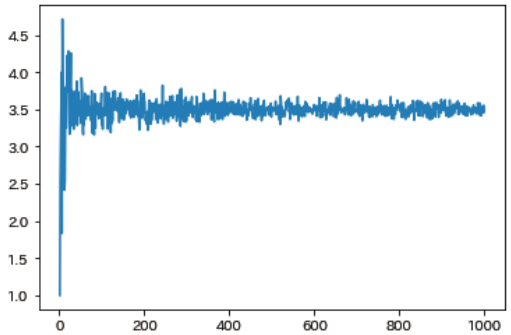


参照（Wikipedia）:
https://ja.wikipedia.org/wiki/%E5%A4%A7%E6%95%B0%E3%81%AE%E6%B3%95%E5%89%87

大数の法則のコンピュータ・シミュレーション
サイコロの目の平均値

本演習のゴール

サイコロを1回、2回、、、N回と多数投げ、出た目の平均値を計算していきます。
1回目からN回目までの平均値を図示することにより、試行回数を増やすにしたがってサイコロの目の平均値が3.5に収束することを確認します。



演習1：サイコロの目の平均値の計算

- 任意の回数サイコロを振り、出た目の平均値を求める関数を実装してください。
- サイコロを任意の回数振り、出た目の平均値を確認してください。

演習2：平均値の図示

- サイコロを1回振ったときの値、2回振ったときの平均値、3回振ったときの平均値、、、N回振ったときの平均値を図示する関数を実装してください。

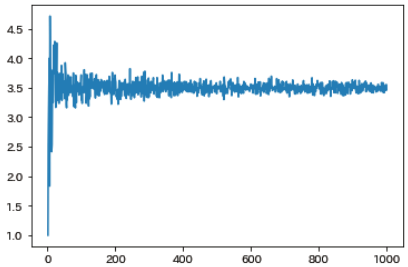
演習3：平均値の収束の確認

- サイコロを1,000回まで投げる試行を3回繰り返し、いずれも平均値が3.5に近づくことを確認してください。

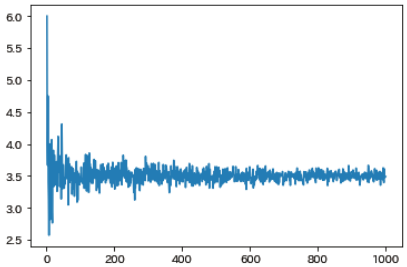
サイコロの目の平均値の収束

- サイコロを1,000回投げるまでの平均値の変化を図示した結果です。
- 3回テストを繰り返し、いずれも平均値が3.5に収束していることが確認できます。

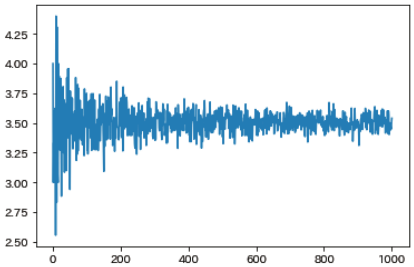
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目

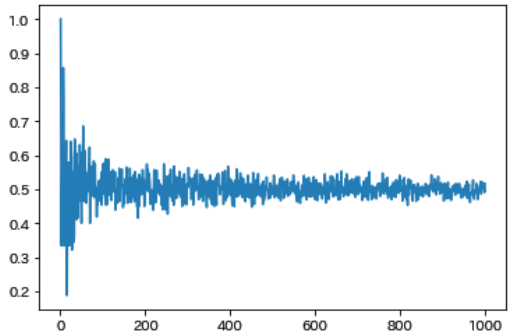


3回目

大数の法則のコンピュータ・シミュレーション
コインの裏表

本演習のゴール

コインを1回、2回、、、N回と多数投げ、表が出た場合を1、裏が出た場合を0として、表裏の平均値を計算していきます。
1回目からN回目までの平均値を図示することにより、試行回数を増やすにしたがってコインの裏表の平均値が0.5に収束することを確認します。



演習1：コインの裏表の平均値の計算

- 任意の回数コインを投げ、表が出た時を1、裏が出た時を0として、表裏の平均値を求める関数を実装してください。
- コインを任意の回数投げ、表裏の平均値を確認してください。

演習2：平均値の図示

- コインを1回投げたときの値、2回投げたときの平均値、3回投げたときの平均値、、、N回投げたときの平均値を図示する関数を実装してください。

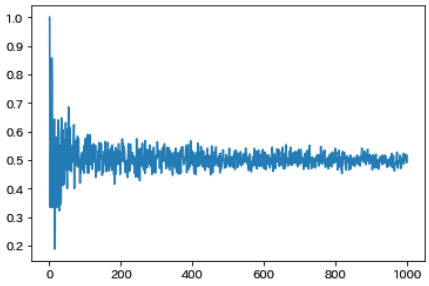
演習3：平均値の収束の確認

- コインを1,000回まで投げる試行を3回繰り返し、いずれも平均値が0.5に近づくことを確認してください。

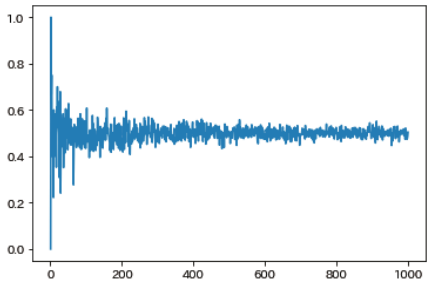
コインの裏表の平均値の収束

- コインを1,000回投げるまでの平均値の変化を図示した結果です。
- 3回テストを繰り返し、いずれも平均値が0.5に収束していることが確認できます。

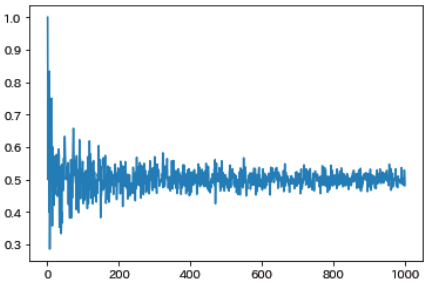
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

第6回：中心極限定理

アジェンダ

- 中心極限定理
- 中心極限定理のコンピュータ・シミュレーション（一様分布からのサンプリング）
- 中心極限定理のコンピュータ・シミュレーション（任意の分布からのサンプリング）

中心極限定理

中心極限定理とは

母集団の確率分布によらず、標本の大きさが十分に大きければ和や標本平均の分布は正規分布に従うという定理です。

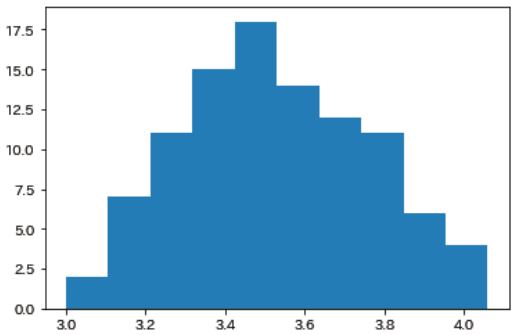
サンプル数をn、母集団の平均（母平均）をμ、分散（母分散）をσ^2とすると、$N(\mu, \sigma^2/n)$ という正規分布になります。

中心極限定理のコンピュータ・シミュレーション
一様分布からのサンプリング

中心極限定理のコンピュータ・シミュレーション

サイコロをN回振ったときの平均値の算出をM回繰り返し、平均値の分布を図示することにより、平均値の分布が正規分布に近づくことを確認します。

サイコロの目なので、母集団は一様分布となります。



演習1：サイコロの目の平均値の計算

- 任意の回数サイコロを振り、出た目の平均値を求める関数を実装してください。
- サイコロを任意の回数振り、出た目の平均値を確認してください。

演習2：平均値の分布の図示

- サイコロをN回振ったときの平均値の算出をM回繰り返し、その結果の分布を図示する関数を実装してください。

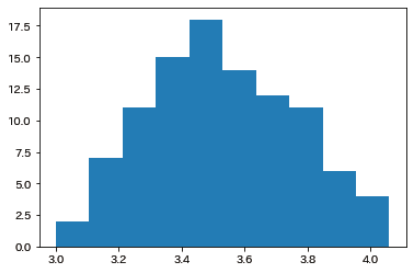
演習3：平均値の分布の形状の確認

- 「サイコロを50回投げて平均値を求めることを100回繰り返し散布図を図示する」ことを3回繰り返し、標本平均の分布が正規分布に近づくことを確認してください。

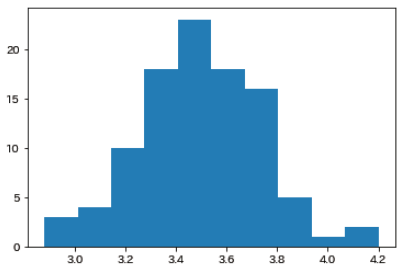
サイコロの目の平均値の分布

- サイコロを50回投げて平均値を求めることを100回繰り返したときの平均値の分布を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

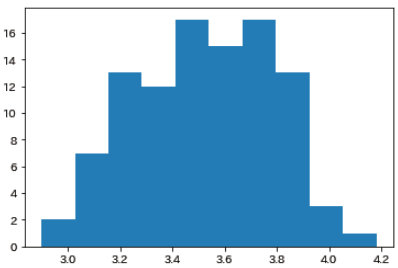
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

中心極限定理のコンピュータ・シミュレーション
任意の分布からのサンプリング

中心極限定理のコンピュータ・シミュレーション

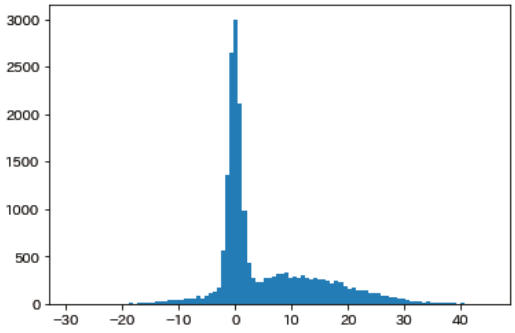
中心極限定理が成立する条件は、「サンプリング元の母集団の分布によらない」とされています。
前ページまでのシミュレーションではサイコロの目を扱ったため、母集団は一様分布にしたがっていました。

以降のシミュレーションでは、正規分布を2つ重ね合わせた分布を母集団とし、中心極限定理が成り立つことを確認します。

※中心極限定理が成立しない例外的な分布も存在しますが、ここでは取り扱いません。
https://ja.wikipedia.org/wiki/中心極限定理

演習1：母集団の生成

- 正規分布を2つ重ねた母集団データを作成してください。



演習2：抽出データの平均値の計算

- 任意の回数サンプルを抽出し、抽出したデータの平均値を求める関数を実装してください。
- 抽出したデータの平均値を確認してください。

演習3：平均値の分布の図示

- N回サンプリングしたデータの平均値の算出をM回繰り返し、その結果の分布を図示する関数を実装してください。

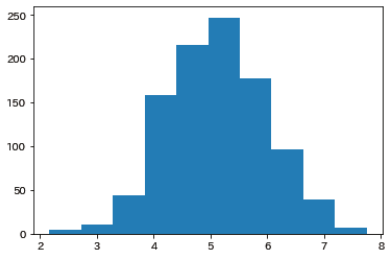
演習4：平均値の分布の形状の確認

- 「100回サンプリングして平均値を求めることを1,000回繰り返し散布図を図示する」ことを3回繰り返し、標本平均の分布が正規分布に近づくことを確認してください。

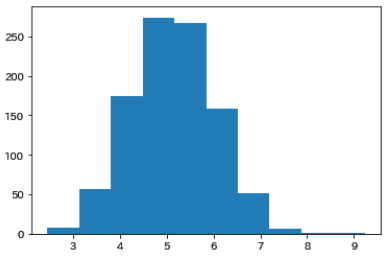
抽出データの平均値の分布

- 100回サンプリングして平均値を求めることを1,000回繰り返し散布図を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

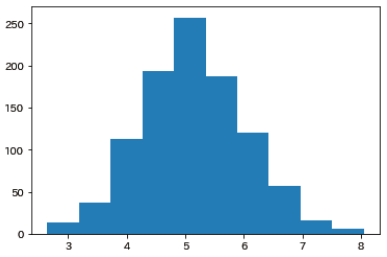
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

第7回：サンプリングと統計量

アジェンダ

- 記述統計学と推測統計学
- 母集団、標本、サンプリング
- 様々な統計量

記述統計学と推測統計学

記述統計学とは

記述統計学とは、入手済みのデータを集計する方法を学ぶ学問体系です。
また、データの特徴を簡単に表現する方法とも言えます。

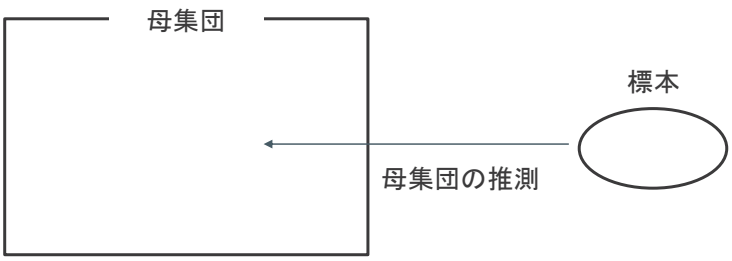
例えば、右下の表には2019年における東京都の市区町村の一部と、
人口を記載しています。
60あまりのデータがあるのですが、そのデータの概要を大まかに掴むために
人口の平均値、頻度分布などを作ったりします。

都道府県名	市区町村名	人口
東京都	中央区	154,851
東京都	港区	237,369
東京都	新宿区	303,094
東京都	文京区	210,681
東京都	台東区	183,859
東京都	墨田区	259,214
東京都	江東区	489,007
東京都	品川区	381,658
東京都	目黒区	270,240
東京都	大田区	705,335
東京都	世田谷区	887,528
東京都	渋谷区	215,955
東京都	中野区	312,332
東京都	杉並区	551,410
東京都	豊島区	259,285
東京都	北区	329,355
東京都	荒川区	196,835
東京都	板橋区	540,131
東京都	練馬区	712,780
東京都	足立区	656,806

推測統計学とは

推計統計学において、手元のデータは母集団の標本であると考えます。
この標本から母集団を推測することを試みます。

記述統計学では母集団と標本を区別しないため、標本に含まれないデータは取り扱うことができません。



母集団、標本、サンプリング

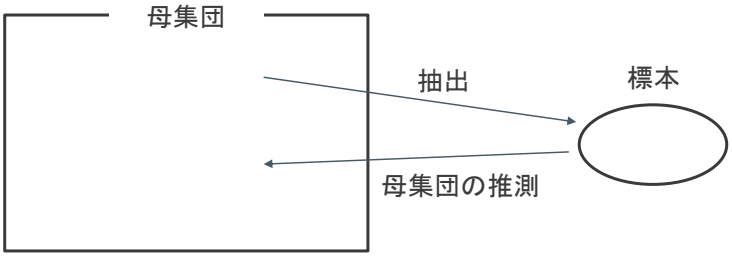
母集団と標本

日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。

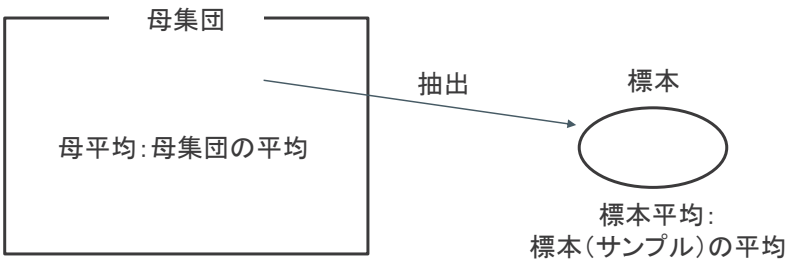


様々な統計量

母平均と標本平均

母集団全体の平均を母平均、サンプルの平均を標本平均といいます。

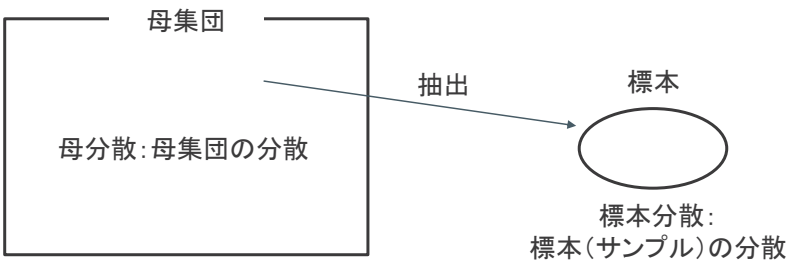
母集団を固定した場合、母平均は1つの確定した値になります。
標本平均は抽出の方法に依存した値となります。



母分散と標本分散

母集団全体の分散を母分散、サンプルの分散を標本分散といいます。

母集団を固定した場合、母分散は1つの確定した値になります。
標本分散は抽出の方法に依存した値となります。



推定量と不偏性と一致性

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。

不偏性（偏りがない）とは、推定量の期待値が、真の母数の値となることです。
これを満たす推定量を不偏推定量といいます。

一致性とは標本サイズが大きくなるほど、推定量が母集団の真の値に近づいていくことです。

不偏分散とは

標本したn個のデータから計算した分散が標本分散で、以下のように計算します。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

母集団に比べ標本数が少ない時は、標本分散が母分散よりも小さくなります。
そこで、標本分散が母分散に等しくなるように補正したものを不偏分散といい、以下のように計算できます。

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

演習問題

演習1：母平均と標本平均

母集団のサンプル数（体重）が5だとします。
それぞれの数値が42kg, 48kg, 53kg, 59kg, 71kgだとします。

- 母平均を求めてください。
- 体重が軽い方から3つのデータを抽出した標本の標本平均を求めてください。
- 体重が重い方から2つのデータを抽出した標本の標本平均を求めてください。

演習2：標本分散と不偏分散

- [data/tokyo_shikuchoson.tsv]の人口に対して標本分散を求めてください。
- 上のデータに対して不偏分散を求めてください。

演習3：記述統計（各種統計量）

- pandasのdescribe関数を使用し、平均値等の各種統計値を出力してください。

演習4：記述統計（ヒストグラム）

- pandasのplot関数を使用し、ヒストグラムを出力してください。

第8回：標本平均の標本分布

アジェンダ

- 区間推定
- 母分散既知の区間推定の例
- 母分散未知の区間推定の例
- Studentのt分布

区間推定

区間推定とは

母集団が正規分布に従うと仮定できるときに、標本から得られた値からある区間でもって母平均などの母数を推定する方法を区間推定といいます。

このときの区間のことを「信頼区間」といいます（「CI」と表記されることがあります）。

母分散既知/母分散未知と区間推定

母平均の区間推定では、母分散が分かっている場合と分からない場合とで、その算出方法が異なります。
母分散が分かっている場合を「母分散既知」、母分散が分からない場合を「母分散未知」といいます。

母分散既知の場合
- 母分散の値を使い、標準正規分布を用いて信頼区間を算出します。

母分散未知の場合
- 不偏分散の値を使い、t分布を用いて信頼区間を算出します。

母分散既知/母分散未知と区間推定

母平均は分からず母分散だけは分かっている、という状況は現実にはほとんどありません。
したがって、母平均の区間推定を行う場合にはt分布が用いられることがほとんどです。

母平均の区間推定では「95%信頼区間（95%CI）」を求めることが多々あります。
これは「母集団から標本平均を求めるという作業を100回実施した際、95回はその標本平均を含んでいると考えられる区間のこと」です。

このように、ある区間に母数が含まれる確率のことを「信頼係数」あるいは「信頼度」といいます。

標本平均の標本分布

母集団が正規分布している（正規母集団）とし、正規母集団から標本を取り出すことを考えます。

標本は何回も取り直せるため、標本の平均（標本平均）も異なる値をとります。よって、標本平均も確率変数と考えることができ、その分布を考えることができます。
この分布を標本平均の標本分布と言います。

母分散既知の区間推定の例

母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。
以上の条件で、母平均の95%信頼区間を求めます。

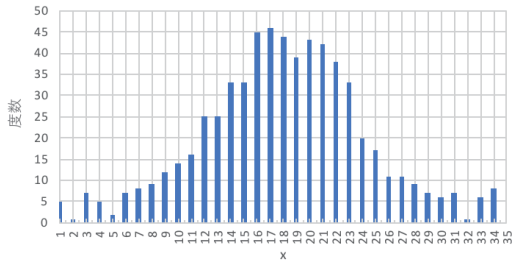
まず、標本平均$\bar{x}$を求めます。

$$\bar{x} = \frac{(5+7+5+\cdots+6+8)}{14} = 22.64$$

サンプル数をnとするとき、標本平均$\bar{x}$は
下記の式で標準化できます。

$$\frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}}$$

母集団



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

母分散既知の区間推定

ここで、標準正規分布を考えます。
右下の図のように、面積が95%となる上限値/下限値を求めます。
標準正規分布表[https://bellcurve.jp/statistics/course/8888.html]から、面積が95%となる上限値と下限値はそれぞれ+1.96、−1.96だとわかります。

正規化した標本平均は下記ようになります。

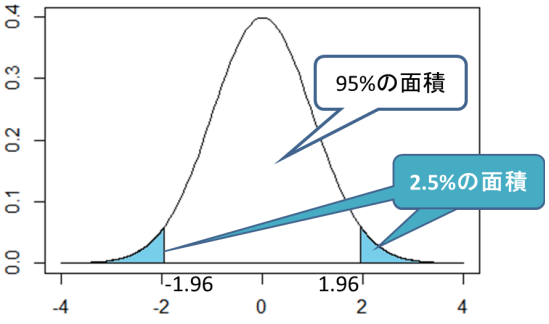
$$-1.96 \leq \frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96$$

この式を変形すると、母平均の95%信頼区間が求められます。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$22.64 - 1.96\sqrt{\frac{222.64}{14}} \leq \mu \leq 22.64 + 1.96\sqrt{\frac{222.64}{14}}$$

$$14.83 \leq \mu \leq 30.46$$



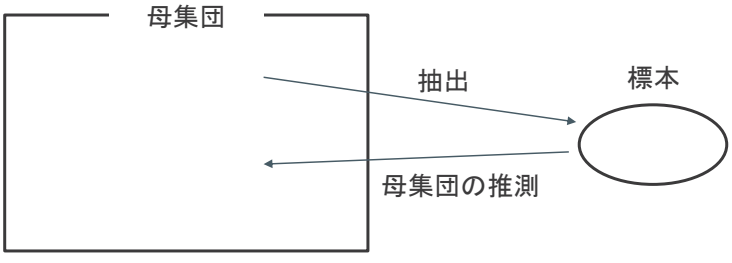
参照：https://bellcurve.jp/statistics/course/8888.html

母分散未知の区間推定の例

母分散未知の区間推定

前ページまでで、母分散既知の母平均の信頼区間の算出方法について学びました。
しかし、母平均が分からないのに母分散だけは分かっているという状況はほとんどありません。

ここからは、母分散未知の場合について学習します。
母集団未知の場合、母平均の区間推定を行うにはt分布（あるいはStudentのt分布ともいいます）を用います。



母分散未知の区間推定

母分散既知のときは下記の式を使用しましたが、母分散未知ではこの式は使用できません。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

母分散未知のときは、母分散σ^2の代わりに不偏分散s^2を使用します。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{14}(x_i - \bar{x})^2 = 310.09$$

また、母分散既知の際は1.96を求めるのに標準正規分布表を使用しましたが、母分散未知の場合はt分布表を使用します。

母分散未知の区間推定

t分布表[https://bellcurve.jp/statistics/course/8970.html]を参照します。
14個のデータから母平均の95%信頼区間を求めるので、

$$\alpha = 0.025$$
$$v = 13$$

が交わるところを参照すると、2.160であることがわかります。

$$\bar{x} - 2.160\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + 2.160\sqrt{\frac{s^2}{n}}$$

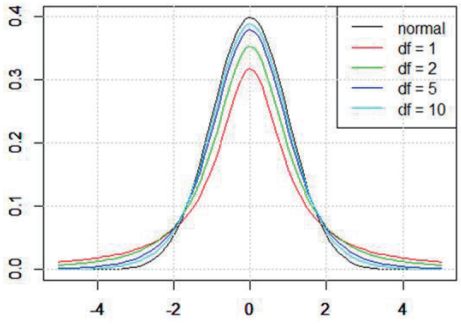
標本平均、不偏分散、サンプル数をそれぞれ代入すると、母平均の95%信頼区間を求めることができます。

$$22.64 - 2.160\sqrt{\frac{310.09^2}{14}} \leq \mu \leq 22.64 + 2.160\sqrt{\frac{310.09^2}{14}}$$
$$12.48 \leq \mu \leq 32.81$$

Studentのt分布

Studentのt分布は標準正規分布とよく似た形の分布で、「自由度」をパラメータとして分布の形が変わるという特徴を持っています。

自由度（グラフ中ではdfで表示しています）を変化させた時のt分布の形を右下の図に示します。
自由度が1、2、5、10と大きくなるにつれ、標準正規分布（黒線：normal）に近づくことが分かります。



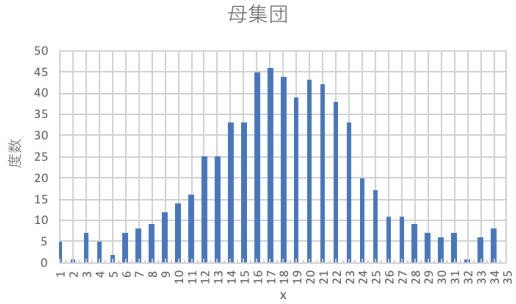
参照：https://bellcurve.jp/statistics/course/8968.html

演習問題

演習1：母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。

母平均の96%信頼区間を求めてください。

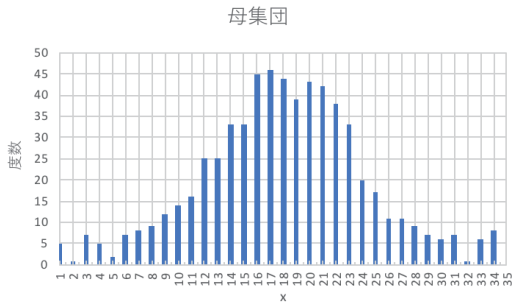


No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

演習2：母分散未知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、ランダムに14個のデータを抽出しました。

母分散未知として、母平均の90%信頼区間を求めてください。



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

第9回：統計学Ⅱ総復習 その1

第1回：微分

微分の定義

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

$f(x)=x^2$のとき、上式にしたがって微分を実施すると、以下のようになります。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}=\lim_{h\to 0}\frac{(x+h)^2-x^2}{h}=\lim_{h\to 0}\frac{x^2+2hx+h^2-x^2}{h}=\lim_{h\to 0}(2x+h)=2x$$

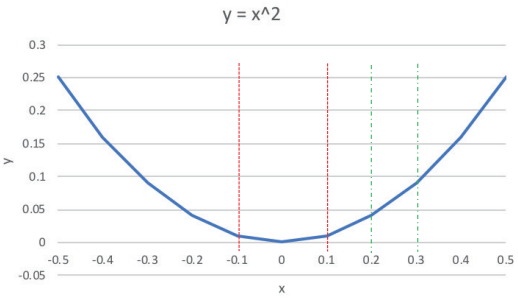
ここで求めた導関数を基に計算すると、

x=0.2のとき0.4

x=0.3のとき0.6

となります。「微分の例：2次関数」ではxが0.2から0.3に増えるときの傾きを計算し0.5となりました。

これは上の数字のちょうど真ん中にあることが確認できます。



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

可微分性

関数$f(x)$の任意の場所aにおける微分は、以下の式で表せました。

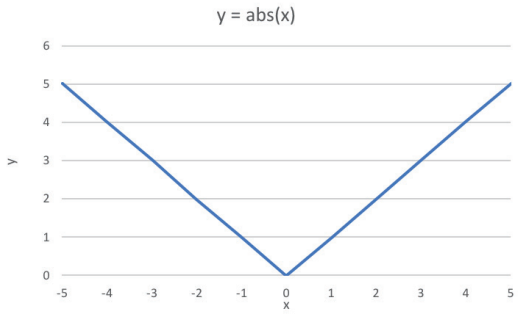
$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

絶対値関数$f(x)=|x|$の、x=0における傾きは

h＞0のときは1

h＜0のときは−1

となり、x=0で微分可能ではありません。



偏微分とは

偏微分とは，変数が複数ある関数の「特定の変数以外は定数だとみなして」微分することです。
2変数（X, Y）の場合、偏微分は以下のように表されます。

$\frac{\partial f(x,y)}{\partial x}, f_x$: xによる偏微分
$\frac{\partial f(x,y)}{\partial y}, f_y$: yによる偏微分

偏微分の計算例

「$f(x,y) = x^2 + y^3 + 5y + xy$」についての偏微分を取り扱います。

上式をxで偏微分します（xのみ変数として扱い、yは定数とみなします）。

$$\frac{\partial f(x,y)}{\partial x} = 2x + y$$

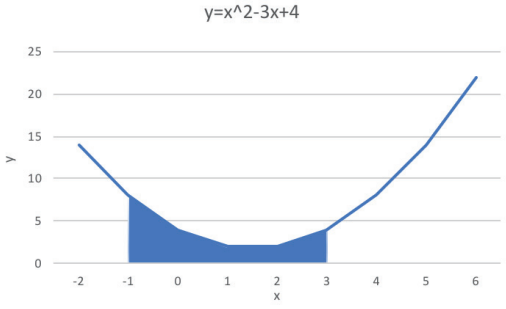
上式をyで偏微分します（yのみ変数として扱い、xは定数とみなします）。

$$\frac{\partial f(x,y)}{\partial y} = 3y^2 + 5 + x$$

第2回：積分

積分とは

- 積分とは、任意の関数f(x)で囲まれた部分の面積を求めることを意味しています。
  - $\int_a^b f(x)dx$
- 例えば$f(x) = x^2 - 3x + 4$、a=−1、b=3の場合、下図の青い部分の面積を求めることができます。



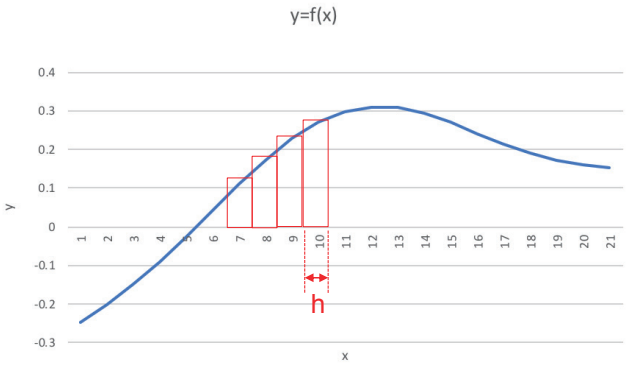
複雑な関数の積分

これまでの例では、導関数の原始関数を解析的に求めることができました。

解析的に求めることができない関数に対して面積を算出する際は、コンピュータプログラムなどで以下のようにして面積の近似値を求めます。

$$\sum_{i=a}^{b} f(i)h$$

hの間隔を徐々に狭くしていけば、上式の値は真の値に近づいていきます。



微分と積分の関係

「複雑な関数の積分」のページでは、面積の近似値をプログラムで求める方法を記載しましたが、より厳密に面積を求めていきます。

下図のオレンジ色の面積は、青い部分より大きく、青＋赤より小さいことがわかります。

$$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$$

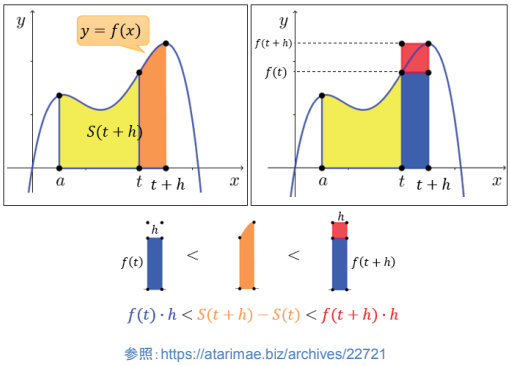
$$f(t) < \frac{S(t+h)-S(t)}{h} < f(t+h)$$

hの極限をとると、

$$f(t) < \lim_{h\to 0}\frac{S(t+h)-S(t)}{h} < \lim_{h\to 0} f(t+h) = f(t)$$

$$\lim_{h\to 0}\frac{S(t+h)-S(t)}{h} = f(t)$$

最後の式は、f(x)の導関数を求める式と同じであることが確認できます。



$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$

参照：https://atarimae.biz/archives/22721

第3回：確率変数

確率変数とは

ある現象がいろいろな値を取り得るとき、取り得る値全体を確率変数といいます。

例えば、サイコロを振ったときに出る目は[1, 2, 3, 4, 5, 6]のいずれかとなります。
この場合、確率変数Xは

$$\mathrm{X} = 1, 2, 3, 4, 5, 6$$

と表します。

確率変数をXと置くことで、サイコロの目が取りうる値の確率を、以下のように記載することができます。

$$\mathrm{P}(x) = \frac{1}{6}(\mathrm{X} = 1, 2, 3, 4, 5, 6)$$

サイコロを振って4が出る確率は以下のように書きます。

$$\mathrm{P}(x = 4) = \frac{1}{6}$$

離散型の確率変数

離散型確率変数は、「とびとびの値」を指します。
隣り合った数値の間には、数値は存在しません。
例えばサイコロの目、コインの裏表、ルーレットの番号などが該当します。

連続型の確率変数

連続型確率変数は、「連続した値」を指します。
例えば速度であれば、5km/hと6km/hの間には5.1km/hや5.01km/h、5.0001km/hなど無数の値が存在します。

その他の連続確率変数には温度、湿度、高度、体重などがあります。

確率分布とは

確率変数のそれぞれの値に対し、その確率変数をとる確率の分布のことです。

離散型確率変数に対する確率分布として、以下のような確率分布があります。

- ポアソン分布
- 二項分布
- 幾何分布
- 一様分布

連続型確率変数に対する確率分布として、以下のような確率分布があります。

- 正規分布
- 指数分布
- 一様分布

期待値とは

期待値とは、1回の試行で得られる値の平均値のことです。
得られうるすべての値（すべての確率変数）とそれが起こる確率の積を足し合わせて計算できます。

分散とは

分散とは、「[確率変数の全ての値と期待値（平均値）の差]の2乗」と「確率」との積を、全て足し合わせたものです。
分散は英語でVarianceと表記するので、頭文字を使って$V(X)$と表記します。

第4回：多次元の確率分布

離散型同時確率分布とは

2つの離散型確率変数XとYが、それぞれある値をとるときの確率を表したものを「離散型同時確率分布」といいます。

例えば、男子20名、女子20名のあるクラスがあるとします。生徒の居住地区を表にしてみました。
性別をX、居住地区をYとすると、2つの離散型確率変数とみなせます。

	A地区	B地区	C地区	D地区	計
男子	4	6	6	4	20
女子	6	8	4	2	20
				総合計	40

全生徒40人に対する各マスの数字の割合を計算してみました。これは2つの離散型確率変数XとYがそれぞれの値を同時にとる、離散型同時確率分布となります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

離散型同時確率分布とは

2つの確率変数からなる同時確率分布は、以下のように表記します。

$$f(x_i, y_j) = P(X = x_i, Y = y_j) \quad i = 1,2,3,,,;j = 1,2,3,,,$$

例えば、男子でD地区に住む生徒の確率は、以下のようになります。

$$P(X = \text{男子}, Y = \text{D地区}) = 0.1$$

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

ここで$f(x_i, y_j)$のことを同時確率関数といいます。各iとjについて全ての確率を足すと総和は1になります。

$$\sum_i \sum_j f(x_i, y_j) = 1$$

周辺確率分布

性別X、居住地区Yのそれぞれの値について、確率の合計を計算してみます。
男子の割合は0.5、A地区に居住する生徒の割合は0.25であることがわかります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5
計	0.25	0.35	0.25	0.15	1.00

このようにある1つの確率変数を固定し、別の確率変数が取りうる全ての確率を合計したものを周辺確率分布といいます。

$$f_x(x_i) = \sum_j f(x_i, y_j) = P(X = x_i) \quad i = 1,2,3,,,$$
$$f_y(y_j) = \sum_i f(x_i, y_j) = P(Y = y_j) \quad j = 1,2,3,,,$$

ここで、$f_x(x_i)$と$f_y(y_j)$をそれぞれXとYの周辺確率関数といいます。

連続型同時確率分布とは

XとYが連続型確率変数であるとき、それぞれある値をとるときの確率を表したものを「連続型同時確率分布」といいます。

XとYの同時確率分布を表す関数を「同時確率密度関数」といいます。

$$P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$$

確率の総和は1になるため、同時確率密度関数に関して以下の式が成り立ちます。

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

連続型確率変数XとYの周辺確率密度関数は、以下の式で求めることができます。

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

同時確率変数XとYの全範囲についての確率を求めてみます。

$$P(0\le x\le 1, 0\le y\le 1)=\int_0^1\int_0^1 (x+y)\,dx\,dy=\int_0^1\left[\frac{x^2}{2}+yx\right]_0^1 dy=\int_0^1\left(\frac{1}{2}+y\right)dy$$

$$=\left[\frac{y}{2}+\frac{y^2}{2}\right]_0^1=1$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Xの周辺確率密度関数を求めてみます。

$$f_x(x)=\int_0^1 (x+y)\,dy=\left[x+\frac{y^2}{2}\right]_0^1=x+\frac{1}{2}$$

独立な確率変数とは

2つの確率変数XとYの同時確率分布（同時確率密度関数）f(x, y)が、それぞれの確率変数の周辺確率分布（周辺確率密度関数）g(x)とh(y)の積に分解できる時、その2つの確率変数は独立(independent)であると言います。

$$f(x, y) = g(x)h(y)$$

直感的な理解としては、「XとYの動きは、お互いに影響を及ぼさない」ということです。

第5回：大数の法則

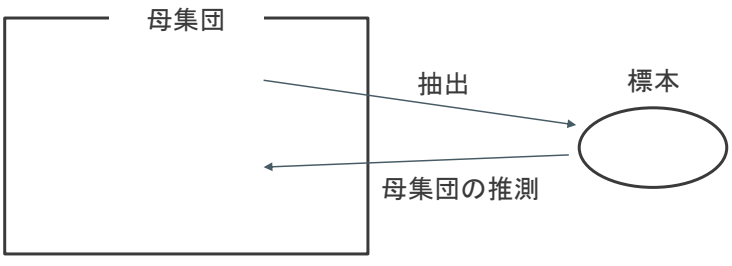
母集団と標本

日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。



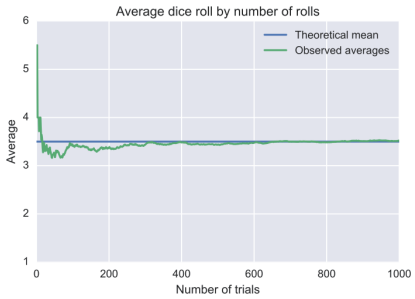
大数の法則とは

大数の法則とは、「ある独立した試行において、試行回数が大きくなるにつれて標本平均は母平均（期待値）に収束する」ということを意味します。

サイコロを何度も投げ続けることを考えます。サイコロの目の期待値は

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

なので、試行を繰り返すと標本平均は3.5に近づいていきます。



参照（Wikipedia）:
https://ja.wikipedia.org/wiki/%E5%A4%A7%E6%95%B0%E3%81%AE%E6%B3%95%E5%89%87

第6回：中心極限定理

中心極限定理とは

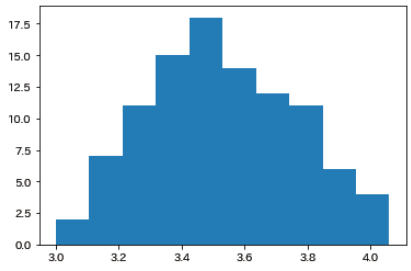
母集団の確率分布によらず、標本の大きさが十分に大きければ和や標本平均の分布は正規分布に従うという定理です。

サンプル数をn、母集団の平均（母平均）をμ、分散（母分散）をσ^2とすると、$N(\mu, \sigma^2/n)$ という正規分布になります。

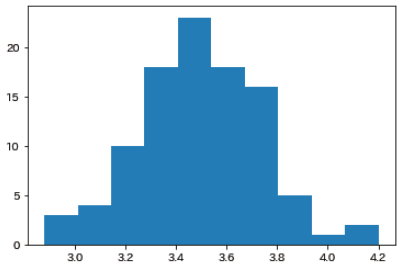
サイコロの目の平均値の分布

- サイコロを50回投げて平均値を求めることを100回繰り返したときの平均値の分布を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

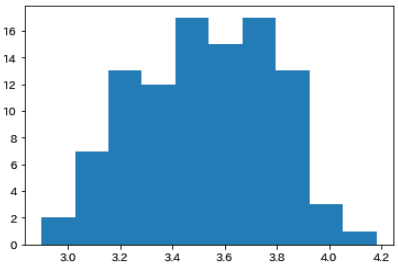
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

第7回：サンプリングと統計量

記述統計学とは

記述統計学とは、入手済みのデータを集計する方法を学ぶ学問体系です。
また、データの特徴を簡単に表現する方法とも言えます。

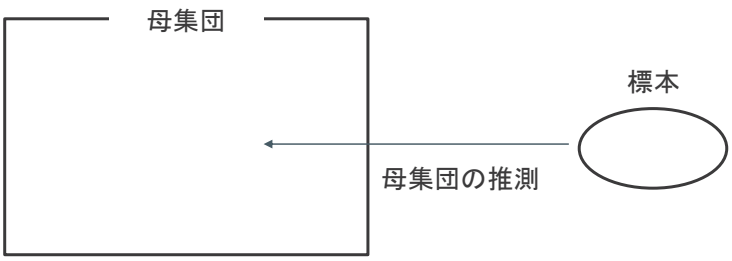
例えば、右下の表には2019年における東京都の市区町村の一部と、
人口を記載しています。
60あまりのデータがあるのですが、そのデータの概要を大まかに掴むために
人口の平均値、頻度分布などを作ったりします。

都道府県名	市区町村名	人口
東京都	中央区	154,851
東京都	港区	237,369
東京都	新宿区	303,094
東京都	文京区	210,681
東京都	台東区	183,859
東京都	墨田区	259,214
東京都	江東区	489,007
東京都	品川区	381,658
東京都	目黒区	270,240
東京都	大田区	705,335
東京都	世田谷区	887,528
東京都	渋谷区	215,955
東京都	中野区	312,332
東京都	杉並区	551,410
東京都	豊島区	259,285
東京都	北区	329,355
東京都	荒川区	196,835
東京都	板橋区	540,131
東京都	練馬区	712,780
東京都	足立区	656,806

推測統計学とは

推計統計学において、手元のデータは母集団の標本であると考えます。
この標本から母集団を推測することを試みます。

記述統計学では母集団と標本を区別しないため、標本に含まれないデータは取り扱うことができません。



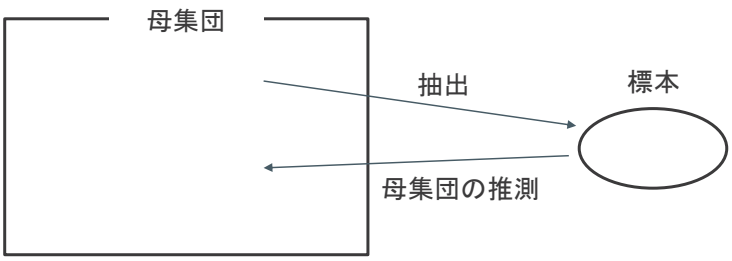
母集団と標本

日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

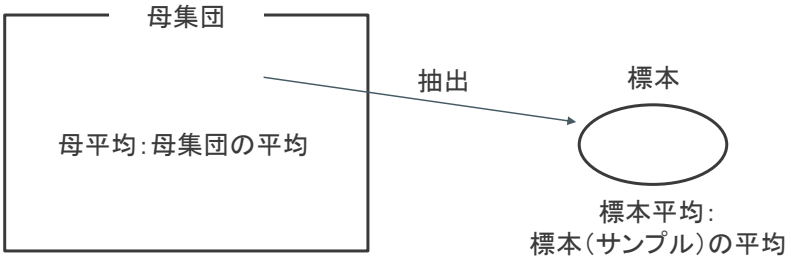
ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。



母平均と標本平均

母集団全体の平均を母平均、サンプルの平均を標本平均といいます。

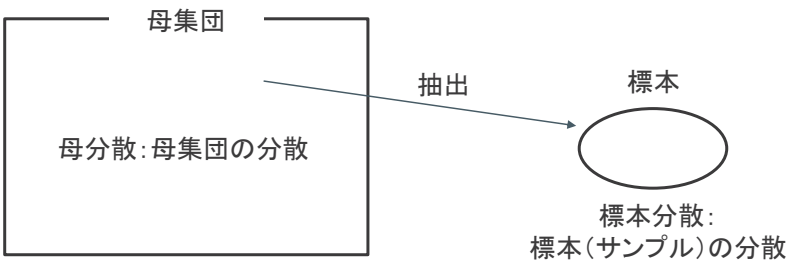
母集団を固定した場合、母平均は1つの確定した値になります。
標本平均は抽出の方法に依存した値となります。



母分散と標本分散

母集団全体の分散を母分散、サンプルの分散を標本分散といいます。

母集団を固定した場合、母分散は1つの確定した値になります。
標本分散は抽出の方法に依存した値となります。



推定量と不偏性と一致性

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。

不偏性（偏りがない）とは、推定量の期待値が、真の母数の値となることです。
これを満たす推定量を不偏推定量といいます。

一致性とは標本サイズが大きくなるほど、推定量が母集団の真の値に近づいていくことです。

不偏分散とは

標本したn個のデータから計算した分散が標本分散で、以下のように計算します。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

母集団に比べ標本数が少ない時は、標本分散が母分散よりも小さくなります。
そこで、標本分散が母分散に等しくなるように補正したものを不偏分散といい、以下のように計算できます。

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

第8回：標本平均の標本分布

区間推定とは

母集団が正規分布に従うと仮定できるときに、標本から得られた値からある区間でもって母平均などの母数を推定する方法を区間推定といいます。

このときの区間のことを「信頼区間」といいます（「CI」と表記されることがあります）。

母分散既知/母分散未知と区間推定

母平均の区間推定では、母分散が分かっている場合と分からない場合とで、その算出方法が異なります。
母分散が分かっている場合を「母分散既知」、母分散が分からない場合を「母分散未知」といいます。

母分散既知の場合

- 母分散の値を使い、標準正規分布を用いて信頼区間を算出します。

母分散未知の場合

- 不偏分散の値を使い、t分布を用いて信頼区間を算出します。

母分散既知/母分散未知と区間推定

母平均は分からず母分散だけは分かっている、という状況は現実にはほとんどありません。
したがって、母平均の区間推定を行う場合にはt分布が用いられることがほとんどです。

母平均の区間推定では「95%信頼区間（95%CI）」を求めることが多々あります。
これは「母集団から標本平均を求めるという作業を100回実施した際、95回はその標本平均を含んでいると考えられる区間のこと」です。

このように、ある区間に母数が含まれる確率のことを「信頼係数」あるいは「信頼度」といいます。

標本平均の標本分布

母集団が正規分布している（正規母集団）とし、正規母集団から標本を取り出すことを考えます。

標本は何回も取り直せるため、標本の平均（標本平均）も異なる値をとります。よって、標本平均も確率変数と考えることができ、その分布を考えることができます。
この分布を標本平均の標本分布と言います。

母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。
以上の条件で、母平均の95%信頼区間を求めます。

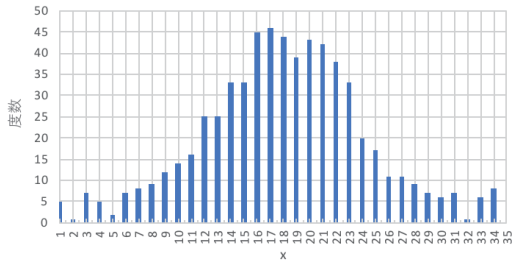
まず、標本平均$\bar{x}$を求めます。

$$\bar{x} = \frac{(5+7+5+\cdots+6+8)}{14} = 22.64$$

サンプル数をnとするとき、標本平均$\bar{x}$は
下記の式で標準化できます。

$$\frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}}$$

母集団



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

母分散既知の区間推定

ここで、標準正規分布を考えます。
右下の図のように、面積が95%となる上限値/下限値を求めます。
標準正規分布表[https://bellcurve.jp/statistics/course/8888.html]から、面積が95%となる上限値と下限値はそれぞれ+1.96、−1.96だとわかります。

正規化した標本平均は下記ようになります。

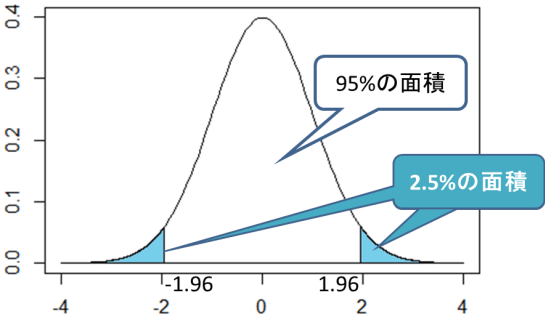
$$-1.96 \leq \frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96$$

この式を変形すると、母平均の95%信頼区間が求められます。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$22.64 - 1.96\sqrt{\frac{222.64}{14}} \leq \mu \leq 22.64 + 1.96\sqrt{\frac{222.64}{14}}$$

$$14.83 \leq \mu \leq 30.46$$

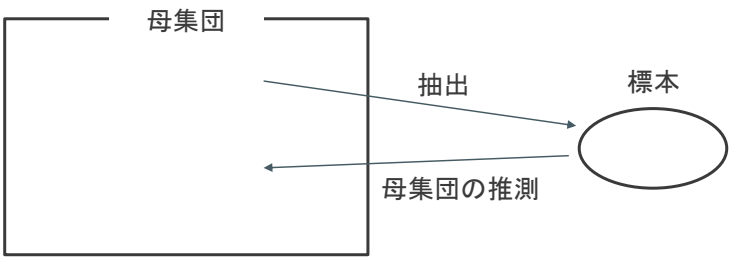


参照：https://bellcurve.jp/statistics/course/8888.html

母分散未知の区間推定

前ページまでで、母分散既知の母平均の信頼区間の算出方法について学びました。
しかし、母平均が分からないのに母分散だけは分かっているという状況はほとんどありません。

ここからは、母分散未知の場合について学習します。
母集団未知の場合、母平均の区間推定を行うにはt分布（あるいはStudentのt分布ともいいます）を用います。



母分散未知の区間推定

母分散既知のときは下記の式を使用しましたが、母分散未知ではこの式は使用できません。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

母分散未知のときは、母分散σ^2の代わりに不偏分散s^2を使用します。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{14}(x_i - \bar{x})^2 = 310.09$$

また、母分散既知の際は1.96を求めるのに標準正規分布表を使用しましたが、母分散未知の場合はt分布表を使用します。

母分散未知の区間推定

t分布表[https://bellcurve.jp/statistics/course/8970.html]を参照します。
14個のデータから母平均の95%信頼区間を求めるので、

$$\alpha = 0.025$$
$$v = 13$$

が交わるところを参照すると、2.160であることがわかります。

$$\bar{x} - 2.160\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + 2.160\sqrt{\frac{s^2}{n}}$$

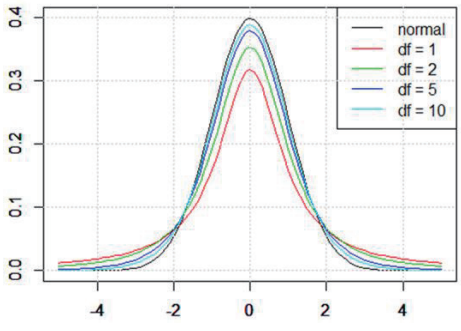
標本平均、不偏分散、サンプル数をそれぞれ代入すると、母平均の95%信頼区間を求めることができます。

$$22.64 - 2.160\sqrt{\frac{310.09^2}{14}} \leq \mu \leq 22.64 + 2.160\sqrt{\frac{310.09^2}{14}}$$
$$12.48 \leq \mu \leq 32.81$$

Studentのt分布

Studentのt分布は標準正規分布とよく似た形の分布で、「自由度」をパラメータとして分布の形が変わるという特徴を持っています。

自由度（グラフ中ではdfで表示しています）を変化させた時のt分布の形を右下の図に示します。
自由度が1、2、5、10と大きくなるにつれ、標準正規分布（黒線：normal）に近づくことが分かります。



参照：https://bellcurve.jp/statistics/course/8968.html

第10回：点推定

アジェンダ

- 点推定と区間推定
- 点推定
- 最尤法

点推定と区間推定

母集団と標本

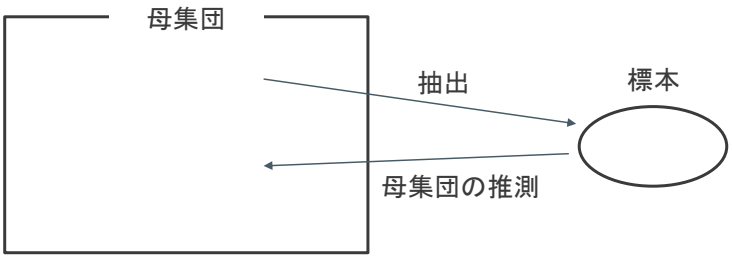
日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。

実際には母集団全てを知り得ることができないことが一般的であるため、標本から母集団を推定することとなります。

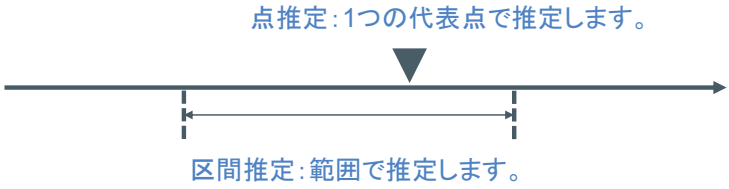


点推定と区間推定

推定には大別して点推定と区間推定があります。

点推定とは、標本から求められる一つの代表点によって母数を推定する方法です。
例えば、標本平均をそのまま母平均の推定値とするのも点推定の1つです。

区間推定とは、標本から2つの値を計算し、その間に母数が含まれていると考える推定方法です。
例えば、95%信頼区間は区間推定の一つです。

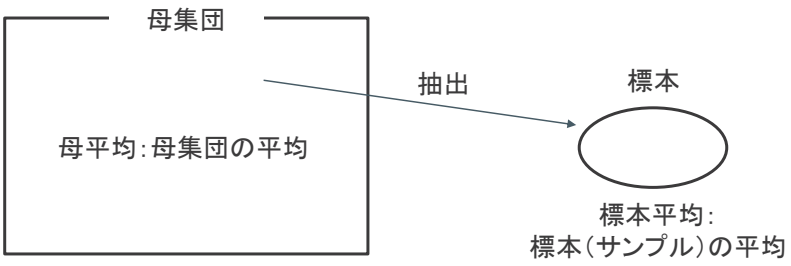


点推定

母平均と標本平均

母集団全体の平均を母平均、サンプルの平均を標本平均といいます。

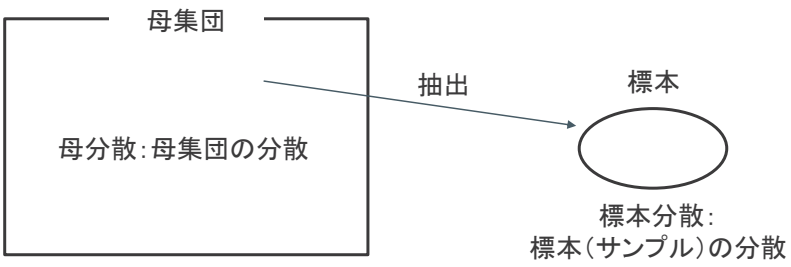
母集団を固定した場合、母平均は1つの確定した値になります。
標本平均は抽出の方法に依存した値となります。



母分散と標本分散

母集団全体の分散を母分散、サンプルの分散を標本分散といいます。

母集団を固定した場合、母分散は1つの確定した値になります。
標本分散は抽出の方法に依存した値となります。



推定量と推定値

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。
例えば平均は以下のように表します。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

統計量やその関数のことを推定量といいます。

また、実際に試行し得た値より計算した数値のことを、推定値といいます。

不偏性と一致性

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。

不偏性（偏りがない）とは、推定量の期待値が、真の母数の値となることです。
これを満たす推定量を不偏推定量といいます。

一致性とは標本サイズが大きくなるほど、推定量が母集団の真の値に近づいていくことです。

不偏分散とは

標本したn個のデータから計算した分散が標本分散で、以下のように計算します。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$：標本の平均

母集団に比べ標本数が少ない時は、標本分散が母分散よりも小さくなります。
そこで、標本分散が母分散に等しくなるように補正したものを不偏分散といい、以下のように計算できます。

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$：標本の平均

最尤法

最尤法とは

[Wikipediaより]最尤推定（さいゆうすいてい、英: maximum likelihood estimation、略してMLEともいう）や最尤法（さいゆうほう、英: method of maximum likelihood）とは、統計学において、与えられたデータからそれが従う確率分布の母数を点推定する方法である。

観測されたデータからそれを生んだ母集団を説明しようとする際に広く用いられる。生物学では塩基やアミノ酸配列のような分子データの置換に関する確率モデルに基づいて系統樹を作成する際に、一番尤もらしくデータを説明する樹形を選択するための有力な方法としても利用される。機械学習ではニューラルネットワーク（特に生成モデル）を学習する際に最尤推定（負の対数尤度最小化として定式化）が用いられる。

最尤法の例：反復試行の確率

確率pで成功する試行を独立にn回反復して実施したとき、n回のうちk回成功する確率は以下のように表します。

$$ {}_nC_k p^k (1-p)^{n-k} $$

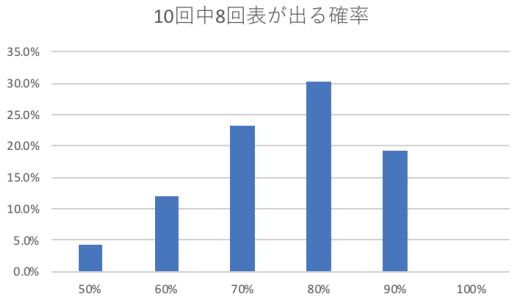
上の式を応用し、次の問題を最尤法で考えてみます。

ある特殊なコインがあるとします。このコインは表の出る確率が1/2ではありません。
このコインを10回投げたところ、8回表が出ました。
この特殊なコインの表が出る確率はいくつだと考えるのが妥当でしょうか。

最尤法の例：反復試行の確率

コインの表が出る確率が50%～100%としたときの、「10回中8回表がでる確率」を計算したものが下の表です。
10%刻みで計算すると、表が出る確率が80%とした場合に「10回中8回表がでる確率」が最も大きくなることがわかります。
したがって、表がでる確率が80%とすることが尤もらしいと言えそうです。

表が出る確率	10回中8回表が出る確率
50%	4.4%
60%	12.1%
70%	23.3%
80%	30.2%
90%	19.4%
100%	0.0%



演習問題

演習1：母平均と標本平均

母集団のサンプル数（体重）が5だとします。
それぞれの数値が42kg, 48kg, 53kg, 59kg, 71kgだとします。

- 母平均を求めてください。
- 体重が軽い方から3つのデータを抽出した標本の標本平均を求めてください。
- 体重が重い方から2つのデータを抽出した標本の標本平均を求めてください。

演習2：標本分散と不偏分散

- [data/tokyo_shikuchoson.tsv]の人口に対して標本分散を求めてください。
- 上のデータに対して不偏分散を求めてください。

演習3：最尤法

- ある特殊なコインがあるとします。このコインは表の出る確率が1/2ではありません。このコインを10回投げたところ、6回表が出ました。この特殊なコインの表が出る確率はいくつだと考えるのが妥当でしょうか。

第11回：区間推定

アジェンダ

- 点推定と区間推定
- 区間推定
- 母分散既知の区間推定の例
- 母分散未知の区間推定の例

点推定と区間推定

母集団と標本

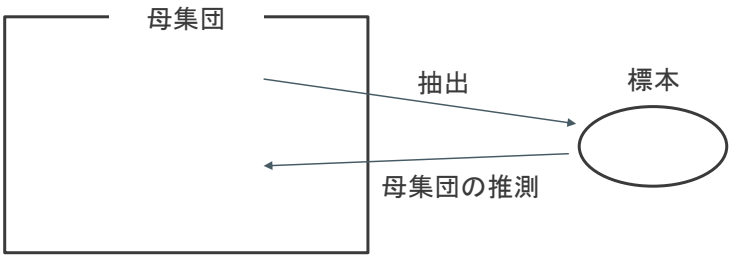
日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。

実際には母集団全てを知り得ることができないことが一般的であるため、標本から母集団を推定することとなります。

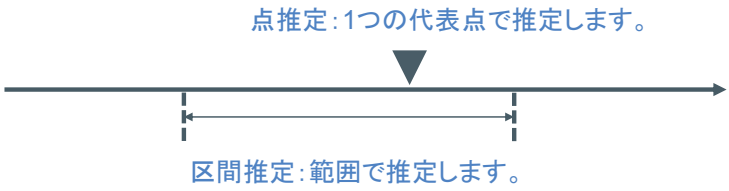


点推定と区間推定

推定には大別して点推定と区間推定があります。

点推定とは、標本から求められる一つの代表点によって母数を推定する方法です。
例えば、標本平均をそのまま母平均の推定値とするのも点推定の1つです。

区間推定とは、標本から2つの値を計算し、その間に母数が含まれていると考える推定方法です。
例えば、95%信頼区間は区間推定の一つです。



区間推定

区間推定とは

母集団が正規分布に従うと仮定できるときに、標本から得られた値からある区間でもって母平均などの母数を推定する方法を区間推定といいます。

このときの区間のことを「信頼区間」といいます（「CI」と表記されることがあります）。

母分散既知/母分散未知と区間推定

母平均の区間推定では、母分散が分かっている場合と分からない場合とで、その算出方法が異なります。
母分散が分かっている場合を「母分散既知」、母分散が分からない場合を「母分散未知」といいます。

母分散既知の場合
- 母分散の値を使い、標準正規分布を用いて信頼区間を算出します。

母分散未知の場合
- 不偏分散の値を使い、t分布を用いて信頼区間を算出します。

母分散既知/母分散未知と区間推定

母平均は分からず母分散だけは分かっている、という状況は現実にはほとんどありません。
したがって、母平均の区間推定を行う場合にはt分布が用いられることがほとんどです。

母平均の区間推定では「95%信頼区間（95%CI）」を求めることが多々あります。
これは「母集団から標本平均を求めるという作業を100回実施した際、95回はその標本平均を含んでいると考えられる区間のこと」です。

このように、ある区間に母数が含まれる確率のことを「信頼係数」あるいは「信頼度」といいます。

標本平均の標本分布

母集団が正規分布している（正規母集団）とし、正規母集団から標本を取り出すことを考えます。

標本は何回も取り直せるため、標本の平均（標本平均）も異なる値をとります。よって、標本平均も確率変数と考えることができ、その分布を考えることができます。
この分布を標本平均の標本分布と言います。

母分散既知の区間推定の例

母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。
以上の条件で、母平均の95%信頼区間を求めます。

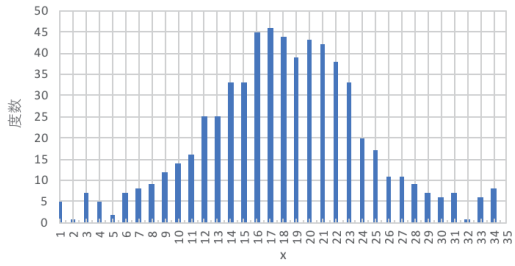
まず、標本平均$\bar{x}$を求めます。

$$\bar{x} = \frac{(5+7+5+\cdots+6+8)}{14} = 22.64$$

サンプル数をnとするとき、標本平均$\bar{x}$は
下記の式で標準化できます。

$$\frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}}$$

母集団



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

母分散既知の区間推定

ここで、標準正規分布を考えます。
右下の図のように、面積が95%となる上限値/下限値を求めます。
標準正規分布表[https://bellcurve.jp/statistics/course/8888.html]から、面積が95%となる上限値と下限値はそれぞれ+1.96、−1.96だとわかります。

正規化した標本平均は下記ようになります。

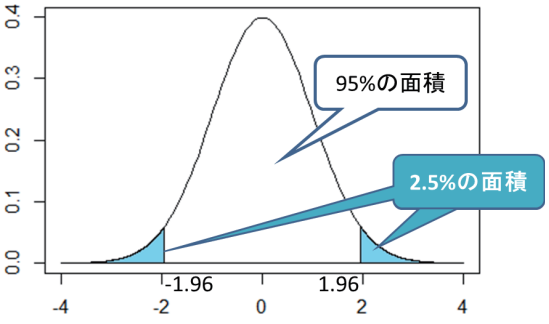
$$-1.96 \leq \frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96$$

この式を変形すると、母平均の95%信頼区間が求められます。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$22.64 - 1.96\sqrt{\frac{222.64}{14}} \leq \mu \leq 22.64 + 1.96\sqrt{\frac{222.64}{14}}$$

$$14.83 \leq \mu \leq 30.46$$



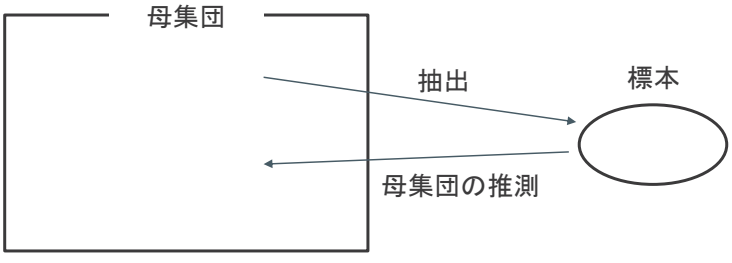
参照：https://bellcurve.jp/statistics/course/8888.html

母分散未知の区間推定の例

母分散未知の区間推定

前ページまでで、母分散既知の母平均の信頼区間の算出方法について学びました。
しかし、母平均が分からないのに母分散だけは分かっているという状況はほとんどありません。

ここからは、母分散未知の場合について学習します。
母集団未知の場合、母平均の区間推定を行うにはt分布（あるいはStudentのt分布ともいいます）を用います。



母分散未知の区間推定

母分散既知のときは下記の式を使用しましたが、母分散未知ではこの式は使用できません。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

母分散未知のときは、母分散σ^2の代わりに不偏分散s^2を使用します。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{14}(x_i - \bar{x})^2 = 310.09$$

また、母分散既知の際は1.96を求めるのに標準正規分布表を使用しましたが、母分散未知の場合はt分布表を使用します。

母分散未知の区間推定

t分布表[https://bellcurve.jp/statistics/course/8970.html]を参照します。
14個のデータから母平均の95%信頼区間を求めるので、

$$\alpha = 0.025$$
$$v = 13$$

が交わるところを参照すると、2.160であることがわかります。

$$\bar{x} - 2.160\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + 2.160\sqrt{\frac{s^2}{n}}$$

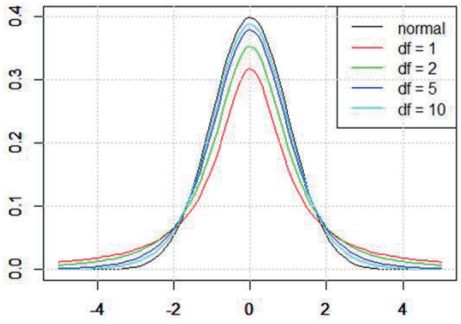
標本平均、不偏分散、サンプル数をそれぞれ代入すると、母平均の95%信頼区間を求めることができます。

$$22.64 - 2.160\sqrt{\frac{310.09^2}{14}} \leq \mu \leq 22.64 + 2.160\sqrt{\frac{310.09^2}{14}}$$
$$12.48 \leq \mu \leq 32.81$$

Studentのt分布

Studentのt分布は標準正規分布とよく似た形の分布で、「自由度」をパラメータとして分布の形が変わるという特徴を持っています。

自由度（グラフ中ではdfで表示しています）を変化させた時のt分布の形を右下の図に示します。
自由度が1、2、5、10と大きくなるにつれ、標準正規分布（黒線：normal）に近づくことが分かります。



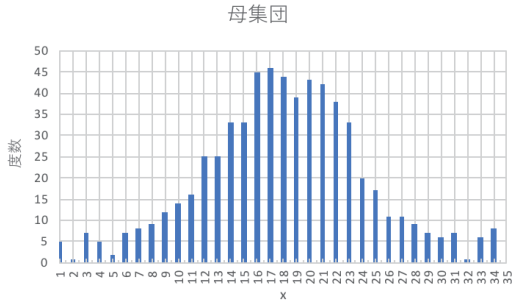
参照：https://bellcurve.jp/statistics/course/8968.html

演習問題

演習1：母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。

母平均の96%信頼区間を求めてください。

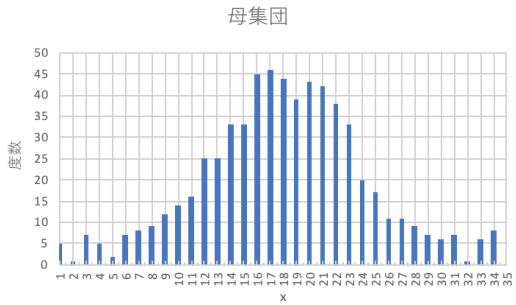


No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

演習2：母分散未知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、ランダムに14個のデータを抽出しました。

母分散未知として、母平均の90%信頼区間を求めてください。



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

演習3：Pythonによる区間推定

- scipyのstats関数を使用し、母分散未知の際の母平均を推定してください。
- 参考[演習/Chapter11_interval_estimation.ipynb]

第12回：仮説検定

アジェンダ

- 統計的仮説検定と有意性
- 帰無仮説と対立仮説
- 仮説の棄却
- 有意水準
- 第一種の誤りと第二種の誤り

統計的仮説検定と有意性

統計的仮説検定とは、母集団についての仮説を標本に基づいて検証することです。
具体的には、標本から得られる数値と仮説の差が意味を持つのか、誤差の範囲内なのかを検証します。

仮説との差が意味を持つのであれば、その差を「有意差がある」といいます。

帰無仮説と対立仮説

帰無仮説とは、「ある仮説」が正しいかどうか判断するために立てる仮説です。
たいていは、否定されることを期待した仮説を立てます。

対立仮説とは、帰無仮説とは対立している「証明したい仮説」のことです。

帰無仮説と対立仮説の例

コインを20回投げ、15回表が出たとします。
このコインは通常のコインではなく、表がでやすいコインなのではないかと疑っています。

証明したい仮説である対立仮説H_1を立てます。
コインで表が出る確率Pが1/2ではないので、以下のようになります。

対立仮説：コインで表がでる確率Pは1/2ではない。-> $H_1 : P \neq 1/2$

次に、対立仮説とは反対の意味の、帰無仮説H_0を考えます。
対立仮説とは反対の意味になるので、コインで表が出る確率Pが1/2になるとします。

帰無仮説：コインで表が出る確率Pは1/2である。-> $H_0 : P = 1/2$

ここで着目するのは、帰無仮説は「P = 1/2」とイコールが使われているため、帰無仮説が正しいと仮定したときに「コインを20回投げ、15回表が出る」が起こりうる確率を計算することができます。
しかし、対立仮説は「≠」が使われているため、確率の計算ができません。

仮説の棄却

前ページでは、「対立仮説は確率の計算ができず、帰無仮説は確率の計算ができる。」ということを記載しました。
証明したい対立仮説を直接計算することができないため、いったん反対の意味を持つ帰無仮説を立て、帰無仮説が正しいと仮定したときに事象が起こりうる確率を計算します。

帰無仮説H_0が正しいときに20回中15回表が出た事象がかなりまれであった場合、統計学的仮説検定ではまれな出来事は起こらないと判断するので、「帰無仮説H_0は正しい」を棄却（reject）し、対立仮説H_1が正しいと判断します。

有意水準

帰無仮説H_0を正しいと仮定したとき、20回中15回以上表が出る確率は0.0207となります。
この確率を基に、帰無仮説を棄却するかどうかを検討します。

確率がある一定の基準値を下回った場合に、この事象は稀にしか起こらないということで、帰無仮説を棄却します。
この基準値のことを有意水準αといいます。

有意水準には

$$\alpha = 0.1,\ 0.05,\ 0.01$$

などがよく用いられます

有意水準が0.05であるとした場合、事象は稀にしか起こらないとみなされ、帰無仮説は棄却されます。
有意水準が0.01であるとした場合、事象はよく起こることとみなされ、帰無仮説は棄却されません。

第一種の誤りと第二種の誤り

仮説検定は確率を基にしていますので、間違える可能性があります。仮説検定から得るのは、ある種の判断（judgment）であって、真実（truth）ではありません。
仮説検定で生じる間違いには2パターンがあります。

コイン投げの例でいえば、コインで表が出る確率が1/2（帰無仮説が正しい）なのに、帰無仮説を棄却してしまうことがあります。これを第一種の誤りといいます。
一方、コインで表が出る確率が1/2でない（対立仮説が正しい）のに、帰無仮説を棄却しない場合を第二種の誤りといいます。

		真実	
		帰無仮説が正しい	対立仮説が正しい
検定の結果	帰無仮説を棄却しない	正しい	第二種の誤り（β）
	帰無仮説を棄却する	第一種の誤り（α）	正しい

演習問題

演習1：対立仮説と帰無仮説

コインを20回投げ、14回表が出たとします。
このコインは通常のコインではなく、表がでやすいコインなのではないかと疑っています。

- 対立仮説と帰無仮説を立ててください。
- 帰無仮説が正しいと仮定した場合に、「コインを20回投げ、14回表が出る」確率を求めてください。
- 有意水準$\alpha = 0.1$のとき、帰無仮説が棄却されるか否かを判定してください。
- 有意水準$\alpha = 0.01$のとき、帰無仮説が棄却されるか否かを判定してください。

演習2：第一種の誤りと第二種の誤り

第一種の誤りと第二種の誤りについて、事例を2つ挙げてください。

第13回：検定統計量

アジェンダ

- 検定統計量
- 統計量z
- 統計量t
- 棄却域と採択域
- 両側検定と片側検定

検定統計量

「コインを20回投げ、15回表が出る」などはその事象が発生する確率を計算できますが、直接は確率が計算できない場合があります。
例えば身長などは、「身長が175cmになる確率」というのはわかりません。
このような場合、身長の値を「検定するための値」に変換する必要があります。この検定するための値を、検定統計量といいます。

統計量z

統計量z（z値）とは、平均が0、分散が1となるようにデータを標準化した値のことです。
例えば、標本平均を標準化した値は次の式で算出できます。

$$z = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$

$\bar{x}$：データの平均
μ：母平均
σ^2：母分散
n：サンプルサイズ

統計量zは標準正規分布に従うため、統計量zを用いた検定を行う際には標準正規分布を使います。

統計量t

統計量t(t値)とは、母分散が未知の場合に不偏分散を使用して計算した値のことです。

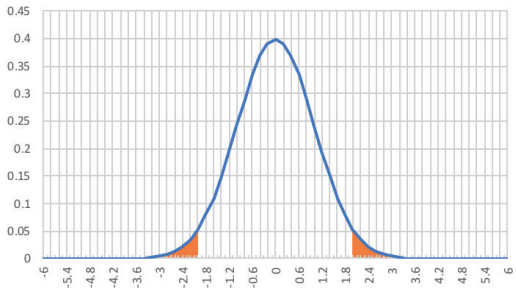
$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$:データの平均
μ:母平均
s^2:不偏分散
n:サンプルサイズ

統計量tは、サンプルサイズがnの場合は自由度(n − 1)のt分布に従います。

統計量tによる検定の例

日本人男性100人をランダム抽出し、身長を測定しました。
標本平均が174cm、不偏分散が100となりました。
身長の分布が正規分布に従うと仮定する場合、日本人の男性の平均身長は178cmと言ってよいでしょうか。



統計量tによる検定の例

以下のように帰無仮説と対立仮説を立てます。

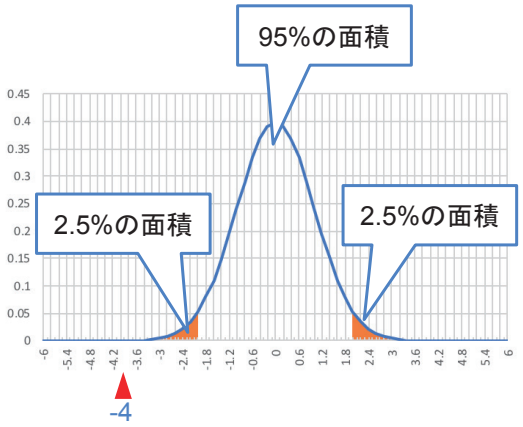
帰無仮説H_0：日本人の平均身長は178cmである。

対立仮説H_1：日本人の平均身長は178cmでない。

母分散は未知のため、不偏分散を用いて統計量tを算出します。

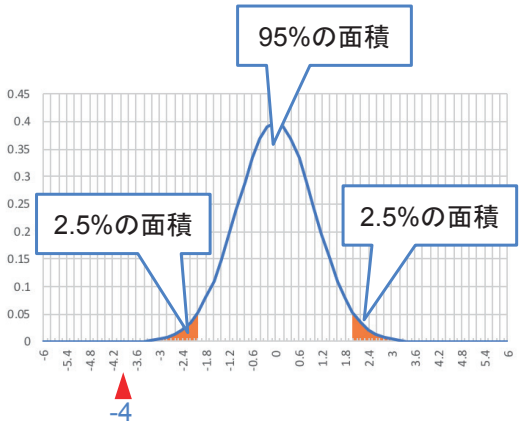
$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{174-178}{\sqrt{\frac{100}{100}}} = -4$$

有意水準を5%とする場合、t値がオレンジの領域にあれば
帰無仮説H_0は棄却されます。
今回t値は−4なので、帰無仮説は棄却されます。
日本人の平均身長は178cmではない、ということになります。



棄却域と採択域

オレンジの領域は帰無仮説が棄却される領域なので、「棄却域」といいます。
また、オレンジでない部分は帰無仮説が棄却されない領域なので、「採択域」といいます。



両側検定と片側検定

第12回で取り上げたコイン投げの事例では、以下のように対立仮説を立てることもできます。

1. 対立仮説：コインで表がでる確率Pは1/2ではない。-> $H_1 : P \neq 1/2$
2. 対立仮説：コインで表がでる確率Pは1/2より大きい。-> $H_1 : P > 1/2$
3. 対立仮説：コインで表がでる確率Pは1/2より小さい。-> $H_1 : P < 1/2$

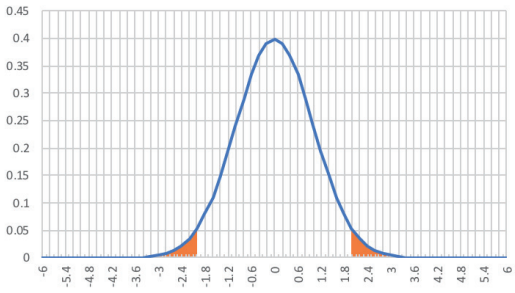
1番目の対立仮説を立てる場合、「20回中15回以上表が出る確率」に加えて、「20回中5回以下表が出る確率」も考慮する必要があります。
帰無仮説H_0を正しいと仮定したとき、20回中15回以上表が出る確率は0.0207で、20回中5回以下表が出る確率は0.0207です。
この場合、P = 0.0207 + 0.0207 = 0.0414が有意水準$\alpha = 0.05$を下回ることが、帰無仮説を棄却するためには必要です。
この様に、表が出やすい場合だけでなく表が出にくい場合も検定することを、両側検定といいます。

2番目や3番目の仮説のように、どちらか一方のみ検定することを片側検定といいます。

演習問題

演習1：統計量tによる検定

日本人男性100人をランダム抽出し、身長を測定しました。
標本平均が177cm、不偏分散が100となりました。
身長の分布が正規分布に従うと仮定する場合、日本人の男性の平均身長は178cmと言ってよいでしょうか。



第14回：Studentのt検定

アジェンダ

- t検定の種類
- 1標本問題のt検定
- 2標本問題のt検定

t検定の種類

t検定は3種類に大別できます。
1. 正規分布に従う一つの母集団の、母平均が特定の値と等しいかの検定（1標本問題）

2. 正規分布に従う、二つの母集団の母平均の差（有意差が認められるか否か）に関する検定（2標本問題）
- 2つの標本の母分散が等しいと仮定した上で行う検定。
- 2つの標本の等分散性を仮定出来ない時に行う検定。
- 検定の対象となる2つの標本において、標本の一つ一つが対になっている、もしくは何らかの関係が認められるときに行う検定。（例：受験者が同じのテストで2回調査するとき）

3. 回帰分析における回帰直線の回帰係数が0であるかに関する検定

第14回の学習内容

本講義では以下の検定について取り扱います。
- 1標本問題のt検定
- 2標本問題のt検定（2つの標本の母分散が等しいと仮定した上で行う検定）

その他の2標本問題や回帰分析における問題については、webサイトなどを参考にしてください。

1標本問題のt検定

統計量t

統計量t(t値)とは、母分散が未知の場合に不偏分散を使用して計算した値のことです。

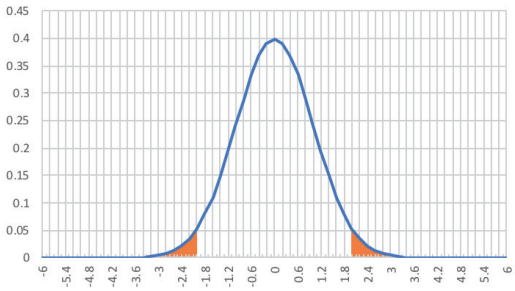
$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$:データの平均
μ:母平均
s^2:不偏分散
n:サンプルサイズ

統計量tは、サンプルサイズがnの場合は自由度(n − 1)のt分布に従います。

統計量tによる検定の例

日本人男性100人をランダム抽出し、身長を測定しました。
標本平均が174cm、不偏分散が100となりました。
身長の分布が正規分布に従うと仮定する場合、日本人の男性の平均身長は178cmと言ってよいでしょうか。



統計量tによる検定の例

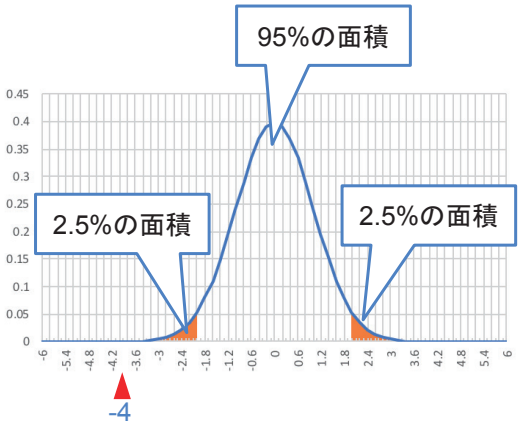
以下のように帰無仮説と対立仮説を立てます。

帰無仮説H_0：日本人の平均身長は178cmである。
対立仮説H_1：日本人の平均身長は178cmでない。

母分散は未知のため、不偏分散を用いて統計量tを算出します。

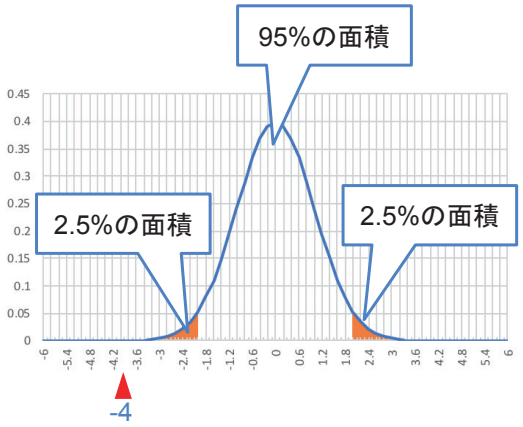
$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{174-178}{\sqrt{\frac{100}{100}}} = -4$$

有意水準を5%とする場合、t値がオレンジの領域にあれば
帰無仮説H_0は棄却されます。
今回t値は−4なので、帰無仮説は棄却されます。
日本人の平均身長は178cmではない、ということになります。



棄却域と採択域

オレンジの領域は帰無仮説が棄却される領域なので、「棄却域」といいます。
また、オレンジでない部分は帰無仮説が棄却されない領域なので、「採択域」といいます。



2標本問題のt検定
（2つの標本の母分散が等しいと仮定した上で行う検定）

2標本問題のt検定

2つの標本群X（m個の標本）とY（n個の標本）があるとします。
それぞれの平均値$\bar{X}$と$\bar{Y}$とすると、2群を合わせた分散s^2とt値は以下のように表せます。

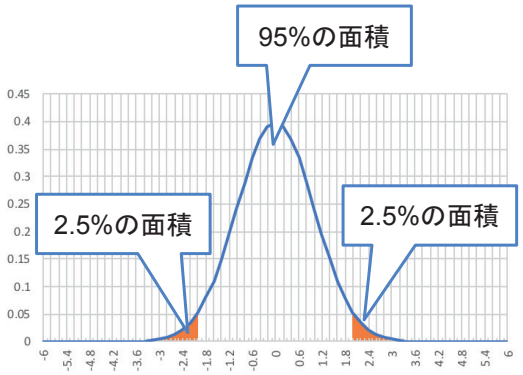
$$s^2 = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_j-\bar{Y})^2}{m+n-2}$$

$$t\text{値} = \frac{\bar{X}-\bar{Y}}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

棄却域と採択域

オレンジの領域は帰無仮説が棄却される領域なので、「棄却域」といいます。
また、オレンジでない部分は帰無仮説が棄却されない領域なので、「採択域」といいます。

2標本問題においても1標本問題と同じ様に、t値とt分布を参照して
t値が採択域にあるのか、棄却域にあるのかを判断します。



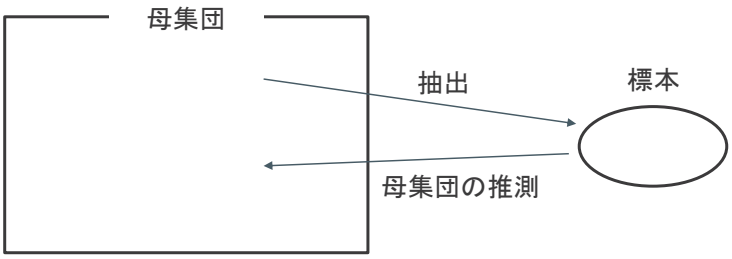
第15回：統計学Ⅱ総復習 その2

第10回：点推定

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。

実際には母集団全てを知り得ることができないことが一般的であるため、標本から母集団を推定することとなります。

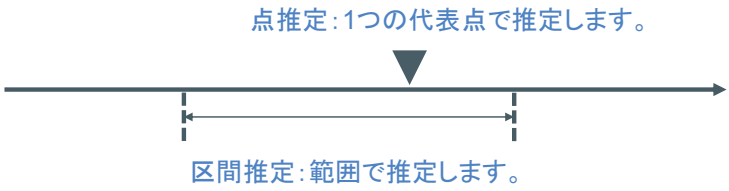


点推定と区間推定

推定には大別して点推定と区間推定があります。

点推定とは、標本から求められる一つの代表点によって母数を推定する方法です。
例えば、標本平均をそのまま母平均の推定値とするのも点推定の1つです。

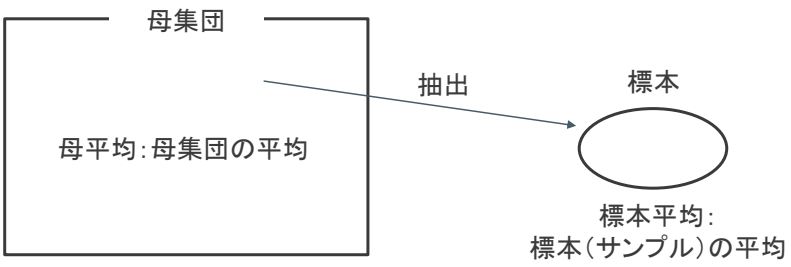
区間推定とは、標本から2つの値を計算し、その間に母数が含まれていると考える推定方法です。
例えば、95%信頼区間は区間推定の一つです。



母平均と標本平均

母集団全体の平均を母平均、サンプルの平均を標本平均といいます。

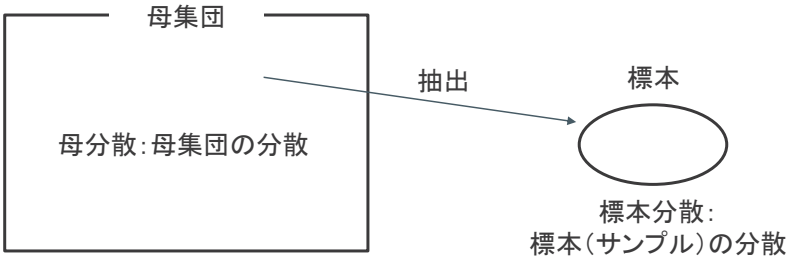
母集団を固定した場合、母平均は1つの確定した値になります。
標本平均は抽出の方法に依存した値となります。



母分散と標本分散

母集団全体の分散を母分散、サンプルの分散を標本分散といいます。

母集団を固定した場合、母分散は1つの確定した値になります。
標本分散は抽出の方法に依存した値となります。



推定量と推定値

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。
例えば平均は以下のように表します。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

統計量やその関数のことを推定量といいます。

また、実際に試行し得た値より計算した数値のことを、推定値といいます。

不偏性と一致性

標本の統計量から母集団のパラメータを推定するとき、標本の統計量のことを推定量といいます。

不偏性（偏りがない）とは、推定量の期待値が、真の母数の値となることです。
これを満たす推定量を不偏推定量といいます。

一致性とは標本サイズが大きくなるほど、推定量が母集団の真の値に近づいていくことです。

不偏分散とは

標本したn個のデータから計算した分散が標本分散で、以下のように計算します。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

母集団に比べ標本数が少ない時は、標本分散が母分散よりも小さくなります。
そこで、標本分散が母分散に等しくなるように補正したものを不偏分散といい、以下のように計算できます。

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$\bar{X}$:標本の平均

第11回：区間推定

区間推定とは

母集団が正規分布に従うと仮定できるときに、標本から得られた値からある区間でもって母平均などの母数を推定する方法を区間推定といいます。

このときの区間のことを「信頼区間」といいます（「CI」と表記されることがあります）。

母分散既知/母分散未知と区間推定

母平均の区間推定では、母分散が分かっている場合と分からない場合とで、その算出方法が異なります。
母分散が分かっている場合を「母分散既知」、母分散が分からない場合を「母分散未知」といいます。

母分散既知の場合
- 母分散の値を使い、標準正規分布を用いて信頼区間を算出します。

母分散未知の場合
- 不偏分散の値を使い、t分布を用いて信頼区間を算出します。

母分散既知/母分散未知と区間推定

母平均は分からず母分散だけは分かっている、という状況は現実にはほとんどありません。
したがって、母平均の区間推定を行う場合にはt分布が用いられることがほとんどです。

母平均の区間推定では「95%信頼区間（95%CI）」を求めることが多々あります。
これは「母集団から標本平均を求めるという作業を100回実施した際、95回はその標本平均を含んでいると考えられる区間のこと」です。

このように、ある区間に母数が含まれる確率のことを「信頼係数」あるいは「信頼度」といいます。

標本平均の標本分布

母集団が正規分布している（正規母集団）とし、正規母集団から標本を取り出すことを考えます。

標本は何回も取り直せるため、標本の平均（標本平均）も異なる値をとります。よって、標本平均も確率変数と考えることができ、その分布を考えることができます。
この分布を標本平均の標本分布と言います。

母分散既知の区間推定

正規分布にノイズを乗せて作成したダミーデータの区間推定を取り扱います。
35のデータがあり、母平均$\mu = 18.14$、母分散は$\sigma^2 = 222.64$です。
そのうち、ランダムに14個のデータを抽出しました。
以上の条件で、母平均の95%信頼区間を求めます。

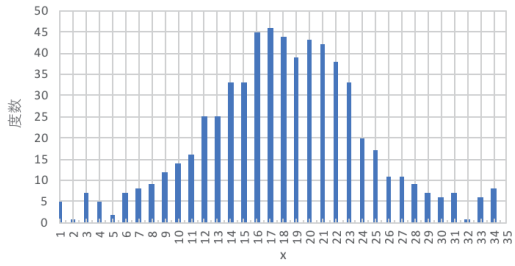
まず、標本平均$\bar{x}$を求めます。

$$\bar{x} = \frac{(5+7+5+\cdots+6+8)}{14} = 22.64$$

サンプル数をnとするとき、標本平均$\bar{x}$は
下記の式で標準化できます。

$$\frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}}$$

母集団



No	数値
1	5
3	7
4	5
6	7
12	25
16	45
17	46
20	43
21	42
22	38
23	33
31	7
33	6
34	8

母分散既知の区間推定

ここで、標準正規分布を考えます。
右下の図のように、面積が95%となる上限値/下限値を求めます。
標準正規分布表[https://bellcurve.jp/statistics/course/8888.html]から、面積が95%となる上限値と下限値はそれぞれ+1.96、−1.96だとわかります。

正規化した標本平均は下記ようになります。

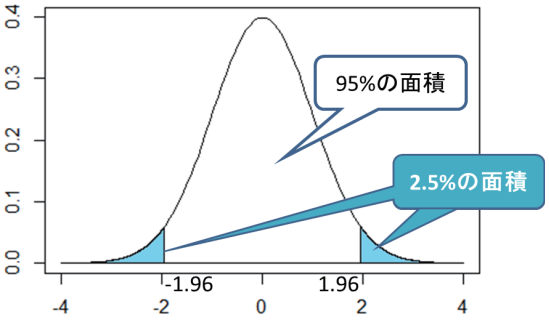
$$-1.96 \leq \frac{(\bar{x}-\mu)}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96$$

この式を変形すると、母平均の95%信頼区間が求められます。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

$$22.64 - 1.96\sqrt{\frac{222.64}{14}} \leq \mu \leq 22.64 + 1.96\sqrt{\frac{222.64}{14}}$$

$$14.83 \leq \mu \leq 30.46$$

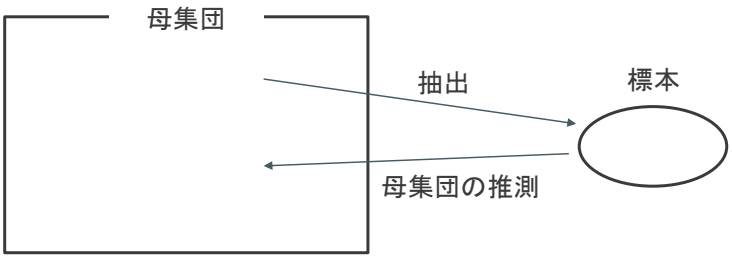


参照：https://bellcurve.jp/statistics/course/8888.html

母分散未知の区間推定

前ページまでで、母分散既知の母平均の信頼区間の算出方法について学びました。
しかし、母平均が分からないのに母分散だけは分かっているという状況はほとんどありません。

ここからは、母分散未知の場合について学習します。
母集団未知の場合、母平均の区間推定を行うにはt分布（あるいはStudentのt分布ともいいます）を用います。



母分散未知の区間推定

母分散既知のときは下記の式を使用しましたが、母分散未知ではこの式は使用できません。

$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}}$$

母分散未知のときは、母分散σ^2の代わりに不偏分散s^2を使用します。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{14}(x_i - \bar{x})^2 = 310.09$$

また、母分散既知の際は1.96を求めるのに標準正規分布表を使用しましたが、母分散未知の場合はt分布表を使用します。

母分散未知の区間推定

t分布表[https://bellcurve.jp/statistics/course/8970.html]を参照します。
14個のデータから母平均の95%信頼区間を求めるので、

$$\alpha = 0.025$$
$$v = 13$$

が交わるところを参照すると、2.160であることがわかります。

$$\bar{x} - 2.160\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + 2.160\sqrt{\frac{s^2}{n}}$$

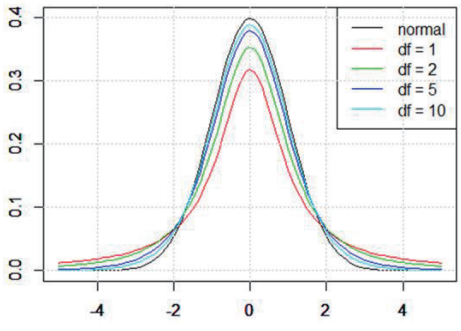
標本平均、不偏分散、サンプル数をそれぞれ代入すると、母平均の95%信頼区間を求めることができます。

$$22.64 - 2.160\sqrt{\frac{310.09^2}{14}} \leq \mu \leq 22.64 + 2.160\sqrt{\frac{310.09^2}{14}}$$
$$12.48 \leq \mu \leq 32.81$$

Studentのt分布

Studentのt分布は標準正規分布とよく似た形の分布で、「自由度」をパラメータとして分布の形が変わるという特徴を持っています。

自由度（グラフ中ではdfで表示しています）を変化させた時のt分布の形を右下の図に示します。
自由度が1、2、5、10と大きくなるにつれ、標準正規分布（黒線：normal）に近づくことが分かります。



参照：https://bellcurve.jp/statistics/course/8968.html

第12回：仮説検定

統計的仮説検定と有意性

統計的仮説検定とは、母集団についての仮説を標本に基づいて検証することです。
具体的には、標本から得られる数値と仮説の差が意味を持つのか、誤差の範囲内なのかを検証します。

仮説との差が意味を持つのであれば、その差を「有意差がある」といいます。

帰無仮説と対立仮説

帰無仮説とは、「ある仮説」が正しいかどうか判断するために立てる仮説です。
たいていは、否定されることを期待した仮説を立てます。

対立仮説とは、帰無仮説とは対立している「証明したい仮説」のことです。

帰無仮説と対立仮説の例

コインを20回投げ、15回表が出たとします。
このコインは通常のコインではなく、表がでやすいコインなのではないかと疑っています。

証明したい仮説である対立仮説H_1を立てます。
コインで表が出る確率Pが1/2ではないので、以下のようになります。

対立仮説：コインで表がでる確率Pは1/2ではない。-> $H_1 : P \neq 1/2$

次に、対立仮説とは反対の意味の、帰無仮説H_0を考えます。
対立仮説とは反対の意味になるので、コインで表が出る確率Pが1/2になるとします。

帰無仮説：コインで表が出る確率Pは1/2である。-> $H_0 : P = 1/2$

ここで着目するのは、帰無仮説は「P = 1/2」とイコールが使われているため、帰無仮説が正しいと仮定したときに「コインを20回投げ、15回表が出る」が起こりうる確率を計算することができます。
しかし、対立仮説は「≠」が使われているため、確率の計算ができません。

仮説の棄却

前ページでは、「対立仮説は確率の計算ができず、帰無仮説は確率の計算ができる。」ということを記載しました。
証明したい対立仮説を直接計算することができないため、いったん反対の意味を持つ帰無仮説を立て、帰無仮説が正しいと仮定したときに事象が起こりうる確率を計算します。

帰無仮説H_0が正しいときに20回中15回表が出た事象がかなりまれであった場合、統計学的仮説検定ではまれな出来事は起こらないと判断するので、「帰無仮説H_0は正しい」を棄却（reject）し、対立仮説H_1が正しいと判断します。

有意水準

帰無仮説H_0を正しいと仮定したとき、20回中15回以上表が出る確率は0.0207となります。
この確率を基に、帰無仮説を棄却するかどうかを検討します。

確率がある一定の基準値を下回った場合に、この事象は稀にしか起こらないということで、帰無仮説を棄却します。
この基準値のことを有意水準αといいます。

有意水準には

$$\alpha = 0.1,\ 0.05,\ 0.01$$

などがよく用いられます

有意水準が0.05であるとした場合、事象は稀にしか起こらないとみなされ、帰無仮説は棄却されます。
有意水準が0.01であるとした場合、事象はよく起こることとみなされ、帰無仮説は棄却されません。

第一種の誤りと第二種の誤り

仮説検定は確率を基にしていますので、間違える可能性があります。仮説検定から得るのは、ある種の判断（judgment）であって、真実（truth）ではありません。
仮説検定で生じる間違いには2パターンがあります。

コイン投げの例でいえば、コインで表が出る確率が1/2（帰無仮説が正しい）なのに、帰無仮説を棄却してしまうことがあります。これを第一種の誤りといいます。
一方、コインで表が出る確率が1/2でない（対立仮説が正しい）のに、帰無仮説を棄却しない場合を第二種の誤りといいます。

		真実	
		帰無仮説が正しい	対立仮説が正しい
検定の結果	帰無仮説を棄却しない	正しい	第二種の誤り（β）
	帰無仮説を棄却する	第一種の誤り（α）	正しい

第13回：検定統計量

統計量z

統計量z(z値)とは、平均が0、分散が1となるようにデータを標準化した値のことです。
例えば、標本平均を標準化した値は次の式で算出できます。

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

$\bar{x}$:データの平均
μ:母平均
σ^2:母分散
n:サンプルサイズ

統計量zは標準正規分布に従うため、統計量zを用いた検定を行う際には標準正規分布を使います。

統計量t

統計量t(t値)とは、母分散が未知の場合に不偏分散を使用して計算した値のことです。

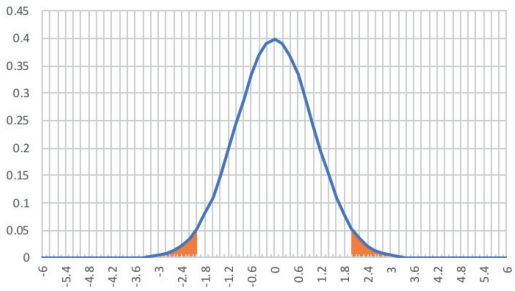
$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$:データの平均
μ:母平均
s^2:不偏分散
n:サンプルサイズ

統計量tは、サンプルサイズがnの場合は自由度(n − 1)のt分布に従います。

統計量tによる検定の例

日本人男性100人をランダム抽出し、身長を測定しました。
標本平均が174cm、不偏分散が100となりました。
身長の分布が正規分布に従うと仮定する場合、日本人の男性の平均身長は178cmと言ってよいでしょうか。



統計量tによる検定の例

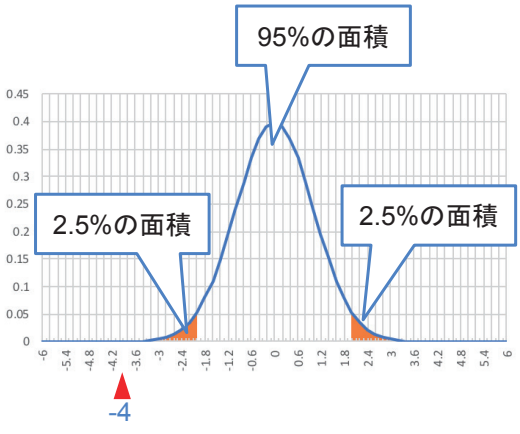
以下のように帰無仮説と対立仮説を立てます。

帰無仮説H_0：日本人の平均身長は178cmである。
対立仮説H_1：日本人の平均身長は178cmでない。

母分散は未知のため、不偏分散を用いて統計量tを算出します。

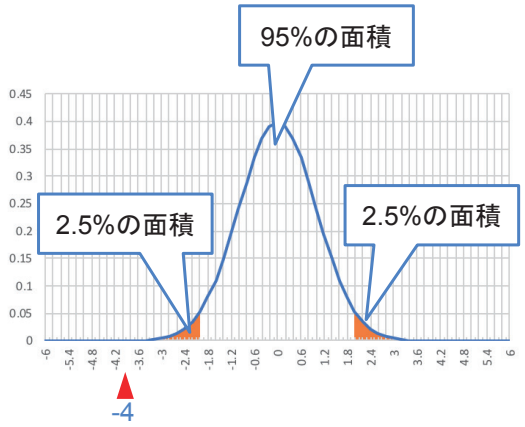
$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{174-178}{\sqrt{\frac{100}{100}}} = -4$$

有意水準を5%とする場合、t値がオレンジの領域にあれば
帰無仮説H_0は棄却されます。
今回t値は−4なので、帰無仮説は棄却されます。
日本人の平均身長は178cmではない、ということになります。



棄却域と採択域

オレンジの領域は帰無仮説が棄却される領域なので、「棄却域」といいます。
また、オレンジでない部分は帰無仮説が棄却されない領域なので、「採択域」といいます。



第14回：Studentのt検定

t検定の種類

t検定は3種類に大別できます。
1. 正規分布に従う一つの母集団の、母平均が特定の値と等しいかの検定（1標本問題）

2. 正規分布に従う、二つの母集団の母平均の差（有意差が認められるか否か）に関する検定（2標本問題）
- 2つの標本の母分散が等しいと仮定した上で行う検定。
- 2つの標本の等分散性を仮定出来ない時に行う検定。
- 検定の対象となる2つの標本において、標本の一つ一つが対になっている、もしくは何らかの関係が認められるときに行う検定。（例：受験者が同じのテストで2回調査するとき）

3. 回帰分析における回帰直線の回帰係数が0であるかに関する検定

第14回の学習内容

本講義では以下の検定について取り扱います。
- 1標本問題のt検定
- 2標本問題のt検定（2つの標本の母分散が等しいと仮定した上で行う検定）

その他の2標本問題や回帰分析における問題については、webサイトなどを参考にしてください。

2標本問題のt検定

2つの標本群X（m個の標本）とY（n個の標本）があるとします。
それぞれの平均値$\bar{X}$と$\bar{Y}$とすると、2群を合わせた分散s^2とt値は以下のように表せます。

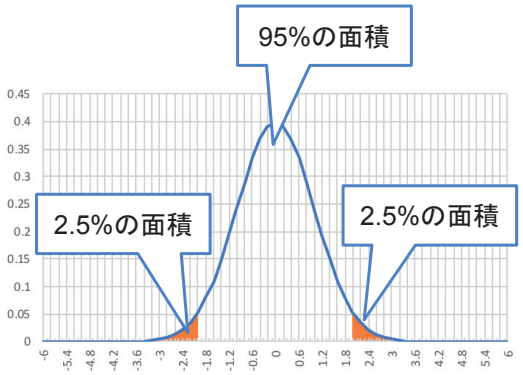
$$s^2 = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_j-\bar{Y})^2}{m+n-2}$$

$$t\text{値} = \frac{\bar{X}-\bar{Y}}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

棄却域と採択域

オレンジの領域は帰無仮説が棄却される領域なので、「棄却域」といいます。
また、オレンジでない部分は帰無仮説が棄却されない領域なので、「採択域」といいます。

2標本問題においても1標本問題と同じ様に、t値とt分布を参照して
t値が採択域にあるのか、棄却域にあるのかを判断します。



令和２年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

令和２年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

統計学Ⅱ

令和３年２月

学校法人電子学園（日本電子専門学校）
〒169-8522　東京都新宿区百人町 1-25-4
TEL 03-3369-9333　FAX 03-3363-7685